IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning

[1]Archana Agarwal, [2] Ravi Singh Bajetha, [3]Saurav Dhyani, [4]Anmol Pal, [5]Akshit Bhatt
[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student
Dept. of Computer Science and Engineering
Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India

Abstract: Heart Disease poses a major risk to humans, and late diagnosis often affects a patient's chances of survival. Factors such as age, gender, cholesterol levels, diabetes and heart rate can affect heart problems. However, due to the many variables, it is difficult for professionals to evaluate each patient individually. The medical industry produces a lot of information about patients and diseases every day. But doctors do not use this information well. Predicting heart disease requires complex and large amounts of data that cannot be processed and analyzed with modern tools. Today healthcare industry is rich in data but not using the data efficiently. There are many machine learning techniques and tools that can be used to extract useful information from data and use this data for accurate diagnosis and decision making. This study provides a solution that uses machine learning to predict heart disease in patients. "Heart Disease Prediction System" is based on predictive modeling and uses five algorithms to predict the occurrence of diseases based on user symptoms: KNN, Random Forest, Naive Bayes, Logistic regression, and decision tree classification algorithms. This cardiovascular disease prediction can improve health and reduce costs. This study highlights the importance of predictive models in measuring and reducing cardiovascular disease in the population.

***Index Terms* -** Heart Disease, Machine Learning, Predictive Modeling, KNN, Random Forest, Naïve Bayes, Logistic Regression, Decision Tree Classification, Cardiovascular Disease (CVD), Cardiovascular Disease (CVD)

I. INTRODUCTION

According to the World Health Organization, 18 million people worldwide die from heart disease every year. Heart disease is one of the leading causes of death in the world population. An estimated 17.9 million people died of CVD in 2019, accounting for 32% of all deaths worldwide. Of these deaths, 85% were caused by heart attack and stroke. More than three-thirds of deaths from heart disease occur in low- and middle-income countries. In 2016, India recorded that 63% of all deaths were caused by non-communicable diseases (NCDs), with 27% specifically linked to cardiovascular diseases (CVD). CVDs were responsible for 45% of deaths among individuals aged 40-69. Increasing age Smoking, diabetes, high cholesterol, high blood pressure, lack of physical activity, overweight, work stress and poor eating habits are the major causes of the heart diseases. Even young adults are affected, probably due to bad habits, lack of sleep and genetic factors. Diagnosing heart disease is complex and involves manual checks taking into account various factors. Cardiovascular disease prediction is considered one of the most important subjects in the data analysis section. The burden of cardiovascular disease has been increasing rapidly worldwide in the last few years. Much research has been done to determine the most influential factors in heart disease and to accurately predict overall risk. Medical organizations around the world collect data on various health-related issues. This data can be leveraged using various machine learning techniques to extract useful information. However, the data collected is very massive and often this data can be very noisy. These data sets, which are too overwhelming for the human mind to comprehend, can be easily explored using various machine learning techniques. These algorithms have recently become very useful for accurately predicting the presence or absence of heart-related diseases. Early diagnosis of heart disease plays a vital role in making decisions about lifestyle changes in high-risk patients and subsequently reduces complications. Machine learning has proven to be effective in making decisions and making predictions from the large amount of data produced by the healthcare industry. This project aims to predict future heart disease by analyzing patient data to classify whether they have had heart disease or not. By collecting data from various sources, classifying them under appropriate names and finally analyzing them to extract the required data, we can say that this technique can be very well adapted to predict heart diseases. The aim of this project is to verify whether a patient is likely to be diagnosed with any cardiovascular heart disease based on their medical attributes such as gender, age, chest pain, fasting blood sugar, etc. The dataset was selected from the UCI repository and included the patient's medical history and characteristics. We use this data set to predict whether a patient may have heart disease or not. We predict the likelihood of heart disease based on 14 health factors in a patient. If the analysis classifies these factors, we can say whether a person is likely to have heart disease. These medical attributes are trained according to five different algorithms: Logistic Regression, Naïve Bayes, KNN, Random Forest, Decision Tree Classifier. And finally, we classify patients who are at risk of heart disease or not, and also this method is completely cost-effective.

II. Dataset Description

The dataset utilized in this project comprises 14 different variables. The independent variable to be predicted is the "diagnosis," which determines whether a person is healthy or has heart disease.

