Predicting Heart Disease: A Machine Learning Approach to Early Detection

Hasti D. Patel[1], Nirali Borad[2]
[1]*Student, Atmiya University*
[2]*Assistant Professor, Atmiya University*

***Abstract*— Heart disease is a major global health issue, emphasizing the need for early detection. With advancements in biotechnology generating vast data, such as genetic and clinical records, machine learning (ML) has emerged as a promising tool for heart disease prediction. Identifying diseases based on symptoms alone is challenging and often reliant on doctors' expertise, which may not always yield accurate diagnoses. This research evaluates seven techniques: Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Naïve Bays, and Support Vector Machine (SVM). Additionally, it proposes a data mining approach that leverages advanced techniques to facilitate early detection and prevention, offering significant benefits to both patients and doctors.**

***Keywords*— Heart Disease Prediction, Machine Learning (ML), Data Mining, Early Detection, Healthcare Technology.**

I. INTRODUCTION

The heart is an essential organ responsible for maintaining life by pumping blood and distributing it throughout the body. Heart disease has emerged as a major global health concern, posing a serious risk to human well-being. According to World Health Organization estimates, heart disease accounts for approximately 12 million deaths worldwide annually. Heart ailment has overtaken cancer as the leading cause of death worldwide in recent years, and it is now a major public health issue, not just in India but around the world[4].

Despite its high incidence, heart disease can be prevented through dietary and lifestyle adjustments, medical interventions, and technological advancements[1]. The noticeable upsurge in heart disease in adults are mainly due to habits of smoking, work and non-work-related stress, unhealthy diet, lack of sufficient physical activity and excess consumption of salt and packaged food[3]. Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery and with the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved[3]. A healthful lifestyle and early detection of heart disease can save a person from death[2]. There are many data mining techniques are available for extract knowledge and information from large dataset. Most of the dataset is in discrete form for medical analysis. We can use machine learning classification algorithms for diagnosis and prediction of different type of disease[2].Machine learning algorithms were used to develop the models to predict the risk of heart disease[3]. The absence of early detection of many diseases, including cancer, diabetes, and others, is the primary reason why these illnesses are increasingly the cause of fatalities on a global scale[5].

II. DATA MINING AND MACHINE LEARNING IN HEALTHCARE SECTOR

Data mining and machine learning are transforming the healthcare sector by extracting meaningful insights from complex and extensive datasets. These cutting-edge technologies enable predictive analytics, enhance disease diagnosis, refine treatment approaches, and streamline resource management.

III. APPLICATIONS IN HEALTHCARE SECTOR

- Predicting and Diagnosing Diseases
- Tailored Medical Treatments
- Discovering and Developing New Drugs
- Monitoring Patients and Detecting Issues
- Analyzing Health Records

ML algorithms have revolutionized decision-making, offering tremendous potential for innovation and improvement in various industries[1]. Heart disease prediction, Heart Beat makes it easy for a general user to self-assess their chances of being diagnosed with heart disease and

the precautions that needs to be taken by the user[3]. If heart disease of a patient is predicted on early stage then the patient will have the better treatment for their heart disease and prevent their health from terrible consequences[2].

IV. THEORATICAL FRAMEWORK

In this research, seven techniques from machine learning algorithms were employed to analyze and predict outcomes effectively. These techniques include Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (SVM). Each of these methods brings unique strengths to the analysis, enabling a comprehensive comparison of their performance in the context of the study.

1. Linear Regression:

Linear regression is a supervised machine learning algorithm that learns from labeled data. It identifies the optimal linear function that aligns with the data points, enabling it to predict results for new data. The algorithm establishes a linear relationship between the dependent variable and one or more independent features by fitting a linear equation to the observed data. Linear regression is used to predict continuous output values based on input variables.

For instance, when predicting the price of a house, various factors like the house's age, distance from the main road, location, area, and the number of rooms are taken into account. Linear regression evaluates these factors and determines the relationship between them and the house's price, making predictions based on this linear connection.

2. Logistic Regression:

Logistic Regression is a widely used supervised machine learning algorithm primarily designed for binary classification problems, where the target variable has two possible outcomes. It models the relationship between one or more independent variables (predictors) and a binary dependent variable by estimating probabilities using a logistic function, also known as the sigmoid function. Unlike linear regression, which predicts continuous outcomes, logistic regression predicts the probability of the dependent variable belonging to a particular class. The output is a value between 0 and 1, which can be interpreted as a probability.

