International Journal of Computer
Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 13, Issue 1, Jan–Jun 2024; 1–10
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

HEART DISEASE PREDICTION USING MACHINE LEARNING

Jayavani Vankara[1], Reeni Pragnya Mishra[2], Shaik Elijah[3], Karri Ajay[4] & Arrepu Aditya[5]

[1]Assistant Professor, Dept. of CSE, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India

[2,3,4,5]Student, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India

ABSTRACT

A heart is essential to all living things. A greater degree of accuracy, precision, and correctness is needed when diagnosing and forecasting heart-related disorders because even a minor error could have serious consequences. Heart-related deaths are on the rise, and it leads to an individual's mortality or fatigue issues. Machine learning is widely applicable everywhere in the world. There are no exceptions in the world of healthcare. Machine learning is useful in predicting heart problems, different types of diseases, and anomalies in the locomotor system. Informed by this kind of data, doctors can adjust patient diagnoses and treatment regimens when it is anticipated. Our goal with machine learning techniques is to forecast impending heart attacks. This research assesses the performance of several classifiers, such as Random Forest, SVM, naive Bayes, decision tree models, and logistic regression. Additionally, we present an ensemble classifier that can process large amounts of training and validation data by combining the best characteristics of both weak and solid classifiers to perform hybrid classification. In the medical field, there are no exceptions. Neural networks have a lot to offer predictive analytics. Diseases including heart problems, anomalies of the locomotor system, and others. If anticipated, this data can provide medical practitioners with critical insights that allow them to adjust their diagnosis and treatment plan for patients. Our goal is to predict impending heart attacks with the use of machine learning techniques. [1]

KEYWORDS: *Forecast, Precision, Excellence, Repercussions, Deaths, Illnesses, Logistic Regression, Support Vector Machines, Random Forest; Ensemble classifier*

Article History

Received: ***04 Apr 2024*** **|** ***Revised:*** ***11 Apr 2024*** **|** ***Accepted:*** ***16 Apr 2024***

INTRODUCTION

The primary objective of the most widespread causes of demise and illness worldwide is coronary artery bypass graft disease. Within data analysis, one of the most important areas is the prediction of cardiovascular disease. Globally, the prevalence of cardiovascular disease has increased dramatically in recent years. Much research has been done to identify other significant risk factors for heart disease as well as to accurately quantify the overall risk. [1] Because heart disease never shows symptoms until a person passes away, it is often known as the silent killer. Early detection is necessary for cardiovascular illness. The difficulties will decrease if high-risk people are assisted in changing their lives. [2]Making predictions and decisions from the vast volumes of data generated by the healthcare sector is made easier by machine learning. The goal of this research is to predict the incidence of heart disease in the future by analyzing patient data and applying a machine-learning algorithm to determine whether a patient has heart disease. Artificial intelligence (AI) tools can be quite useful here. Heart disease is a difficult ailment to diagnose, thus we can conclude that this procedure is highly flexible. The necessary information is extracted by statistical analysis after data is collected from multiple sources and logically categorized. [3]Some technologies are capable of predicting the risk of coronary arteries, but they are either very expensive or inefficient at estimating the likelihood that a person will develop heart disease. Heart disease mortality and its overall effects can be decreased by early identification. Since a doctor cannot see a patient around-the-clock, accurately monitoring patients on a daily basis is not always possible. It also takes more time, energy, and knowledge. Thanks to the wealth of data available today, we can look for hidden patterns using a variety of machine-learning approaches. Based on the underlying trends, medical data may be utilized to diagnose disorders. [4]

CORRELATED PIECES OF WORK

Following are some of the steps that various researchers have taken in order to acquire accurate findings using diverse methods:

The purpose of this literature review is to set the scene for the significance of heart disease as a significant public health concern that contributes significantly to morbidity and mortality. Predicting cardiac disease early and accurately is essential for both preventative and therapy strategies. Talking about machine learning is essential. We discuss our idea as a helpful tool to increase prediction accuracy in this situation. This assessment of the literature looks at the advancement of learning from data applications in the prediction of coronary artery bypass and is based on research publications published between 2018 and 2022. [3]