Study Information:

- Age: Age of the patient in years.
- Gender: Patient's gender (1 = M; 0 = F).
- Chest Pain: Type of pain experienced by the patient on pressure, categorized as 1 = typical angina; 2 = atypical angina; 3 = non-anginal; 4 = asymptomatic.
- Blood Pressure: Resting blood pressure of the patient in mmHg at admission.
- Serum Cholesterol: Serum cholesterol level of the patient in mg/dl.
- Fasting Blood Glucose: Fasting blood glucose level of the patient, categorized as 1 = greater than 120mg/dl; 0 = less than or equal to 120mg/dl.
- ECG: Resting ECG results of the patient, categorized as 0 = normal; 1 = ST-T; 2 = hypertrophy.
- Maximum Heart Rate: Maximum heart rate achieved by the patient.
- Induced Angina: Whether the patient experienced induced angina caused by exercise, categorized as 1 = yes; 0 = no.
- ST Depression: ST depression caused by exercise relative to rest.
- Slope: Slope of the ST segment at peak exercise, categorized as 1 = uphill; 2 = flat; 3 = downward slope.
- Number of vessels: Number of large vessels stained by fluoroscopy (0-3).
- Thalassemia: Thalassemia status of the patient, categorized as 3 = normal; 6 = repair defect; 7 = reversible defect.
- Diagnosis: Predictive features indicating cardiac diagnosis (angiographic disease state), with value 0 indicating less than 50% diameter reduction and value 1 indicating greater than 50% diameter reduction.

III. Literature Review

[1] In 2018, Sanchayita Dhar, Pritha Datta, Ankur Biswas, Tanusree Dey, and Krishna Roy published a paper titled "A Hybrid Machine Learning Approach for Prediction of Heart Disease." The study employed a hybrid machine learning approach utilizing Naive Bayes, Decision Tree, and Random Forest algorithms. The goal was to develop a prediction system capable of envisioning heart disease based on measurements extracted from the ERIC laboratory dataset, which consisted of 209 test cases [Sanchayita Dhar et al., 2018].

[2] In 2019, a paper titled 'Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques' was published by Senthilkumar Mohan, Chandrasekar Thirumalai, and Gautam Srivastava. The paper introduced HRFLM, a hybrid machine learning method that integrates Random Forest (RF) and Linear Method (LM) to predict heart disease. The study demonstrated an accuracy of 88.7% on the Cleveland heart disease dataset, which was higher than previous methods [Senthilkumar Mohan et al., 2019].

[3] In 2020, Rishabh Magar, Rohan Memane, and Suraj Raut published a paper titled "Heart Disease Prediction Using Machine Learning." The heart disease prediction system presented in the study evaluates the risk of disease for a patient based on the accuracy of a particular trained model using user-given data values. The system incorporates four algorithms: Support Vector Machine (SVM), Decision Tree, Naïve Bayes, and Logistic Regression. The report details the implementation of these algorithms, highlighting that the Logistic Regression algorithm exhibited the highest efficiency, achieving an accuracy of 82.89% [Rishabh Magar et al., 2020].

[4] In 2020, Abhay Agrahara published a paper titled "Heart Disease Prediction Using Machine Learning Algorithms." The study employed machine learning methods, specifically Logistic Regression and Decision Trees, for predicting heart disease. The effectiveness of these methods was assessed using metrics such as accuracy, precision, recall, and F-1 score. The results indicated that the Decision Tree Classifier achieved the highest accuracy in predicting heart 7 disease. The paper emphasized the significance of diverse datasets in healthcare applications [Abhay Agrahara, 2020].

[5] In 2021, Armin Yazdani, Kasturi Dewi Varathan, and Asad Waqar Malik published a paper titled "Machine learning–based heart disease prediction system for Indian population" in which they aimed to predict heart disease using various algorithms, including k-Nearest Neighbors (kNN), Naïve Bayes (NB), Logistic Regression (LR), AdaBoost (AB), and Random Forest (RF). The study reported a diagnostic accuracy of 93.8% on a validation set using the Random Forest (RF) algorithm [Armin Yazdani et al., 2021].

[6] In 2021, Harshit Jindal published a paper titled "Heart Disease Prediction Using Machine Learning Algorithms." The study employed logistic regression and k-Nearest Neighbors (KNN) algorithms on a heart disease dataset for prediction. The report revealed an accuracy of 87.5% in detecting cardiovascular disease using the data. Notably, the KNN algorithm outperformed other algorithms, achieving an accuracy of 88.52%. The paper recommended incorporating more data for improved accuracy and highlighted the importance of data cleaning to address dataset issues [Harshit Jindal, 2021].

