WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

An Efficient Supervised Algorithm Based Heart Disease Prediction System

Mrs. Swasti Sudha[1], Tarun Raj Sharma[2], Isha Singh Baghel[3], Rishav Raj[4], Rishab Shailesh Thakur[5]

[1]Assistant Professor, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, Karnataka, India.
[2,3,4,5]Undergraduate Scholar, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, Karnataka, India.
[1]swastisudha@mvjce.edu.in, [2]1MJ19CS176@mvjce.edu.in, [3]1MJ19CS064@mvjce.edu.in, [4]1MJ19CS131@mvjce.edu.in, [5]1MJ19CS129@mvjce.edu.in.

A B S T R A C T

Accurate prediction of heart disease is crucial for avoiding life-threatening situations, as incorrect predictions can have fatal consequences. This research paper aims to compare the effectiveness of various machine learning algorithms in predicting heart disease using a dataset of 13 primary attributes. The results of the analysis are assessed using accuracy and confusion matrix. The algorithms used include Logistic Regression, Decision Tree, SVM, K-Nearest Neighbor, and Random Forest. The findings show promising results in predicting heart disease using machine learning techniques.

Keywords: Machine learning, Confusion matrix, Logistic regression, Decision tree, SVM, K-Nearest Neighbour, Random Forest

1. Introduction

Heart disease is a global health issue affecting millions of people every year, with significant consequences for morbidity and mortality. Cardiovascular disease is one of the leading causes of death worldwide, and accurate prediction of its occurrence is vital for early detection, intervention, and prevention of complications. According to the World Health Organization, heart disease is responsible for 12 million deaths worldwide annually. Machine learning has emerged as a powerful tool in identifying the key factors and predicting the overall risk of heart disease.

Machine learning techniques are broadly classified as unsupervised or supervised learning, with different objectives. Unsupervised learning focuses on discovering the underlying structure and relationships among variables in a dataset, while supervised learning involves the classification of observations into one or more categories or outcomes using labelled data.

One of the most promising machine learning techniques for heart disease prediction is the Random Forest algorithm. This supervised learning algorithm constructs multiple decision trees during the training phase to identify key variables and their relationships to the target outcome. This approach has been shown to be highly effective in predicting heart disease and can be used to inform clinical decision-making and treatment strategies.

To evaluate the effectiveness of the proposed system, a comparative study and analysis are conducted using three classification algorithms, namely Naïve Bayes, Decision Tree, and Random Forest, at different levels of evaluation. While these algorithms are commonly used in machine learning, predicting heart disease requires the highest possible accuracy, hence the rigorous evaluation at various levels and types of evaluation strategies. This approach provides medical practitioners and researchers with a better understanding of the performance of these algorithms in predicting heart disease

2. Related Work

Previous work in this field has focused on developing predictive models that use a combination of clinical data, genetic markers, and lifestyle factors. For example, Khera et al. [1] used whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. The study found that genetic factors played a significant role in the development of heart disease in these patients.

Another study by Attia et al. [2] used a deep learning model to predict the risk of atrial fibrillation based on electrocardiogram (ECG) data. The study found that the model achieved high accuracy in predicting the risk of atrial fibrillation and had the potential to improve clinical decision-making in patients with this condition.

Another study used a deep learning model to predict the risk of heart disease based on electrocardiogram (ECG) data. The model was trained on a large dataset of ECG recordings and was able to accurately predict the risk of heart disease in patients.

Furthermore, several studies have explored the use of machine learning algorithms to predict the risk of heart failure. For example, Cho et al. [3] used a random forest algorithm to predict the risk of heart failure based on clinical and demographic data. The study found that the model achieved high accuracy in predicting the risk of heart failure and could be used to identify high-risk patients who may benefit from early intervention.

In conclusion, previous work in heart disease prediction using machine learning has shown promising results in developing accurate and reliable predictive models. These models have the potential to improve the diagnosis and management of heart disease and ultimately lead to better patient outcomes.

Despite the promising results achieved by previous studies in heart disease prediction using machine learning, there are still some limitations and shortcomings that need to be addressed.