I've been interested in P.S. HIREMANTH AND S.N. PATIL's "A Hybrid Machine Learning Approach for Automated Heart Disease Risk Prediction" (2018). This study was released in 2018 and focuses on expert systems with applications. In order to forecast cardiac disease, risk factor assessment is crucial, and deep learning can help with this. For example, I am aware of this. [5] In 2020, Vijeta Sharma, Shrinkhala Yadav, and Manjari Gupta conducted a study. naïve Bayes, decision tree techniques, and support vector machines were the three methods used. Their endeavour achieved ninety, eighty-five, and ninety-six percentile accuracy. [6] Archana Singh and Rakesh Kumar finished a project in 2020. They produced 83% and 78% accuracy rates, respectively, using the two different methodologies of SVM and linear regression. [7] Dr. Geeta S and Mr. Santhana Krishnan J implemented two distinct algorithms, Naive Bayes and Decision Tree, to construct a project in 2019 that provided accuracy rates of 87% and 91%, respectively. [8]Garima Choudhary and Dr. Shailendra Narayan Singh used the Adaboost algorithm to complete a project in 2020 and received an accuracy rating of 89.88%. [9]

ALGORITHM USED

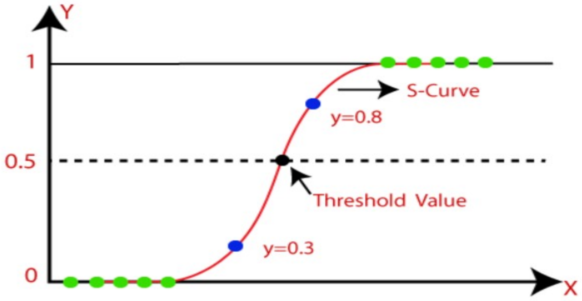
The goal of the artificial intelligence field known as "algorithmic learning" (ML) is to build machines that can learn from and enhance the data they process. [4]The phrase "intelligent technology" refers to a broad category of apparatuses or systems that emulate human intelligence. Though the phrases artificial intelligence and machine learning are not interchangeable, they are frequently discussed and even used in the same sentence. There is an important distinction to be made between artificial intelligence and machine learning: although appliance learning is a part of AI, artificial brain is not the same as artificial life.

Logistically Backward

When utilizing the regression analysis predictive algorithm to predict heart disease, LR models are first trained with five splitting conditions and then assessed using test data to get the best accuracy and understand the model behavior. The algorithm returns a category of 1 or 0 in the event of cardiac sickness. [11]

Equation

$\text{Probability}(y=1|x)=1+e-(\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn)1$



Random Forest with Kernel Numerical Method

The most potent and extensively utilized predictive algorithm is called random forest. Regulated automated learning is where it fits in best. It can be used in machine learning applications for both classification and regression. To create decision trees and gather data, random forests employ a variety of samples. For the decision trees, an average is determined. It can process the categorical variable dataset and ignore missing values, albeit being slower than a single decision tree. [12]

KNN seeks to identify patterns in the values within the dataset and predictions. Since a specific set of parameters for a given functional form cannot be found, KNN employs a non-parametric approach. It doesn't make any assumptions about the dataset's properties or output. Because it memorizes the training set instead of precisely learning and updating the weights, KNN is also known as a lazy classifier. [3]This means that rather than during training, the majority of the computer effort is done during categorization. KNN frequently attempts to determine which class it is closest to in order to locate the neighboring class during the implementation of a new feature. [20]

APPROACH

To produce results, we gather patient data in this system and apply several machine learning techniques, including feature scaling, data pre-processing, model development, and one-hot coding. We will make use of a number of Python modules, including NumPy and Pandas, to visualize the data. To get at our ultimate result, we take five different models and choose the most correct one. The architecture of the Random Forest disease detection system consists of a number of processes, such as picture capture, segmentation, preprocessing, feature extraction, classification, and performance assessment. [10] All of these elements are necessary for the suggested methodology to be effective. With the use of our algorithm, we systematically predict heart illness. Encoding categorical variables, normalizing numerical properties, and addressing absent values in the dataset are our first steps. Afterwards, feature selection techniques are used to determine which characteristics are most important for predicting coronary artery bypass surgery. We then train different machine learning models, including Decision Tree Classifier, Gradient Boosting Classifier, MLP Classifier, Naive Bayes, Logistic

Regression, the k-n Neighbors, and Vector Machines for Support on the preprocessed dataset. After being trained on the training data, each model is put to the test using assessment parameters such as the F1-score, recall, accuracy, and precision. [5]