[7] In 2021, Baban U. Rindhe, Nikita Ahire, Rupali Patil, and Shweta Gagare published a paper titled "Heart Disease Prediction Using Machine Learning." The study utilized Support Vector Classifier, Neural Network, and Random Forest Classifier on a heart disease dataset for prediction. The project involved an analysis of a heart disease patient dataset with thorough data processing. Three different models were trained and tested, resulting in the following maximum accuracy scores: Support Vector Classifier: 84.0%, Neural Network: 83.5%, Random Forest Classifier: 80.0% [Baban U. Rindhe et al., 2021].

[8] In 2022, Apurb Milan SaiDundigalla and Dr. Poonam Ghuli published a paper titled "Heart Disease Prediction using Machine Learning." The study employed Naive Bayes, Decision Tree, Logistic Regression, and Random Forest algorithms on a heart disease dataset for prediction. The results indicated that the Random Forest algorithm was the most accurate among the tested algorithms for predicting heart disease [Apurb Milan Sai Dundigalla et al., 2022].

IV. Proposed Methodologies

1) KNN (K-Nearest Neighbors)

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm for classification and reconstruction. It is a type of model-based learning, also known as lazy learning, in which the model does not learn specifically in training. Instead, it remembers the training data and makes predictions based on the similarity between the new data and existing examples. We can implement the KNN model by following the steps below:

1. Load the data

2. initial value of k is

3. To obtain the predicted class, iterate from 1 to all points of the data set, calculating the distance between the test data and each line of information. Here we will use Euclidean distance as the distance measure because it is the most popular method. Another measure that can be used is Chebyshev, cosine, etc. Separate the distance according to the order of movement according to the distance value b. Get the first k rows from the sorted array c. Get the highest rank in this line d. Back to category predictions

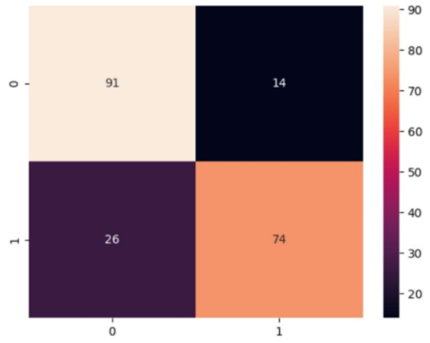

Figure 1: Confusion Matrix of KNN

2) Random Forest Classifier

Random Forest Classifier is a powerful and versatile machine learning algorithm that can be used for classification and propagation. They essentially determine the forest of trees where each tree makes predictions based on the rules. The final prediction of the random forest is made by voting the majority (for classification) or average (for regression) of the predictions of all trees in the forest, Random Forest pseudo code:

1. Randomly select “k” features among all “m” features. Here k <<m
2. Among the “k” features, use the best split point to calculate node “d”
3. Split the node into child nodes using consensus splitting.
4. Repeat steps 1 to 3 until nodes "l" are reached.
5. Create a forest by repeating steps 1 through 4 “n” times to create “n” trees.

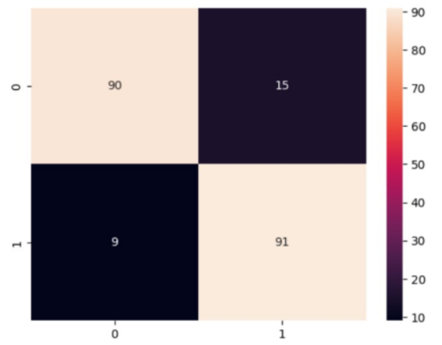

Figure 2: Confusion Matrix of Random Forest

3) Logistic Regression

Logistic Regression is a classification algorithm widely used in binary classification problems. In logistic regression, the logistic regression algorithm uses a logistic function to compress the output of a linear equation between 0 and 1, rather than fitting a straight line or hyperplane. There are 13 different methods that make logistic regression suitable for distribution are independent.

