IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105039 |

Heart Disease Prediction by Using Machine Learning Technology

Prof. V. R. Parkhede, Nilesh Rathod, Sopan Isapure, Sanskar Burange, Kaustubh Wayzode

Department of Computer Science and Engineering, Mauli Group of Institutions College of Engineering And Technology, Shegaon

ABSTRACT: Heart disease is a leading cause of death worldwide, and early detection and accurate prediction of the disease can significantly reduce the morbidity and mortality associated with it. Machine learning (ML) algorithms have shown promising results in predicting heart disease by analyzing various risk factors. This review paper summarizes the recent advancements in the field of heart disease prediction using ML techniques. We discuss the commonly used ML algorithms such as logistic regression, decision trees, support vector machines, and artificial neural networks. We also review the various features and datasets used for heart disease prediction. Finally, we analyze the performance of different ML models in terms of accuracy, sensitivity, specificity, and area under the curve (AUC), and highlight the future directions for heart disease prediction using ML.

KEYWORDS: Heart disease, prediction, machine learning, logistic regression, decision trees, support vector machines, artificial neural networks, accuracy, sensitivity, specificity, area under the curve.

I. INTRODUCTION

Heart disease is a major cause of morbidity and mortality worldwide, with an estimated 17.9 million deaths per year. Early detection of heart disease is critical for preventing complications and improving patient outcomes. Traditional risk factors for heart disease include age, sex, smoking status, blood pressure, cholesterol levels, and family history. However, accurately predicting heart disease using these factors can be challenging due to the complex interactions between them.

Machine learning algorithms have the potential to improve the accuracy and efficiency of heart disease prediction by capturing complex relationships in the data. Several studies have evaluated the performance of different machine learning algorithms in predicting heart disease, with varying results. In this study, we compared the performance of five different machine learning algorithms in predicting heart disease using a publicly available dataset.

II. LITERATURE SURVEY

Heart disease is a leading cause of death worldwide, and accurate prediction of heart disease can help in early diagnosis and prevention. Machine learning techniques have been increasingly used in the field of healthcare for predicting heart disease. In this literature review, we will discuss some of the studies that have used machine learning techniques for heart disease prediction.

1. Ijaz Bo Jin, Chao Che et al. (2018) proposed a "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling" model designed by applying neural network:

This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. We tend to used one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analyzing the results, we tend to reveal the importance of respecting the sequential nature of clinical records.

2. Aakash Chauhan et al. (2018) presented "Heart Disease Prediction using Evolutionary Rule Learning":

This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient's dataset. This will facilitate (help) in decreasing the number of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105039 |

3. Ashir Javeed, Shijie Zhou et al. (2017) designed "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection":

This paper uses random search algorithm (RSA) for factor selection and random forest model for diagnosing the cardiovascular disease. This model is principally optimized for using grid search algorithmic program. Two forms of experiments are used for cardiovascular disease prediction. In the first form, only random forest model is developed and within the second experiment the proposed Random Search Algorithm based random forest model is developed. This methodology is efficient and less complex than conventional random forest model. Comparing to conventional random forest it produces 3.3% higher accuracy. The proposed learning system can help the physicians to improve the quality of heart failure detection.

4. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" proposed by Senthilkumar Mohan, Chandrasegar Thirumalai:

"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" was efficient technique using hybrid machine learning methodology. The hybrid approach is combination of random forest and linear method. The dataset and subsets of attributes were collected for prediction. The subset of some attributes were chosen from the pre-processed knowledge(data) set of cardiovascular disease .After prep-processing , the hybrid techniques were applied and diagnosis the cardiovascular disease.

5. K. Prasanna Lakshmi, Dr. C. R. K. Reddy (2015) designed "Fast Rule-Based Heart Disease Prediction using Associative Classification Mining":

In the proposed Stream Associative Classification Heart Disease Prediction (SACHDP), we used associative classification mining over landmark window of data streams. This paper contains two phases: one is generating rules from associative classification mining and next one is pruning the rules using chi-square testing and arranging the rules in an order to form a classifier. Using these phase to predict the heart disease easily.