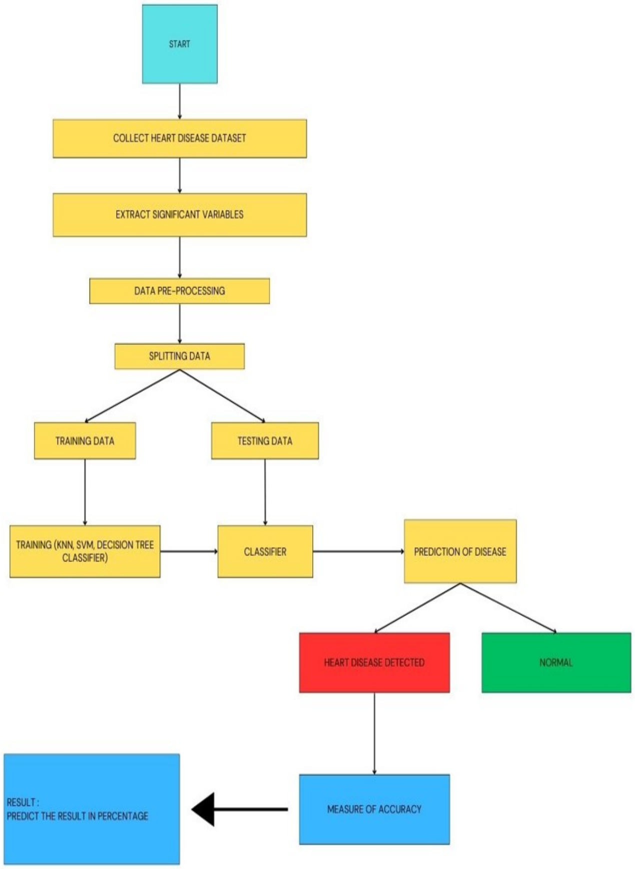
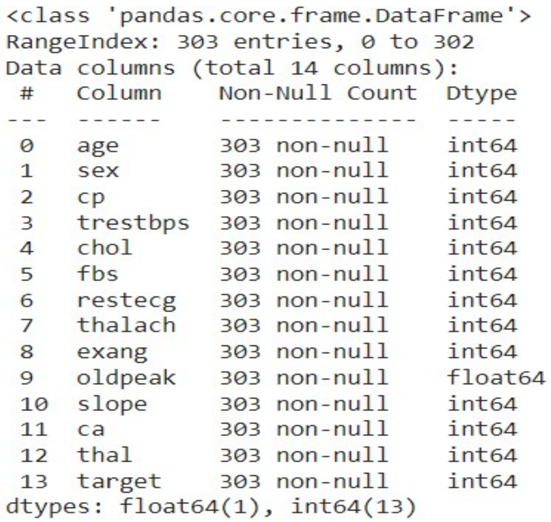


Figure 1: Diagram Illustrating the Process of Contrasting the Machine Learning Model With and Without our Tool's Image Preprocessing.

We compute ROC curves, space under the curve (AUC), confusion matrices, accuracy scores [7], and ROC curves to compare the model performance. By assessing the predictive abilities and discrimination strength of each model, we can utilize this study to identify which machine learning technique for heart disease prediction is the most accurate and practical.

Our approach, which utilizes sophisticated machinery learning algorithms and the Kaggle dataset [2], is intended to advance the field of heart disease prediction science.

Improved early diagnosis, tailored risk assessment, and improved patient outcomes are possible benefits of this research. [6]