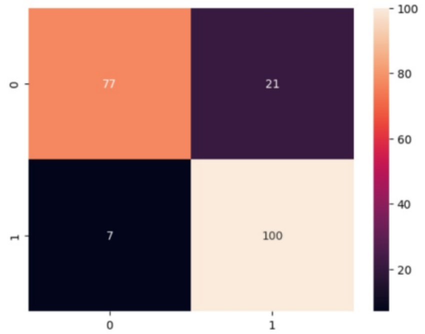

Figure 3: Confusion Matrix of Logistic Regression

4) Naïve Bayes

Naive Bayes is a family of probabilistic algorithms that use probability theory and Bayes' theorem to predict the tag of text (such as a newspaper or customer review). They are based on probability; that is, they calculate the probability of each tag given in the text and then display the highest possible text. The method for obtaining these results is to use Bayes' theorem, which describes the probability of a property based on prior knowledge of the events that will be associated with that property. Bayes theorem is useful when dealing with probability and for this we use the following formula:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

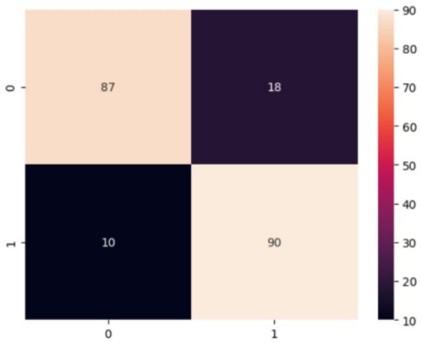

Figure 4: Confusion Matrix of Naïve Bayes

5) Decision Tree

Decision trees are a valuable decision support tool that employ a tree-like model consisting of nodes and branches to represent a sequence of tests performed on a given record or object. The input of the decision tree is an object described by a set of attributes, and the output is a decision with a predicted output value for that input. The attributes can be either discrete or continuous. As the decision tree progresses, it performs a sequence of tests and eventually reaches a decision. Each non-leaf node in the tree represents a test for a relevant attribute value and the branches from the node are labelled with the possible outcomes of the test, Finally, each leaf node specifies the value or decision to be returned if that leaf is reached. There are several decision tree implementation algorithms, including J48, Random Forest and Logistic Tree Model.

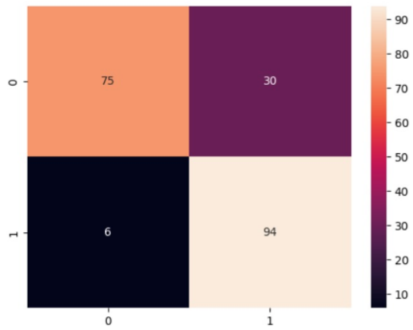

Figure 5: Confusion Matrix of Decision Tree

V. Result and Analysis

This section shows the results obtained by applying Random Forest, Decision Tree, Naive Bayes and Logistic Regression, KNN. The metrics used for performance evaluation are precision, precision (P), regression (R), and F measure. Precision provides a measure of accuracy. Recall defines the true measure of quality. F test accuracy.

Sensitivity = (TP) / (TP + FP)

Return = (TP) / (TP + FN)

F– Measurement = (2 * Degree of Sensitivity * Recall) / (Sensitivity + Recall)

- TP True Positive: The patient has the disease and also the test is positive.
- FP False Positive: The patient does not have the virus but the test is positive.
- TN True Negative: The patient has no disease and the test is negative.
- FN False Negative: The patient has the virus but the test is negative.

Algorithm	True Positive	False Positive	False Negative	True Negative
KNN	74	14	26	91
Random Forest	91	15	9	90
Logistic Regression	100	21	7	77
Decision Tree	94	30	6	75
Naïve Bayes	90	18	10	87

Table I: Values obtained for confusion matrix using different algorithm analysis of machine learning algorithm

In the experiment, the pre-processed data set was used as an experiment and the above algorithm was researched and applied. The above performance measurements were obtained from confusion matrices. The confusion matrix describes the performance of the model. The proposed model focuses on the correct score obtained from the KNN method obtained from different algorithms such as conflict matrix random forest, decision tree, logistic regression and Naive Bayes classification.

Algorithm	Precision	Recall	F1-Score	Accuracy
KNN	0.906	0.728	0.808	81.95
Random Forest	0.878	0.943	0.909	90.24%
Logistic Regression	0.826	0.934	0.877	86.34%
Decision Tree	0.844	0.859	0.851	84.39%
Naïve Bayes	0.834	0.897	0.864	85.37%