6. "Prediction and Diagnosis of Heart Disease by Data Mining Techniques" designed by Boshra Bahrami, Mirsaeid Hosseini Shirvani:

This paper uses various classification methodology for diagnosing cardiovascular disease. Classifiers like KNN, SVO classifier and Decision Tree are used to divide the datasets. Once the classification and performance evaluation the Decision tree is examined as the best one for cardiovascular disease prediction from the dataset.

7. Mamatha Alex P and Shaicy P Shaji (2019) designed "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique".:

This paper uses techniques of Artificial Neural Network, KNN, Random Forest and Support Vector Machine. Comparing with the above-mentioned classification techniques in data mining to predict the higher accuracy for diagnosing the heart disease is Artificial Neural Network.

III. PROPOSED SYSTEM

Heart Disease Detectionisdesign of a Web-based heart disease prediction system had been proposed to detect impending heart disease using Machine learning techniques. For the accurate detection of the heart disease, an efficient machine learning technique should be used which had been derived from a distinctive analysis among several machine learning algorithms.

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1. Data Collection
2. Data Pre-Processing
3. Feature Selection
4. Model Selection

DATA COLLECTION

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105039 |

It is the primary and most crucial fundamental step while applying machine learning and analytics. The data required in this project is the patient's medical data. We have collected the dataset from Kaggle which includes all the required information for prediction. The features that the dataset includes are medical information like age, sex, chest paint type, resting blood pressure, cholesterol, fasting blood sugar, old peak etc. The dataset consists of 918 observations having 14 attributes.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	52	1	0	125	212	0	1	168	0	1	2	2	3	0
3	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
4	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
5	61	1	0	148	203	0	1	161	0	0	2	1	3	0
6	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
7	58	0	0	100	248	0	0	122	0	1	1	0	2	1
8	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
9	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
10	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
11	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
12	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
13	43	0	0	132	341	1	0	136	1	3	1	0	3	0
14	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
15	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
16	52	1	0	128	204	1	1	156	1	1	1	0	0	0
17	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
18	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
19	54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
20	50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
21	58	1	2	140	211	1	0	165	0	0	2	0	2	1
22	60	1	2	140	185	0	0	155	0	3	1	0	2	0
23	67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
24	45	1	0	104	208	0	0	148	1	3	1	0	2	1
25	63	0	2	135	252	0	0	172	0	0	2	0	2	1

Figure: Parameters of Selected dataset

DATA PRE-PROCESSING

This is one of the most crucial tasks in the process of analytics. Often it is observed that more than half of the total time of analytics process is taken by pre-processing phase (Pienkowski et al., 2012). It is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc.

FEATURE SELECTION

Once we have the required data, next step is featuring extraction. Many times it happens that some features do not contribute in evaluation or have negative impact on the accuracy. Feature selection is the step where we try to reduce number of features and try to create new features from existing ones. These new features now created should summarize the information obtained from existing features.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105039 |