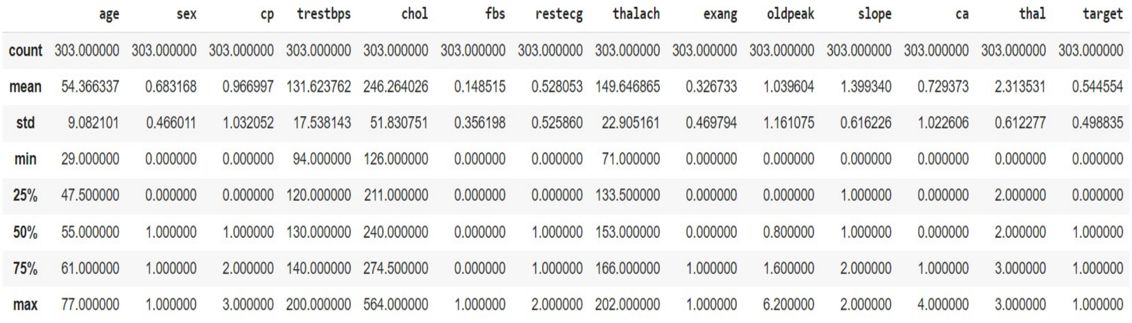


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
```

Figure 2: Variable Information.

I. Description of Datasets

Several clinical and demographic factors are included in a freely accessible heart disease dataset that is used in this study [9]. Age, sex, blood pressure, cholesterol levels, and the presence or absence of heart illness are among the parameters for which the dataset provides information. Detailed information on the dataset is provided.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
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	303.000000	303.000000
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	2.313531	0.544554
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	0.612277	0.498835
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	0.000000	0.000000
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	2.000000	0.000000
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	2.000000	1.000000
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	3.000000	1.000000
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	3.000000	1.000000

Figure 3: Data Descriptions.

II. Data Pre-processing

Prior to being used to train machine learning models, the dataset is put through preparation steps to ensure data quality and consistency. In doing so, numerical features are normalized, missing values are addressed, and categorical variables are encoded. Standard scaling techniques such as normalization or standardization may also be employed to ensure consistency among the input attributes.

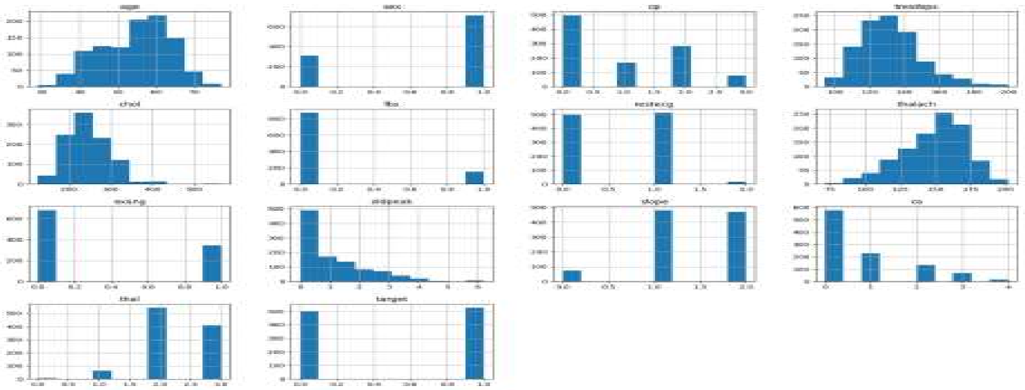


Figure 4: Being Aware of the Variable Distribution of Available Dataset.

III. Feature Selection

Finding the most relevant data for heart disease prediction involves applying feature selection techniques. The goal of these approaches is to reduce dimensionality and enhance model performance by selecting a subset of relevant features. Recursive feature elimination (RFE) and correlation examinations are popular methods for ranking and selecting the most significant features.

IV. Model Training and Evaluation

Numerous algorithms for learning, such as Gradient Boosting Classifier, MLP Classifier, Decision Tree Classifier, Support Vector Machines, Logistic Regression, K-Nearest Neighbors, and Naive Bayes, are used to tackle the problem of cardiac illness prediction. Each model is trained using the preprocessed dataset and assessed using standard metrics like recall, accuracy, and F1-score. [10]

V. Result & Discussion

About coronary artery disease, the prediction result is genuinely individualized, taking into account variables like age, sex, kind of chest pain, resting blood pressure, cholesterol, blood sugar levels after fasting, ECG reading, maximal heart rate, ST depression, and angina during rest. Overall project accuracy was determined to be 91%. The findings show that while some algorithms, including SVC and Decision Tree, produce better results, KNN, Random Forest Classifier, and Logistic Regression outperform these algorithms in the diagnosis of cardiac disease in patients, according to the majority of academics [11]. Our method saves a substantial amount of money and is faster and more accurate than the ones used by previous researchers. Furthermore, the maximum accuracy of 88.5% attained with KNN and Logistic Regression is greater than or almost equivalent to the accuracy of prior research. [13] We discovered that our accuracy has increased due to using the extra medical variables from the dataset.

