JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Novel Approach for Prediction of Heart Disease Using Machine Learning Techniques

[1]Nisha M. Vadodariya
[1]Assistant Professor
[1]Computer Engineering Department,
[1]Atmiya University, Rajkot, India

Abstract : Heart disease remains one of the leading causes of mortality worldwide. Early and accurate prediction of heart disease can significantly improve patient outcomes. In this paper, we propose a novel approach using machine learning techniques to enhance the accuracy of heart disease prediction. The study compares various machine learning algorithms, including logistic regression, decision trees, random forests, and deep learning models, to determine the most effective method for early diagnosis. The proposed model is evaluated using publicly available datasets, and the results demonstrate improved performance in terms of accuracy, precision, and recall.

I. Introduction

Good health is one of the most essential aspects of human life, contributing significantly to overall well-being. Early detection of diseases plays a crucial role in reducing mortality rates and preventing severe health complications. Among various health concerns, heart disease is one of the most prevalent and life-threatening conditions worldwide. It is responsible for a significant number of deaths each year, making early diagnosis and prevention essential.

The heart plays a vital role in circulating blood, supplying oxygen, and delivering essential nutrients throughout the body. Maintaining heart health is crucial to preventing life-threatening conditions such as heart attacks and strokes. According to global health reports, cardiovascular diseases (CVDs) are among the leading causes of death, accounting for approximately 32% of total fatalities worldwide. In the United States alone, one in four individuals succumbs to heart-related illnesses annually, affecting both men and women alike. The World Health Organization (WHO) reported that in 2019, around 17.9 million deaths were attributed to cardiovascular diseases.

Several risk factors contribute to the development of heart disease, including high blood pressure, obesity, high cholesterol levels, diabetes, smoking, sedentary lifestyles, genetic predisposition, stress, and poor dietary habits. Environmental factors such as pollution, excessive alcohol consumption, and inadequate sleep also increase the likelihood of heart-related complications. However, many of these risks can be mitigated through lifestyle modifications, including maintaining a balanced diet, regular exercise, stress management, and routine medical check-ups.

Heart disease manifests in various forms, including coronary artery disease, heart failure, arrhythmias, stroke, and congenital heart conditions. Each type presents distinct symptoms, such as chest pain, shortness of breath, irregular heartbeats, dizziness, and swelling in the limbs. Identifying these warning signs early can lead to timely medical intervention and improved health outcomes.

Advancements in technology have paved the way for innovative approaches to diagnosing and predicting heart disease. Machine learning techniques have emerged as a powerful tool in the healthcare sector, offering efficient and accurate predictions based on medical data. By analyzing key parameters such as blood pressure, cholesterol levels, and lifestyle habits, machine learning models can help detect heart disease at an early stage, enabling preventive measures to be taken promptly.

This research explores a novel machine learning-based approach for heart disease prediction. By leveraging advanced computational techniques, this study aims to enhance the accuracy and reliability of diagnosing cardiovascular conditions, ultimately contributing to better healthcare solutions and improved patient outcomes.

II. Problem Statement

Heart disease remains one of the leading causes of mortality worldwide, posing a significant threat to human health. The complexity of cardiovascular conditions, along with multiple risk factors such as high blood pressure, diabetes, obesity, and genetic predisposition, makes early detection and prevention challenging. Traditional diagnostic methods often require extensive medical expertise, are time-consuming, and may not always provide accurate predictions.

To address this issue, there is a need for an efficient and reliable heart disease prediction system that can analyze multiple health parameters and identify potential risks at an early stage. By leveraging machine learning techniques, an advanced predictive model can be developed to assist healthcare professionals in making accurate diagnoses, improving patient outcomes, and reducing the overall burden of cardiovascular diseases.

III. Research Objectives

Early Detection of Heart Disease: Develop a system to identify heart disease at an early stage, enabling timely medical intervention.