Table II: Analysis of machine learning algorithm

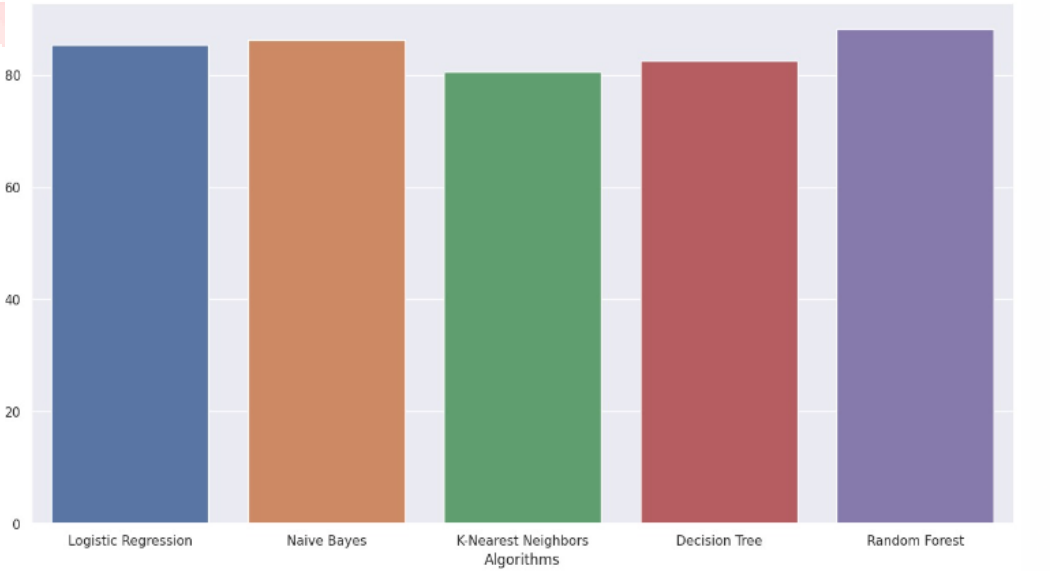

Figure 6: Graphical Representation of Algorithms

VI. Conclusion

This project aims to provide an in-depth understanding of machine learning techniques for classifying heart diseases. The role of a classifier is crucial in the healthcare industry as its results can help predict the most suitable treatment for patients. The study analyzed and compared existing techniques to identify the most efficient and effective systems. Although each of the algorithms tested performed well in some instances, they performed poorly in others. After a thorough analysis, the study concluded that the Random Forest algorithm is the most accurate and efficient, achieving an accuracy score of 90.24% in predicting heart disease.

In the future, the work can be improved by developing a web application based on the Random Forest algorithm. Additionally, using a larger dataset than the one utilized in this analysis can provide better results, which can be valuable for health professionals to predict heart disease more effectively and efficiently.

References

[1] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 70185-70194, 2019. DOI: 10.1109/ACCESS.2019.2923707.

[2] E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, "Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India," Med J Armed Forces India, vol. 77, no. 3, pp. 302-311, Jul. 2021. doi: 10.1016/j.mjafi.2020.10.013.

[3] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and Dr. P. Ghuli, "Cardiovascular disease prediction using machine learning," International Journal of Engineering Research and Technology (IJERT), vol. 9, no. 04, pp. 1-7, April 2020. DOI: 10.1016/j.mjafi.2020.10.013.

[4] A. Agrahara, "Heart Disease Prediction Using Machine Learning Algorithms," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 6, no. 4, pp. 137-149, July-August 2020.

[5] R. Magar, R. Memane, S. Raut, and V. S. Rupnar, "Heart Disease Prediction Using Machine Learning," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 7, no. 6, pp. 2081, June 2020.

[6] B. U. Rindhe, N. Ahire, R. Patil, S. Gagare, and M. Darade, "Heart Disease Prediction Using Machine Learning," *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 5, no. 1, pp. 267, May 2021.

[7] Shinde R, Arjun S, Patil P, & Waghmare J's (2015) research presents a system designed to forecast heart disease by utilizing the Naïve Bayes algorithm and K-Means clustering. This study was published in the International Journal of Computer Science and Information Technologies, 6(1), 637-9

[8] A. Ganna, P. K. Magnusson, N. L. Pedersen, U. de Faire, M. Reilly, J. Ärnlöv, and E. Ingelsson, "Multilocus genetic risk score for cardiovascular disease prediction," Arteriosclerosis, Thrombosis, and Vascular Biology, vol. 33, no. 9, pp. 2267-2272, 2013

[9] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 18-19 December 2020, doi: 10.1109/ICACCCN51052.2020.9362842

[10] R. Katarya thiab S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis," Health and Technology, vol. 11. Tsis muaj. 1 second. 87–97, November 2020, doi: 10.1007/s12553-020-00454-z.