Our findings also show that KNN and logistic regression outperform random forest classifiers in predicting which patient would receive a heart disease diagnosis. This proves that logistic regression and KNN are better at diagnosing heart disease. The numbers of patients who have been categorized and predicted by the classifier based on variables including age group, resting blood pressure, sex, and chest discomfort are plotted in Figures 5 through 8.

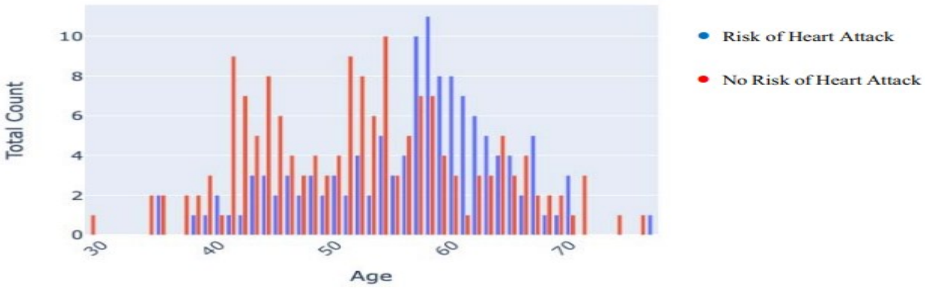


Figure 5: Displays a Person's Age-Based Heart Attack Risk.

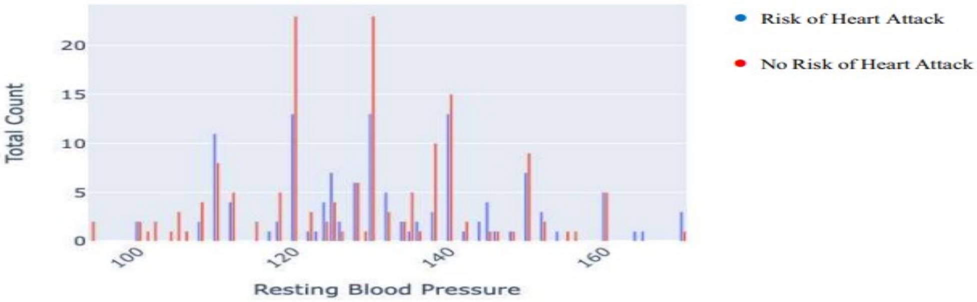


Figure 6: Demonstrates the Patient's Risk of a Heart Attack Based on their Blood Pressure at Rest.

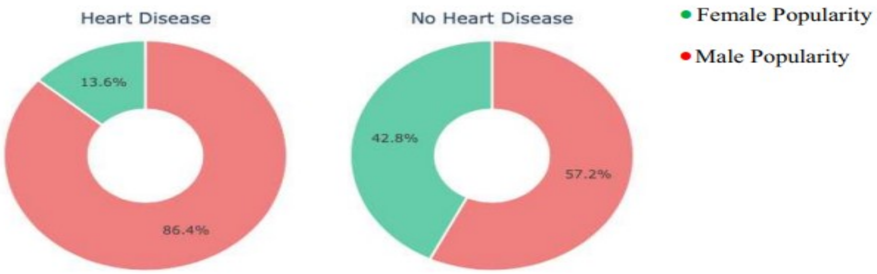


Figure 7: Shows the Status of Patients' Heart Disease Based on their Sex.

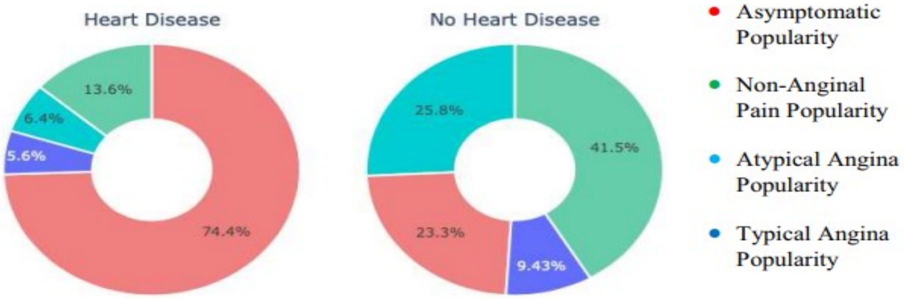


Figure 8: Indicates an Individual's Heart Disease Condition Based on the Kind of Chest Discomfort they are Feeling.

CONCLUSION

The recommended solution is GUI-based, scalable, extendable, and intuitive. The early and rapid diagnosis that the suggested operational paradigm provides also has the added benefit of reducing treatment expenses. Cardiologists, other medical professionals, and students studying medicine can use the model as a soft diagnostic tool and teaching assistance. General practitioners can use this method to diagnose cardiac patients initially. The accuracy and scalability of this prediction system can be improved in several ways, as described in [15]. Now that a universal design has been created, we can use it in an upcoming analysis of various data sets. Another promising direction for future research is coordinating multiple class labels in the prediction process, which might significantly increase the accuracy of medical diagnosis. [13] Future studies may find it challenging to pinpoint and choose crucial factors for a more precise diagnosis of heart disease due to the dimensionality of the cardiac database in DM warehouses. The given model was built using the KNN, Random Forest Classifier, and Logistic Regression approaches. 87.5% is the accuracy of our model. [11] The model will be more accurate in predicting whether or not a given person has heart disease if more training data points are available. These computer-aided solutions substantially decrease costs and enable us to predict patients more accurately and quickly. We may work with several medical datasets, benefiting doctors and patients alike because machine-learning approaches exceed human prediction. Thus, using KNN, logistic regression, and dataset cleaning, our research enables us to forecast the individuals who will receive a heart disease diagnosis. Our estimation accuracy is 87.5% compared to prior models' 85% accuracy in [2].

REFERENCES