Reduction in Mortality Rate: Contribute to lowering the global death rate associated with heart diseases through predictive analysis.

Minimizing Health Risks: Aid in preventing life-threatening situations by providing accurate risk assessments.

Exploring Machine Learning in Healthcare: Study and analyze various machine learning techniques and their applications in medical diagnosis.

Identifying Optimal Prediction Techniques: Evaluate different machine learning algorithms to determine the most effective approach for heart disease prediction.

Developing a Predictive Model: Build and implement a machine learning-based system capable of accurately predicting heart disease based on multiple health parameters.

IV. Machine Learning Techniques for Heart Disease Prediction

Heart disease remains one of the leading causes of mortality worldwide, posing a significant threat to human health. The complexity of cardiovascular conditions, along with multiple risk factors such as high blood pressure, diabetes, obesity, and genetic predisposition, makes early detection and prevention challenging. Traditional diagnostic methods often require extensive medical expertise, are time-consuming, and may not always provide accurate predictions.

Logistic Regression is one of the most widely used supervised machine learning algorithms for classification problems. It is primarily employed to predict categorical outcomes based on independent variables in a dataset. The algorithm estimates the probability that a given input belongs to a specific category, typically expressed as a binary outcome (e.g., Yes/No, True/False, or 0/1). Instead of providing exact values of 0 or 1, logistic regression outputs probabilistic values ranging between 0 and 1.

In the context of heart disease prediction, Logistic Regression is useful for classifying patients into two categories—those at risk of heart disease and those not at risk—based on various medical parameters. The algorithm uses the logistic function (also known as the sigmoid function), which maps input values to a probability range between 0 and 1, making it ideal for binary classification tasks.

Mathematical Representation of Logistic Regression

Logistic Regression is derived from the Linear Regression equation and is represented as follows:
The general equation of a straight line is:

$$Y=b_0+b_1X_1+b_2X_2+b_3X_3+\ldots+b_nX_n \qquad (4.1)$$

Since logistic regression requires the output Y to be between 0 and 1, we modify the equation by expressing it in terms of odds:

$$Y/1\text{-}Y \qquad (4.2)$$

Taking the natural logarithm (logit function) transforms the equation into:

$$\log(1-YY)=b_0+b_1X_1+b_2X_2+b_3X_3+\ldots+b_nX_n \qquad (4.3)$$

This equation helps determine the probability that a given input falls into a particular category.

4.1 Why Logistic Regression for Heart Disease Prediction?

Effective for Binary Classification: Heart disease prediction is typically framed as a classification problem—whether a patient has heart disease (1) or not (0).

Interpretable Model: The coefficients of the logistic regression model provide insight into the contribution of each risk factor.

Handles Multiple Features: The algorithm can incorporate various independent variables such as blood pressure, cholesterol levels, age, smoking habits, and more to predict the likelihood of heart disease.

Computationally Efficient: Logistic Regression is less complex than other machine learning models, making it ideal for real-time medical predictions.

By applying Logistic Regression in this research, we aim to develop a robust predictive model that can assist in the early detection of heart disease and aid medical professionals in making informed decisions.

V. RESEARCH METHODOLOGY

Table 5.1: Literature review

Title of Research Paper	Author	Publication	Year	Dataset Used	Algorithms Used	Accuracy (%)	Technology Used
Heart Disease Prediction using Machine Learning Techniques	Samir Patel, Devansh Shah, Santosh Kumar Bharti	Springer	2020	Cleveland Dataset (303 patients, 76 parameters, 14 used)	Naïve Bayes, Decision Tree, K-Nearest Neighbor, Random Forest	Naïve Bayes: 88.16% Decision Tree: 80.26% KNN: 90.79% Random Forest: 86.84%	WEKA
A Prediction of Heart Disease Using Machine Learning Algorithms	Mohd Faisal Ansari, Bhavya	Springer	2021	UCI Machine Learning Repository (304 patients,	Logistic Regression, Support Vector	Logistic Regression (All attributes): 87% Logistic	-