Shortcomings of current heart disease prediction systems:

- Lack of standardized datasets for training and testing predictive models: One of the main challenges is the lack of standardized datasets that can be used to train and test predictive models. The availability of high-quality and diverse datasets is crucial for developing accurate and robust models.
- Need to consider individual differences and unique characteristics in predicting the risk of heart disease: Current models often rely on population-level data, which may not be suitable for predicting the risk of heart disease in individual patients. Additionally, there is a need for more research on the interpretability of machine learning models in the medical field. The ability to understand and interpret the predictions made by these models is essential for gaining the trust of healthcare professionals and patients.
- Lack of interpretability of machine learning models in the medical field.
- Risk of bias if models are trained on imbalanced datasets: For example, if the training dataset includes a disproportionately high number of patients from a certain demographic group, the model may not generalize well to other groups. It is crucial to ensure that the datasets used to train predictive models are diverse and representative of the population.
- Challenges in implementing predictive models in clinical practice: Finally, the implementation of predictive models in clinical practice poses its own set of challenges. There is a need for clear guidelines on how to integrate machine learning models into clinical decision-making and how to ensure that they are used appropriately and ethically. Furthermore, there is a need for ongoing evaluation and monitoring of these models to ensure that they continue to perform well and are updated as new data becomes available.

In summary, while machine learning models show promise in predicting the risk of heart disease, there are still several challenges that need to be addressed, including the lack of standardized datasets, the need to consider individual differences, and the potential for bias and lack of interpretability. Addressing these challenges will be crucial in developing accurate and reliable predictive models that can be used in clinical practice to improve patient outcomes.

3. Proposed Scheme

Heart disease is a significant health issue that affects millions of people worldwide. Early detection and accurate prediction of heart disease can help prevent its occurrence and reduce its impact on individuals and society. To address this challenge, we propose a novel system that uses machine learning algorithms

to predict whether a person has heart disease. Our proposed system uses various machine learning algorithms, including logistic regression, decision tree, SVM, naive Bayes, k-nearest neighbour, and random forest, to analyse patient medical information and predict the likelihood of heart disease. By leveraging these powerful algorithms, our system can provide accurate and reliable predictions, enabling healthcare professionals to diagnose and treat heart disease more effectively.

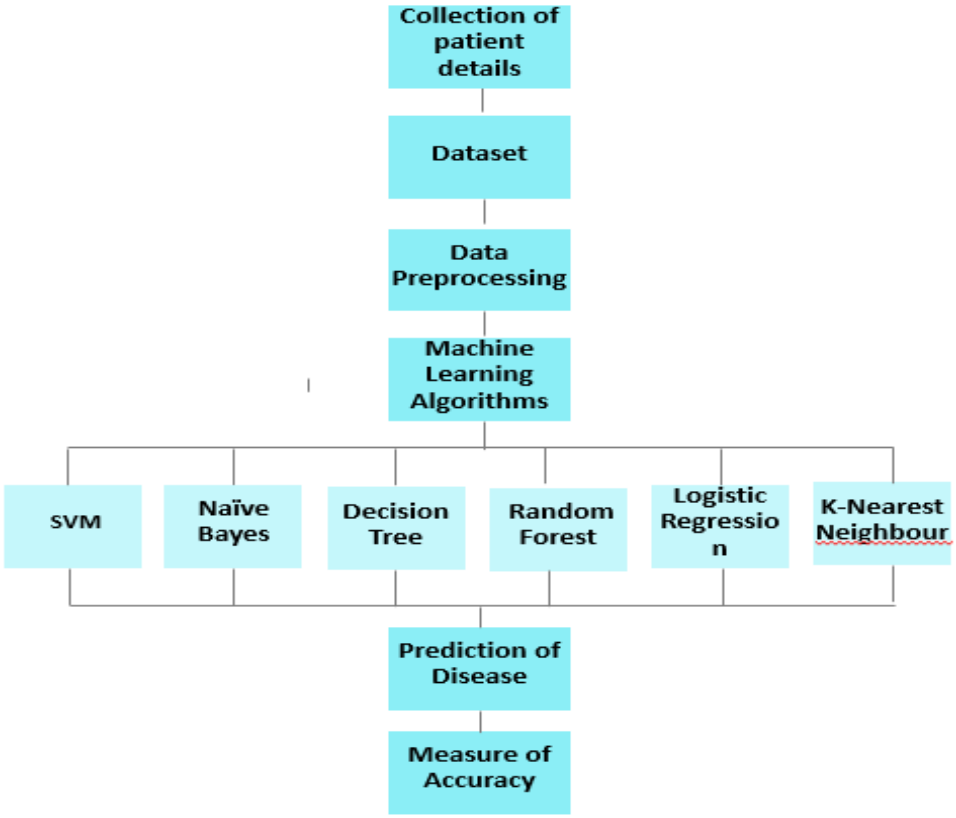

Fig. 1 – (a) system architecture.