1. *Cardiovascular disease (CVD): evaluation prediction and policy implications, S. Rehman, E. Rehman, M. Ikram, and Z. Jianglin, BMC Public Health, vol. 21, no. 1, pp. 1299, 2021.*

2. *The article "Heart Disease Prediction using Machine Learning" was published in the April 2020 issue of the International Journal of Engineering Research and Technology (IJERT) by Apurb Rajdhan, Avi Agarwal, Milan Sai, Dundigalla Ravi, and Poonam Ghuli.*

3. *Heart Disease Prediction Using Machine Learning", International Journal of Advanced Research in Science Communication and Technology, 2021, Baban Rindhe, Nikita Ahire, Patil, Rupali Gagare, Shweta Darade, and Manisha.*

4. *P. Ramprakash, R. Sarumathi, R. Mowriya and S. Nithya Vishnupriya, "Heart Disease Prediction Using Deep Neural Network", International Conference on Inventive Computation Technologies (ICICT), 2020.*

5. *G. Choudhary and S. Narayan Singh, "Prediction of Heart Disease using Machine Learning Algorithms", International Conference on Smart Technologies in Computing Electrical and Electronics (ICSTCEE), 2020.*

6. *Park S W, Yun Y D, Oh D J, Oh B H, Lee S H, Jang Y, and Jee S H (2014). A model for predicting coronary heart disease: the Korean Heart Study. e005025 in BMJ open, 4(5).*

7. *Ganna A, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013); Magnusson P K. Predicting coronary heart disease using multilocus genetic risk scores. Thrombosis, arteriosclerosis, and vascular biology, 33(9), 2267–2272.*

8. *Jabbar M A, Chandra P, & Deekshatulu B L (March 2013). Lazy associative classification for the prediction of heart disease. International Multi-Conference on Automation, Computing, Communication, Control, and Compressed Sensing (iMac4s) (pp. 40–6) came together in 2013. The IEEE.*

9. *M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur and Kushal Ghosh "Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design", September 2014.*

10. *Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin "Analysis of Data Mining Techniques for Heart Disease Prediction", May 2015.*

11. *In the August 2016 issue of IEEE, Soodeh Nikan, Femida Gwadry-Sridhar, and Michael Bauer published "Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis."*

12. *In Bhubaneswar, Odisha, India, at Orissa Engineering College, Sanjay Kumar Sen is an assistant professor of computer science and engineering."Assessing and Forecasting Heart Disease Through Machine Learning Techniques" June 2017, Volume 6, Issue 6 of the International Journal of Engineering and Computer Science*

13. *International Journal of Engineering & Technology, 7 (2.8), April 2018, V.V. Ramalingam, Ayantan Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning tech: A survey."*

14. *O.T. Ali, A.B. Nassif and L.F. Capretz. "Business intelligence solutions in healthcare: a case study: Transform ing OLTP system to BI solution." 2013 3rd International Conference on Communications and Information Technology ICCIT 2013, pp. 209-214, 2013.*

15. *Himanshu Sharma, M.A. Rizvi. "Prediction of Heart Disease using Machine Learning Algorithms: A Survey." International Journal on Recent and Innovation Trends in Computing and Communication, Volume-5, Issue-8, pp.99-104, 2017.*