	Alankar Kaur, Harleen Kaur			14 attributes)	Machines	Regression (PCA): 86% SVM: 68%	
Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms	Kaushalya Dissanayake, Md Gapar Md Johar	Hindawi	2021	Cleveland Dataset (UCI ML Repository)	SVM, Decision Tree, Random Forest, Logistic Regression, KNN	Decision Tree: 81.96% Random Forest: 83.60% Logistic Regression: 83.60% KNN: 63.92%	Python (Jupyter)
Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning	Rohit Bharti, Mohammad Shabaz	Hindawi	2021	Public Health Dataset (Cleveland, Hungary, Switzerland, Long Beach V)	Logistic Regression, KNN, SVM, Random Forest, Decision Tree	Logistic Regression: 83.3% KNN: 84.8% SVM: 83.2% Random Forest: 80.3% Decision Tree: 82.3%	Python
Heart Disease Prediction using Machine Learning and Data Mining	Keshav Srivastava, Dilip Kumar Choubey	IJRTE	2020	Cleveland Heart Disease Dataset	KNN, Naïve Bayes, Decision Tree, SVM	KNN: 87% Naïve Bayes: 83% Decision Tree: 71% SVM: 84%	Python

VI. PROPOSED APPROACH

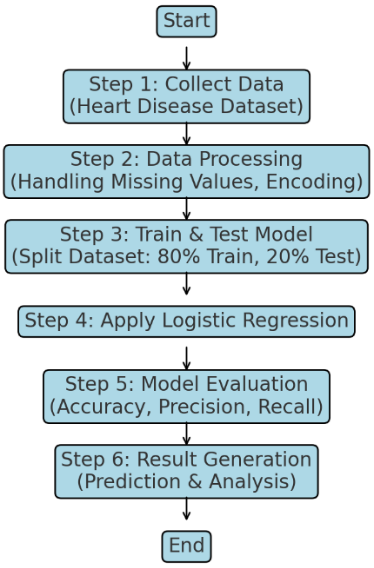
Start
Step 1: Collect Data (Heart Disease Dataset)
Step 2: Data Processing (Handling Missing Values, Encoding)
Step 3: Train & Test Model (Split Dataset: 80% Train, 20% Test)
Step 4: Apply Logistic Regression
Step 5: Model Evaluation (Accuracy, Precision, Recall)
Step 6: Result Generation (Prediction & Analysis)
End

Figure 6.1: Proposed Approach

VII. IMPLEMENTATION

The dataset is processed using Python, with libraries such as Pandas, NumPy, Scikit-learn, and TensorFlow. Feature selection techniques are applied to determine the most significant attributes influencing heart disease. The trained models are compared based on their accuracy and reliability in predicting heart disease.

The dataset used in this study consists of 319,795 samples and includes 18 attributes relevant to heart disease prediction. These attributes encompass various health indicators, lifestyle factors, and demographic details. The primary target variable, Heart Disease, is a binary classification with two possible values: Yes (1) indicating the presence of heart disease and No (0) indicating its absence.

The dataset includes attributes such as BMI, Smoking, Alcohol Consumption, Stroke History, Physical and Mental Health, Walking Difficulty, Gender, Age Category, Race, Diabetic Status, Physical Activity, General Health, Sleep Duration, Asthma, Kidney Disease, and Skin Cancer. Each attribute contributes to analyzing and predicting the likelihood of heart disease, making it a valuable resource for training machine learning models.

This dataset serves as the foundation for implementing predictive models and evaluating their effectiveness in heart disease detection.