Logistic regression is simple yet powerful and works well when there is a linear relationship between the independent variables and the log-odds of the dependent variable. It is commonly used in various fields, including medical research, credit scoring, and social science, for tasks such as disease prediction, customer segmentation, and fraud detection.

3. K-Nearest Neighbors(KNN):

K-Nearest Neighbors (KNN) is a straightforward yet essential classification method in machine learning. Classified as a supervised learning algorithm, it is widely applied in areas such as pattern recognition, data mining, and intrusion detection.

KNN is highly effective in practical scenarios due to its non-parametric approach, which means it does not assume any specific distribution of the data, unlike algorithms like Gaussian Mixture Models (GMM) that rely on the assumption of a Gaussian distribution. The algorithm works by classifying data points into categories based on the features of the training data.

4. Decision Tree:

A decision tree is a diagram used to represent different choices in solving a problem and how various factors are interconnected. It follows a hierarchical structure, starting with a main question at the top (root node) and branching out into potential outcomes. The elements of a decision tree consist of:

Root Node: This is the origin point of the tree, representing the entire dataset.

Branches: These are the connecting lines between nodes, illustrating the flow from one decision to the next.

Internal Nodes: These are locations where decisions are based on input features or criteria.

Leaf Nodes: These terminal nodes at the end of branches represent the final outcomes or predictions.

5. Random Forest:

Random Forest is a supervised machine learning method. This algorithm is a powerful tree-based approach used for making predictions by combining the results (voting) of multiple decision trees. It is frequently used for both classification and regression tasks.

As a classifier, Random Forest constructs several decision trees to predict outcomes. It randomly selects subsets of the dataset to train each individual

tree and then combines the predictions by averaging their results. This technique boosts the accuracy of the predictions. Random Forest is an ensemble learning method in which the combined output from multiple models improves the overall performance.

6. Naïve Bayes:

It is based on Bayes' theorem classification technique which having assumption of independence among predictors.

This technique assumes that the presence of one feature in a particular class is not related to the presence of any other feature. Naïve Bayes model is very easy and quick to build and can be used for a large dataset.

$$P(c/x) = \frac{p(x/c)p(c)}{p(x)}$$

p(c|x): The posterior probability for class (c, target) given predictor (x, attributes).

P(c) : The prior probability of class.

P(x|c) : The probability for predictor given class.

P(x) : The prior probability of predictor.

7. Support Vector Machine (SVM):

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression tasks. Its primary objective is to find an optimal hyperplane that separates data points of different classes in the feature space. SVM focuses on maximizing the margin, which is the distance between the hyperplane and the nearest data points, known as support vectors. This margin maximization enhances the model's ability to generalize, making it effective for both linear and non-linear problems, especially in high-dimensional datasets.

For non-linear datasets, SVM employs the kernel trick, a method that transforms data into a higher-dimensional space where a linear hyperplane can separate classes effectively. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels, allowing SVM to adapt to various data distributions. SVM is widely used in applications such as text classification, image recognition, disease diagnosis, and financial forecasting, making it a versatile tool for solving complex machine learning problems.

V. METHODOLOGY

This section outlines the research methodology, covering the study design, participant details, data collection processes, analytical methods, and study limitations.

A. Study Design

We evaluated the performance of seven machine learning algorithms in predicting heart disease using a publicly available dataset. The dataset, sourced from the UCI Repository, contained information on 300+ patients, including variables such as age, sex, blood pressure, cholesterol levels, and symptoms, with the target variable indicating whether the patients were diagnosed with heart disease.

The machine learning algorithms compared in this study included Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (SVM).

B. Data Analysis

The data analysis involved training the models using 80% of the data and evaluating their performance on the remaining 20% designated as testing data.

1) Model Training:

The study involved training seven machine learning algorithms, namely Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (SVM). These models were implemented using Python, leveraging the following libraries:

Pandas: For data manipulation, analysis, and loading the dataset.

NumPy: For numerical operations and handling arrays.

Matplotlib: A visualization library for creating plots and charts to represent data insights.

Scikit-learn: For implementing machine learning models, data preprocessing (e.g., scaling features), splitting datasets, and evaluation metrics (e.g., accuracy).

2) Evaluation Models:

The performance of the trained models was evaluated using accuracy as the primary metric. Ensemble learning techniques, like Random Forest, were utilized to improve predictive performance by combining multiple decision trees. This comprehensive approach ensured a robust evaluation of each algorithm's capabilities in the context of heart disease prediction.

C. Statistical Analysis

The accuracy of each model was analyzed to identify the top-performing algorithms for predicting heart disease.