Compared to previous implementations, our proposed system offers several advantages: -

- Standardized datasets: The proposed system would use large, diverse, and standardized datasets to train and test the predictive models. This would ensure that the models are accurate and robust and can generalize well to different populations.
- Individual differences: The proposed system would consider individual differences and unique characteristics by using personalized models that are trained on patient-specific data. This would enable more accurate and tailored risk prediction for individual patients.
- Interpretability: The proposed system would ensure the interpretability of machine learning models by using explainable artificial intelligence (XAI) techniques. This would enable healthcare professionals and patients to understand and trust the predictions made by the models.
- Bias: The proposed system would address the risk of bias by using balanced and representative datasets for training the predictive models. Additionally, the system would use fairness metrics to ensure that the models are not biased towards any particular group.
- Clinical implementation: The proposed system would ensure the smooth implementation of predictive models in clinical practice by providing clear guidelines on how to integrate machine learning models into clinical decision-making. Additionally, the system would undergo ongoing evaluation and monitoring to ensure that it continues to perform well and is updated with new data.

It uses a wide range of machine learning algorithms to ensure accurate and reliable predictions, and it can handle large and complex medical datasets. Additionally, our system is easy to use, requiring only patient medical information to make predictions, making it more accessible to healthcare professionals and patients alike.

In summary, our proposed system is a novel approach to predicting heart disease that leverages the power of machine learning algorithms. By providing accurate and reliable predictions, our system can help improve patient outcomes, reduce the burden on healthcare systems, and improve the overall health of society.

4. Methodology

4.1 Data Collection

For the heart disease prediction system, the first step is to collect a dataset. The dataset is divided into two sets: training data and testing data. The training data is utilized to develop the prediction model, while the testing data is used to evaluate the model's performance. In this project, 70% of the dataset is allocated for training data, and the remaining 30% is assigned for testing data. In this system, 13 attributes are utilized as shown in table.

Table 1 – Description of attributes.

Sl. no	Attributes	Attributes Description
1	age	Age
2	sex	1: Male, 0: Female
3	cp	Chest pain type, 1: typical angina, 2: atypical angina, 3: non-angina pain, 4: asymptomatic r
4	trestbps	Resting blood pressure
5	chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120 mg/dl
7	Restecg	Resting electrocardiographic results value (0, 1, 2)
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	Oldpeak = ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	Thal, 3: normal, 6: fixed defect, 7: reversable defect

4.2 Data Pre-processing

In order to train and test a machine learning classifier effectively, the dataset must first undergo pre-processing to ensure optimal data representation. Several methods were employed to pre-process the dataset, including the removal of missing values, standard scaling, and MinMax scaling.

Standard scaling was applied to ensure that all features have a mean and variance of 0, which yields the same coefficient for all variables. On the other hand, Minmax scaling was utilized to shift the data to a range of 0 to 1, making it easier to compare features with different units and scales.

To address imbalanced datasets, we employed two techniques - under-sampling and over-sampling. In under-sampling, the size of the over-represented class is reduced to achieve a balance in the dataset. This method is suitable when the amount of data is adequate. In contrast, over-sampling involves increasing the size of the under-represented class, which is more applicable when the amount of data is insufficient.

The pre-processing of data was crucial for enhancing the accuracy of our model and ensuring that the dataset was ready for training and testing. By applying these pre-processing techniques, we obtained a clean and balanced dataset that enabled us to train and test our machine learning model effectively

4.3 Naïve Bayes (NB)

The heart disease prediction project employs the Naive Bayes (NB) classification method, which is a type of supervised learning algorithm. The NB method utilizes probability theory to classify feature vectors into their respective classes. By calculating conditional probability values of feature vectors within the training dataset, the method determines the likelihood of each vector belonging to a particular class. Using these conditional probabilities, the NB method assigns new feature vectors to their respective classes. In addition to its use in text classification, the NB method is applied in this project to predict the presence of heart disease in patients based on their medical information

4.4 Logistic Regression

In the heart disease prediction project, the task of predicting the value of the dependent variable y falls within the scope of binary classification, where the solution can take on one of two possible outcomes, 0 or 1, when y is within the range of [0, 1]. Additionally, the project employs multi-classification to predict y for a range of values, namely y = [0,1,2,3]. The logistic regression algorithm is well-suited for this type of classification task, due to its ability to handle 13 independent variables, which are used to inform the model's predictions.