Table 7.1: Information of Dataset Attributes

No.	Attribute	Values
1	HeartDisease	Yes:1, No:0
2	BMI	Number
3	Smoking	Yes:1, No:0
4	AlcoholDrinking	Yes:1, No:0
5	Stroke	Yes:1, No:0
6	PhysicalHealth	Number
7	MentalHealth	Number
8	DiffWalking	Yes:1, No:0
9	Sex	Male, Female
10	AgeCategory	18-24: 0, 25-29: 1, 30-34: 2, ... 80 or older: 12
11	Race	White, Hispanic, Black, Other, Asian, American Indian/Alaskan Native
12	Diabetic	Yes, No
13	PhysicalActivity	Yes:1, No:0
14	GenHealth	Poor: 0, Fair: 1, Good: 2, Very good: 3, Excellent: 4
15	SleepTime	Number
16	Asthma	Yes:1, No:0
17	KidneyDisease	Yes:1, No:0
18	SkinCancer	Yes:1, No:0

VIII. RESULTS AND DISCUSSION

This section presents a comparative analysis of machine learning algorithms used for heart disease prediction. The Logistic Regression model serves as a strong baseline, achieving an impressive 91% accuracy in classifying patients at risk of heart disease. Its effectiveness lies in its ability to model relationships between risk factors and disease occurrence using a probabilistic approach.

The findings highlight the importance of feature selection and data preprocessing in enhancing prediction accuracy. By optimizing input variables and ensuring data quality, the Logistic Regression model demonstrates significant potential for early detection and risk assessment in medical applications.

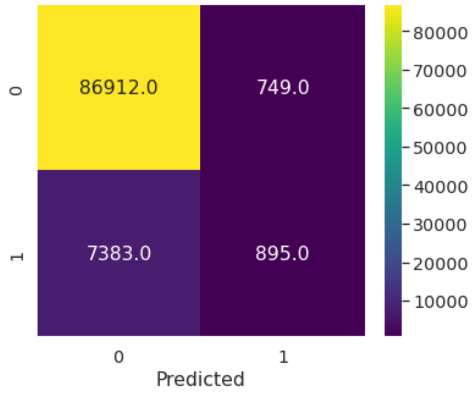

Figure 8.1: Confusion Matrix

IX. References

[1]M. Hoffman, "Picture of the Heart," WebMD, 23 June 2021. Available: https://www.webmd.com/heart/picture-of-the-heart.

[2]"Heart Disease," CDC, 27 September 2021. Available: https://www.cdc.gov/heartdisease/about.htm.

[3]"Heart disease," Mayo Clinic, 9 February 2021. Available: https://www.mayoclinic.org/diseases-conditions/heartdisease/symptoms-causes/syc-20353118.

[4]"About Heart Disease," CDC, 21 September 2021. Available: https://www.cdc.gov/heartdisease/about.htm. Novel approach for prediction of Heart Disease using Machine Learning Technique

[5]J. Beckerman, "Heart Disease: Types, Causes, and Symptoms," WebMD, 14 June 2021. Available: https://www.webmd.com/heart-disease/heart-disease-types-causes-symptoms.

[6]A. Felman, "What to know about coronary artery disease," Medical News Today, 20 July 2021. Available: https://www.medicalnewstoday.com/articles/184130.

[7]"Heart arrhythmia," Mayo Clinic, 22 Aprill 2022. Available: https://www.mayoclinic.org/diseases-conditions/heart-arrhythmia/symptoms-causes/syc-20350668.

[8]"Heart Disease and Pericarditis," webMD, 24 August 2020. Available: https://www.webmd.com/heart-disease/guide/heart-disease-pericardial-disease-percarditis.

[9]"Supervised Machine Learning," Javatpoint. Available: https://www.javatpoint.com/supervised-machine-learning.

[10]"Unsupervised Machine Learning," Javatpoint. Available: https://www.javatpoint.com/unsupervised-machine-learning.

[11]"Reinforcement learning," Geeks for Geeks, 18 November 2021. Available: https://www.geeksforgeeks.org/what-is-reinforcement-learning/.

[12]"Logistic Regression in Machine Learning," Javatpoint. Available: https://www.javatpoint.com/logistic-regression-in-machine-learning.