Feature	Description	Type
age	Age of the individual.	Numerical
sex	Gender of the individual (1 = Male, 0 = Female).	Categorical
cp	Chest pain type: 0 = Typical angina 1 = Atypical angina 2 = Non-anginal pain 3 = Asymptomatic.	Categorical
trestbps	Resting blood pressure (in mm Hg) on admission to the hospital.	Numerical
chol	Serum cholesterol level (mg/dl).	Numerical
fbs	Fasting blood sugar level > 120 mg/dl (1 = True, 0 = False).	Categorical
restecg	Resting electrocardiographic results: 0 = Normal 1 = Having ST-T wave abnormality 2 = Showing probable or definite left ventricular hypertrophy.	Categorical
thalach	Maximum heart rate achieved.	Numerical
exang	Exercise-induced angina (1 = Yes, 0 = No).	Categorical
oldpeak	ST depression induced by exercise relative to rest.	Numerical
slope	Slope of the peak exercise ST segment: 0 = Upsloping 1 = Flat 2 = Downsloping.	Categorical
ca	Number of major vessels (0-3) colored by fluoroscopy.	Numerical
thal	Thalassemia: 1 = Normal 2 = Fixed defect 3 = Reversible defect.	Categorical
target	Presence of heart disease (1 = Heart disease, 0 = No heart disease).	Categorical

Data Set[6]

The dataset, obtained from an unspecified source, includes a binary classification of patients' health metrics and symptoms, along with information indicating whether they have heart disease. It contains 303 patient records and 14 attributes, such as age, gender, blood pressure, cholesterol levels, chest pain type, maximum heart rate achieved, and other relevant factors. The target variable represents a binary classification, identifying whether the individual has been diagnosed with heart disease

VI. RESULTS

The Naïve Bayes achieved the highest accuracy of 93.25%, followed closely by K-Nearest Neighbors (KNN) at 92.25%. Random Forest also performed notably well, attaining an accuracy of 92%. Support Vector Machine (SVM) and Logistic Regression recorded accuracies of 88%, respectively. The Decision Tree algorithm achieved an accuracy of 86%, while Linear Regression had the lowest performance with an accuracy of 55%.

Algorithm	Accuracy (%)
Linear Regression	55.00
Logistic Regression	88.00
K-Nearest Neighbors (KNN)	92.25
Decision Tree	86.00
Random Forest	92.00
Support Vector Machine (SVM)	88.00
Naïve Bayes	93.25

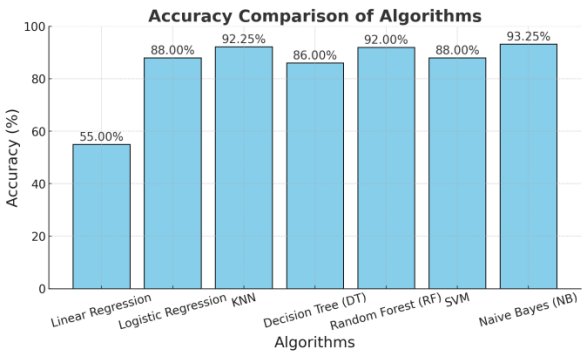

Fig. 1.Visualizing Results with Matplotlib

VII. RESEARCH CHALLENGES & FUTURE WORK

Disease prediction is a critical aspect of the healthcare sector, playing a vital role in saving lives. While numerous algorithms are available for predicting diseases, no single approach is universally effective across all datasets or types of diseases. The integration of data mining and machine learning techniques in healthcare has significantly enhanced prediction accuracy and strengthened data security, ensuring the protection of sensitive patient information. Despite the promising accuracy of machine learning algorithms in predicting heart diseases, challenges persist, particularly when handling large datasets. Addressing these issues requires further research to develop algorithms tailored to specific data types and healthcare needs.

However, this study has certain limitations, including reliance on a single dataset and the exclusive focus on accuracy as the primary metric for evaluating model performance. Expanding future research to explore various feature engineering techniques and experimenting with diverse machine learning approaches could address these constraints and improve the robustness and applicability of prediction models.