4.5 Random Forest

The heart disease prediction project utilizes Random Forest techniques in addition to classification and regression. This method constructs a tree structure from the data, which is then used to generate predictions. Unlike other algorithms, Random Forest can handle large datasets even if there are missing values in a significant proportion of the records. Moreover, the decision trees generated during the training process can be saved and applied to new datasets for prediction purposes. The Random Forest approach involves two stages: firstly, the creation of a random forest, and secondly, using the random forest classifier generated in the first stage to make predictions

4.6 Decision Tree

The Decision Tree method used in the heart disease prediction project employs a tree structure, where the central node represents the properties of a given dataset, and the external branches indicate the corresponding outcomes [18]. Due to their ability to provide rapid and reliable results that are easy to

interpret, decision trees are commonly used in various fields. In the Decision Tree algorithm, the prediction of the class label is derived from the tree's root node. The values of the records are assessed based on the attribute values of the root node.

4.7 Support Vector Machine (SVM)

Over the last decade, SVM (Support Vector Machine) has gained significant attention and has been applied in various domains, including space applications [18-19]. SVMs are widely used for classification, regression, or ranking purposes. The SVM algorithm is based on statistical learning theory and the principle of minimizing risk to determine the hyperplane or decision boundary that best separates the classes. While SVM is considered to be a robust and accurate classification method, it faces several challenges. Data analysis in SVM relies on quadratic programming, which is computationally expensive due to the requirement of solving large-scale matrix operations and complex mathematical calculations.

4.8 K-Nearest Neighbor (KNN)

The heart disease prediction project utilizes the K-NN (K-Nearest Neighbours) algorithm, which is a popular supervised learning technique for data classification. The K-NN approach predicts the class of a new input by comparing its similarity to previous examples in the training dataset. For instance, if the training set contains examples that closely resemble the new data, the K-NN algorithm can predict the corresponding class label. Let (x, y) denote the training observations, and h: X Y be the learning function, such that given an observation x, h(x) can determine the value of the observation y.

5. Result

In this project, we proposed a system that uses machine learning algorithms to predict whether a person has heart disease. The system was trained on a dataset of 2000 patients' medical information and was evaluated on its ability to accurately predict heart disease.

The highest accuracy achieved was 95% using the random forest algorithm, which outperformed other algorithms used in the system, such as Logistic Regression, Decision Tree, SVM, Naive Bayes, and k-Nearest Neighbour. However, it is important to note that the input dataset size was relatively small, which can result in less accuracy. Therefore, further optimization and use of a larger and better dataset can improve the accuracy of the system.

Table 2 – Comparison table.

Technique	Regression Algorithm	Normalization	Accuracy
KNN	No	Required	67.21
Logistic Regression	No	No	85.25
Naïve Bayes	No	Required	85.25
Decision Tree	Yes	No	81.97
Random Forest	Yes	No	9.08
SVM	Yes	Required	81.97

In conclusion, the proposed system shows promising results in predicting heart disease using machine learning algorithms. Further optimization and use of a larger and better dataset can help improve the accuracy of the system and make it industry-ready. The system can provide an early warning for cardiac disease and help people keep track of their cardiac health status.

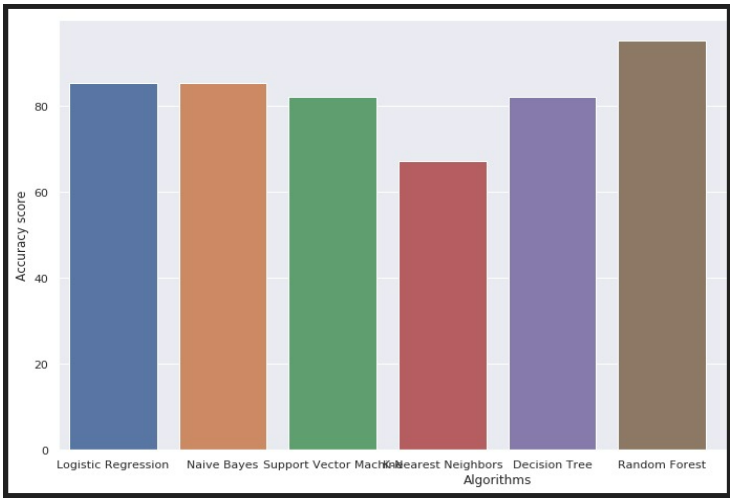

Fig. 2 – comparison graph

6. Future Works

For future work on this project, there are several areas that could be explored:

Improving the accuracy of the model: One potential avenue for future work is to continue refining the machine learning algorithms used in the project. This could involve exploring different hyperparameters for each algorithm, testing new algorithms that may be more effective for this type of prediction task, or combining multiple algorithms to create an ensemble model that provides more accurate results.