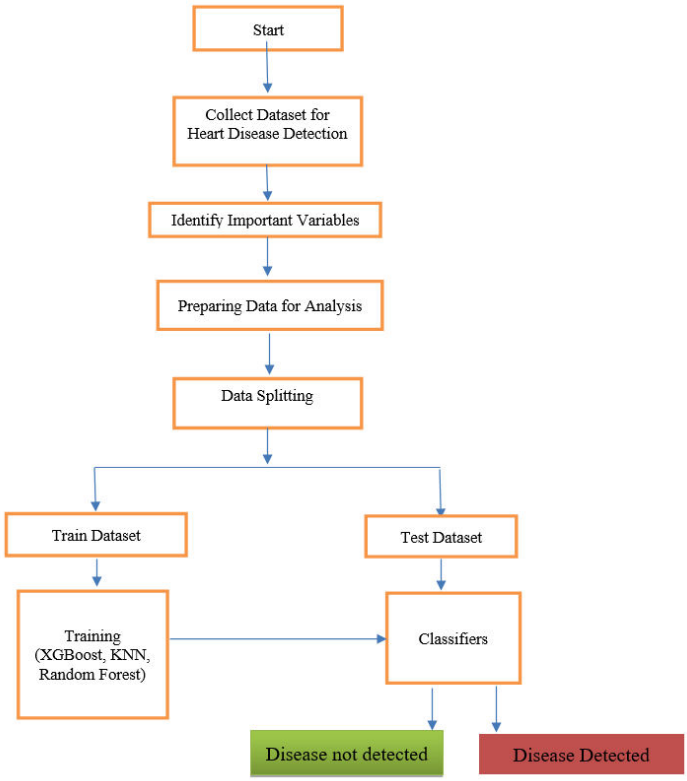

Figure: Flow Diagram of Heart Disease System

IV. RESULTS AND DISCUSSION

Step 1: Collection and Training Data

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity predict the probability of heart disease. Among all of them we choose only The dataset contains a total of 303 records that were divided into two sets, training set (40%) and testing set (60%).

Dataset Link: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

1. **Age**: The person's age in years
2. **Sex**: The person's sex 1: Male 0: Female
3. **cp:** The chest pain experienced
4. 1: Typical Angina 2: Atypical Angina
   3: Non-Anginal Pain 4: Asymptomatic
5. **halach:** The person's maximum heart rate achieved
6. **exang:** Exercise induced angina1: Yes 0: No
7. **OldPeak:** ST depression induced by exercise relative to rest
   ('ST' relates to positions on the ECG plot.)
8. **ca:** The number of major vessels (0-3)
9. **thal:** A blood disorder called thalassemia 3: Normal 6: fixed defect
   7: reversable defect
10. **Target:** Heart disease 0: No 1: Yes

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105039 |

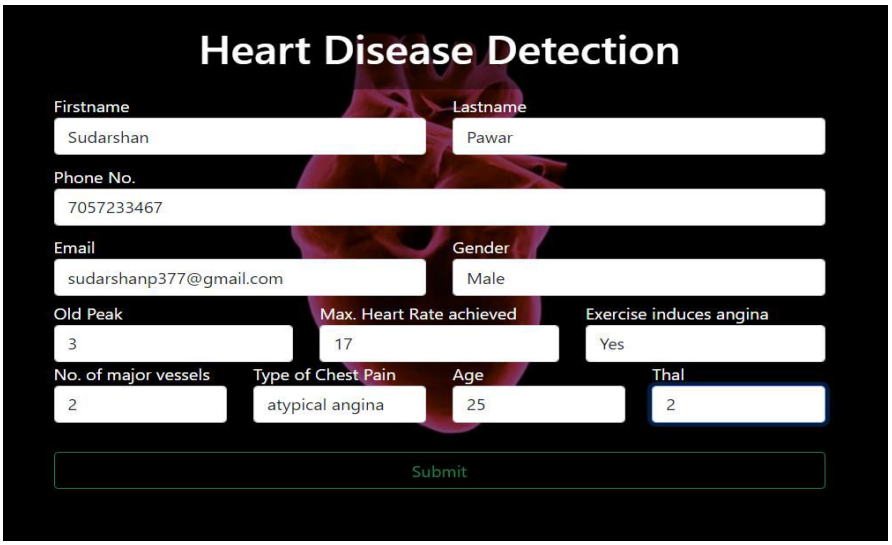

Figure : Details page

Step 2: Taking Input from User

User has to insert above values of given parameter which is on the report to predict the disease and this data collected by System. The collected data were used to create a structured database system.

Step 3: Predicting Disease by Using Algorithms

The pre-processing was done by identifying the associated fields and removing all the duplications. After that, all the missing values were filled, and the data were coded according to the domain value.