VIII. CONCLUTION

The findings demonstrate the potential of machine learning to enhance the early detection and management of heart diseases. These models can also be extended to identify conditions like cardiovascular and coronary artery diseases, ultimately aiming to minimize their impact on individuals and society. Naïve Bayes and K-Nearest Neighbors (KNN) demonstrated outstanding performance with accuracy scores of 93.25% and 92.25%, respectively, while Random Forest also excelled with an accuracy of 92.00%. Support Vector Machine (SVM), Logistic Regression, and Decision Tree showed good performance, achieving accuracy ratings of 88%, 88%, and 86%, respectively. In contrast, Linear Regression, with an accuracy score of 55%, was the least effective algorithm for heart disease prediction in this study. This study compared the accuracy of seven distinct machine learning for heart disease prediction. The insights gained provide valuable comparisons for researchers and healthcare professionals striving to develop accurate and reliable models for heart disease prediction.

REFERENCES

[1] Atul, Pranav Shankar, Srikari Rallabandi, Bhavya Singabhattu, Ashish Shinde"Evaluation of Machine Learning Approaches for Accurate Heart Disease Prediction"*14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*,(IEEE-56998),2023

[2] Abhishek Gupta, Vansh Misra, Ketan Chauhan, Dr. Kumar Manoj"Heart Disease Prediction Using Machine Learning"*International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*,pp.108-112,2023

[3] Sunitha Guruprasad, Valesh Levin Mathias, Winslet Dcunha"Heart Disease Prediction Using Machine Learning Techniques"*International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*,pp.762-766,2021

[4] Kirti Wankhede, Bharati Wukkadada, Sangeetha Rajesh, Sneha Nair"Machine Learning Techniques for Heart Disease Prediction"*Somaiya International Conference on Technology and Information Management (SICTIM)*,pp.28-33,2023

[5] Govindaraj M, V Asha, Binju Saju, Sagar M, Rahul"Machine Learning Algorithms for Disease Prediction Analysis"*5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*,pp.879-888,2023

[6] UCI Machine Learning Repository. https://archive-beta.ics.uci.edu/dataset/45/heart+disease

[7] Priyanka N, Dr.Pushpa RaviKumar"Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree"*International Conference on circuits Power and Computing Technologies (ICCPCT)*,2017

[8] Dhara B. Mehta, Nirali C. Varnagar"Newfangled Approach for Early Detection and Prevention of Ischemic Heart Disease using Data Mining"*Third International Conference on Trends in*

Electronics and Informatics (ICOEI),pp.1158-1162,2019

[9] Mamatha Alex P, Shaicy P Shaji"Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique"*International Conference on Communication and Signal Processing*,pp.848-852,2019

[10] Suhitha Katari, Thanguturu Likith, Muthineni Phani Sai Sree, Venubabu Rachapudi"Heart Disease Prediction using Hybrid ML Algorithms"*International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*,pp.121-125,2023

[11]]Gambhir Singh, Naveen Kumar, Aashish, Prerna Kumari"Heart Disease Prediction using Machine Learning"*International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*,2023

[12] Saumya Bansal, Rakhee"A Smote based Heart Disease Prediction Approach using Conventional Models"*International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*,2023

[13] Nurbaity Sabri, Anis Amilah Shari, Khyrina Airin Fariza Abu Samah, Mohd Rahmat Mohd Noordin, Asma Shazwani Shari, Fadhilah Mohd Ishak, Wan Masnieza Wan Mustapha, Muhammad Fudhail Afiq Nor Rozaini Affendi"Heart Inspect: Heart Disease Prediction of an Individual Using Naïve Bayes Algorithm"*IEEE 11th Conference on Systems, Process & Control(ICSPC)*,2023

[14] Ms. Yamini Ratawal, Sabya, Saurav, Swati"Heart disease prediction using Hybrid machine learning Model"*International Journal of Research in Engineering and Science (IJRES)*,2022

[15] Nikhitha Yathiraju, Arun Sankar, S Sandhiya, Suresh kumar .r , Santhosh K, Rajavetriselvan.S"Cardiac Disease Prediction for Heart Monitoring using Data Mining Techniques"*IEEE International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*,2022

[16] Ignatious K Pious, K Antony Kumar, Y.Cephas Soulwin and E.Nipun Reddy"Heart Disease Prediction Using Machine Learning Algorithms"*International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*,2022

[17] Parizat Binta Kabir and Sharmin Akter"Emphasized Research on Heart Disease Divination Applying Tree Based Algorithms and Feature Selection"*International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*,2021

[18] Likitha KN, Nethravathi R, Nithyashree K, Ritika Kumari, Sridhar N, Venkateswaran K"Heart Disease Detection using Machine Learning Technique"*The Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*,2021

[19] Miss. K.S.Ubale, Dr. P.N.Kalavadekar"Effective Heart Disease Prediction Using Machine Learning Techniques"*International Journal Of Advance Scientific Research & Engineering Trends (ISSN)*,2021