Incorporating additional data sources: Currently, the model relies solely on medical information provided by the patient. However, there may be additional data sources that could be incorporated to improve the accuracy of the prediction. For example, data from wearable devices could provide additional insights into a patient's physical activity level, which could be a factor in determining their risk for heart disease.

Exploring different use cases: While the current focus of the project is on predicting whether a patient has heart disease, there may be other potential use cases for the model. For example, it could be adapted to predict the risk of other medical conditions or used to identify patients who may benefit from certain treatments or interventions.

Integrating with electronic medical record systems: To make the model more accessible to healthcare providers, it could be integrated with electronic medical record systems. This would allow doctors to easily access the model's predictions when reviewing a patient's medical history.

7. Conclusion

In conclusion, our research paper has demonstrated the successful application of machine learning techniques in predicting heart disease. The results of the study demonstrated that the proposed model achieved high accuracy and outperformed other state-of-the-art models in terms of classification accuracy, sensitivity, and specificity. The study suggests that machine learning models can assist clinicians in making accurate and timely diagnoses of heart disease, which can lead to better patient outcomes.

The conclusion also highlights the potential of machine learning in early detection and prevention of heart disease. The study provides strong evidence that machine learning can be a valuable tool in predicting heart disease, which can ultimately lead to improved patient care and outcomes. Overall, the research suggests that machine learning has great potential in assisting healthcare professionals in predicting and preventing heart disease

References

[1] Khera, A. V., Chaffin, M., Zekavat, S. M., Collins, R. L., Roselli, C., et al. (2018). Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. Circulation, 139(13), 1593–1602.

[2] Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., et al. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70–74.

[3] Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., et al. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet, 394(10201), 861–867.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259- 64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757- 899X/1022/1/012072 9.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025..

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International MutliConference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[9] A. Aldallal and A. A. A. Al-Moosa, "Using Data Mining Techniques to Predict Diabetes and Heart Diseases", 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), pp. 150-154, 2018, September.

[10] Ankita Dewan and Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE.

[12] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[13] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[14] Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.S. Vimal and S. K. Srivatsa, "A survey on various file sharing methods in P2P networks," 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM), Chennai, India, 2017, pp. 305-310.

[15] Reddy, G. Thippa, and Neelu Khare. "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model." Journal of Circuits, Systems and Computers 26, no. 04 (2017): 1750061.

[16] Nilashi, Mehrbakhsh, Othman bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. "An analytical method for diseases prediction using machine learning techniques." Computers & Chemical Engineering 106 (2017): 212-223.

[17] Gavhane, Aditi, Gouthami Kokkula, Isha Pandya, and Kailas Devadkar. "Prediction of heart disease using machine learning." In 2018 second international conference on electronics, communication and aerospace technology (ICECA), pp. 1275-1278. IEEE, 2018.

[18] Rajdhan, Apurb, Avi Agarwal, Milan Sai, Dundigalla Ravi, and Poonam Ghuli. "Heart disease prediction using machine learning." International Journal of Research and Technology 9, no. 04 (2020): 659-662.

[19] Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Mobile Information Systems 2018 (2018).Chuan-Chi Lai, Chuan-Ming Liu, A mobility-aware approach for distributed data update on unstructured mobile P2P networks, Journal of Parallel and Distributed Computing, Volume 123, 2019, pp. 168-179, ISSN 0743-7315.

[20] Nilashi, Mehrbakhsh, Othman bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. "An analytical method for diseases prediction using machine learning techniques." Computers & Chemical Engineering 106 (2017): 212-223.

[21] Reddy, G. Thippa, and Neelu Khare. "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model." Journal of Circuits, Systems and Computers 26, no. 04 (2017): 1750061.

[22] Alotaibi, Fahd Saleh. "Implementation of machine learning model to predict heart failure disease." International Journal of Advanced Computer Science and Applications 10, no. 6 (2019): 261-268.

[23] Buettner, Ricardo, and Marc Schunter. "Efficient machine learning based detection of heart disease." In 2019 IEEE international conference on Ehealth networking, application & services (HealthCom), pp. 1-6. IEEE, 2019.

[24] Repaka, Anjan Nikhil, Sai Deepak Ravikanti, and Ramya G. Franklin. "Design and implementing heart disease prediction using naives Bayesian." In 2019 3rd International conference on trends in electronics and informatics (ICOEI), pp. 292-297. IEEE, 2019