[11] D. Bertsimas, L. Mingardi, and B. Stellato, "Machine learning for real-time heart disease prediction," in IEEE Journal of Biomedical and Health Informatics, vol. 25, pp. 3627-3637, Sep. 2021, doi: 10.1109/JBHI.2021.3066347.

[12] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, and R. Sai Suraj, "Heart Disease Prediction using Hybrid Machine Learning Model," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), 20-22 January 2021, doi: 10.1109/ICICT50816.2021.9358597.

[13] S. I. Ayon, M. M. Islam, and M. R. Hossain, authored a paper titled "Coronary roadway heart disease prediction: A study of computational intelligence approaches," which was published in the *International Journal of System Assurance Engineering and Management*, vol. 11, no. 6, pp. 2488-2507, Nov.-Dec.

2020. The paper discusses the use of computational intelligence methods for predicting heart disease. It can be accessed online at doi: 10.1080/03772063.2020.1713916.

[14] S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," in 2020 International Conference for Emerging Technology (INCET), 05-07 June 2020, doi: 10.1109/INCET49848.2020.9154130.

[15] K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, and T. Suryawanshi, "Human Disease Prediction Using Machine Learning Techniques and Real-life Parameters," Department of Computer Science and Engineering, Bharati Vidyapeeth Deemed to be University College of Engineering, Pune, India, vol. 36, no. 06c, pp. 07, 2023. doi: 10.5829/ije.2023.36.06c.07.

[16] Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. BMC Bioinformatics, 21(278), 1-14. doi. :10.1186/s12859-020-03617-7

[17] Gárate- Escamila,A.K., Hajjam El Hassani,A., & Andrès,E.(2020). Bracket models for heart complaint vaticination using point selection and PCA. Informatics in Medicine uncorked, 19, 100330. doi : 10.1016/j.imu.2020.100330

[18] Bharti,R., Khamparia,A., Shabaz,M., Dhiman,G., Pande,S., & Singh,P.(2021). vaticination of heart complaint using a combination of machine learning and deep learning. Journal Name, Volume(Number), Article ID 8387680. doi : 10.1155/2021/8387680

[19] Tougui,I., Jilbab,A., & El Mhamdi,J.(2020). Heart complaint bracket using data mining tools and machine literacy ways. Health and Technology, 10(1137 – 1144). doi :10.1007/s12553-020-00435-2

[20] M. J. Nayeem, S. Rana, and M. R. Islam authored a paper titled "Heart Disease Prediction Using Machine Learning Algorithms," which appeared in the European Journal of Artificial Intelligence and Machine Learning, volume 1, issue 3, pages 23-26, in 2022. The DOI for this article is 10.24018/ejai.2022.1.3.13.

[21] R. Spencer, F. Thabtah, and M. Thompson, authored a paper titled "Exploring feature selection and classification methods for predicting heart disease," which was published in the IEEE Transactions on Biomedical Engineering. This article, found in volume 67, issue 12, pages 2795-2805, was published in December 2020 and can be accessed online with doi : 10.1109/TBME.2020.3005756.

[22] J. Kaur and B. S. Khera, "Hybrid and fuzzy logic approaches for heart disease detection risk: A state-of-the-art review," Journal of the Institute of Engineering: Series B, 2021. doi: 10.1007/s40031-021-00644-3.

[23] A. Kumar, K. Rathor, S. Vaddi, D. Patel, P. Vanjarapu, and M. Maddi, "Electrocardiogram Based Early Heart Attack Prediction Using Neural Networks," in 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), 17-19 August 2022, doi: 10.1109/ICESC54411.2022.9885448.

[24] H. A. Al-Alshaikh, P. P. Poonia, R. C. et al., "Evaluation and performance analysis of machine learning model in heart disease prediction," Scientific Reports, vol. 14, p. 7819, 2024. doi: 10.1038/s41598-024-58489-7.

[25] M. Pal and S. Parija, "Prediction of heart diseases using Random Forest," Journal of Physics: Conference Series, vol. 1817, p. 012009, 2021. doi: 10.1088/1742-6596/1817/1/012009.

[26] M. A. Khan "An IoT Framework for Heart Disease Prediction Grounded on MDCNN Classifier" IEEE Access, vol. 8, pp. 34717- 34727, 2020. doi : 10.1109/ACCESS.2020.2974687.