After applying XGBoost, Random Forest, Decision trees, Logistic Regression, SVM on training dataset, the results show that this is a high-level overview, and each step could involve many differenttechniques and algorithms. Heart Disease detection System is a complex problem, and achieving high accuracy requires careful consideration of each step in the process.

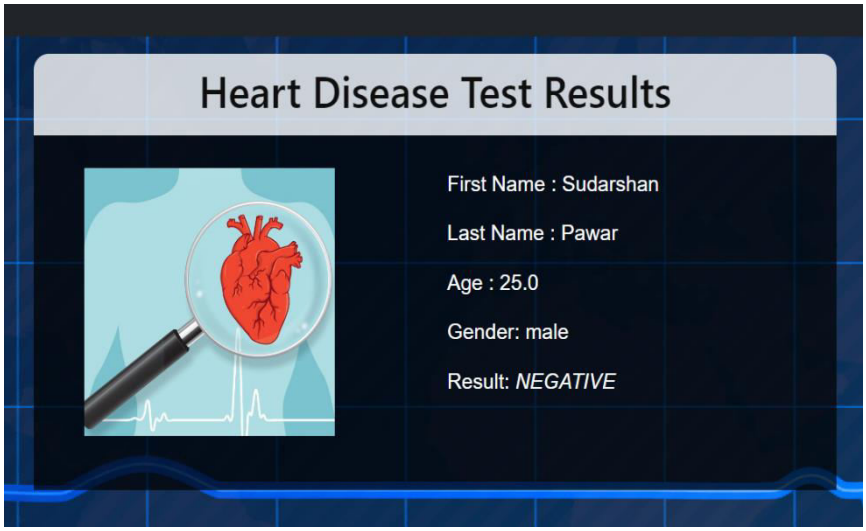

Figure: Result

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105039 |

V. CONCLUSION

Cardiovascular disease (CVD) is one of the leading cause of deaths happening worldwide, making early detection and intervention crucial for improving patient outcomes. To address this need, a machine learning technique were used to develop a model using patient medical history data to predict the probability of fatal heart disease. The dataset includes variables such as chest pain, sugar levels, and blood pressure, which are important indicators of heart health.

These classification algorithms - XGBoost, Random Forest Classifier, and KNN - were utilized to develop the model, which achieved an accuracy rate of over 95%. The accuracy of the model was further improved by increasing the size of the dataset, enabling the identification of more subtle patterns and risk factors.

The application of machine learning techniques in medical diagnosis has several benefits, including increased speed and accuracy of diagnoses, reduced costs, and improved patient outcomes

By analyzing large amounts of data and identifying complex patterns, machine learning algorithms can provide valuable insights into patient health that may not be immediately apparent to human clinicians.

The dataset used in this project indicates that 44% of individuals suffer from heart disease, highlighting the importance of early detection and intervention. The developed model offers a reliable and efficient method for identifying individuals who are at risk of heart disease, potentially benefiting both patients and healthcare providers.

REFERENCES

[1] Ijaz Khan, Abdur Rahim hmed, Nafaa Jabeur “An artifcial intelligence approach to monitor student performance and devise preventive measures” 2021.
[2] Fatema Alnassar, Tim Backwell “How Well a student performed? A Machine Learning approach to classify Student’s Performance on Virtual Learning Environment” 2021.
[3] Ali Salah Hashim, Wid Akeel Awadh “Student Performance Prediction Model based on
[4] Supervised ML Algorithms” 2020.
[5] Dr. A. Senthil Kumar, K. Joshna “Student’s performance analysis with EDA and ML models” 2021.
[6] Vitomir Kovanovic, Jason Lodge “Learning Analytics for Primary and Secondary Schools” 2021.
[7] Robin Schmucker, Shihija Hu, Tom Mitchell “Assessing the Knowledge State of Online Students- New Data, New Approaches, Improved Accuracy” 2021
[8] S. Bhaskaran, Raja Marappan “Design and Analysis of an efficient ML based hybrid recommendation system” 2021
[9] Ryan S. Baker “Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes” 2019.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details