ISSN No:-2456-2165

Heart Disease Prediction

Dr. Chandrasekar Vadivelraju[1] (Professor); Duttala N Sughanditha Reddy[2]; Korrapati Praneeth Kumar Gowd[3]
Katakaraju Vaahini[4]; Pujari Suresh[5]
School of Computer Science and Engineering, Presidency University, Bengaluru, India

Abstract:- The increasing breakthroughs in illness diagnosis classification and identification systems have led to a steady growth in the incorporation of machine learning in medical diagnostics. These systems provide crucial data aiding medical professionals in the early detection of fatal diseases, significantly enhancing patient survival rates. Globally, heart disease stands as the leading cause of death. The escalating rates of heart strokes among juveniles underscore the need for an early detection system to prevent potential incidents. Frequent and costly tests like electrocardiograms (ECG) are impractical for the general population. As a result, a simple and trustworthy method for estimating the risk of heart disease is suggested. This system makes use of machine learning techniques and algorithms including Support Vector Classifier (SVC), Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN). It provides a useful method of heart disease prediction by analyzing several factors that users provide through the frontend interface.

I. INTRODUCTION

In the quick-paced world of today, individuals are grappling with intense stress and anxiety due to hectic schedules and routine assignments. Moreover, Some people become enmeshed in long-term habitual activities, like smoking cigars and drinking gutka, which can result in the development of major health concerns like cancer, heart disease, liver disorders, and renal failure. Treating individuals with chronic diseases has become a significant challenge for renowned doctors, making it a pressing global issue. In response to this challenge, IT professionals are actively providing support to predict and address these diseases early, aiming to help patients recover from chronic conditions.

In the current scenario, each individual possesses unique characteristics and habits, include differences in blood pressure readings and pulse rates. Researchers and medical professionals concur that a healthy individual's blood pressure should range between 120/80 to 140/90 mm Hg, and their pulse rate should be between 60 and 100 beats per minute (bpm). Heart syndrome is a leading cause of sudden or accidental death worldwide, often attributed to poor dietary habits, lack of physical exercise, and activities such as alcohol consumption and smoking.

This article looks at a number of variables, including age, gender, blood pressure, heart rate, and diabetes, in an effort to forecast and assess heart syndrome. However, predicting heart syndrome accurately remains a challenging task for medical practitioners and analysts. Although the health industry employs various machine learning tools and techniques to predict chronic diseases, researchers are still identifying flaws in existing methodologies. Consequently, they are in search of more effective and efficient predictive algorithms to detect human chronic diseases at an early stage, thereby saving lives.

To address this need, the proposed system uses a variety of machine learning techniques and algorithms, such as K-Nearest Neighbors (KNN), Random Forest, Naïve Bayes, and Support Vector Classifier (SVC). Using a front-end interface, the system attempts to forecast cardiac disease based on several user-entered characteristics., providing a potential solution to the early detection of chronic diseases.

➢ *Objective*

Using machine learning techniques and algorithms, such as K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest, and Naïve Bayes, is the main goal of this research. With the use of several parameters that users provide via the front-end interface, this project seeks to forecast heart disease.

➢ *Problem Statement*

The prediction and detection of heart disease pose persistent challenges for healthcare practitioners. Currently, healthcare facilities provide costly therapies and surgeries to address heart diseases, making early detection crucial. Detecting heart disease in its initial stages holds significant value for individuals globally, allowing them to take necessary actions before the condition becomes severe. This project strives to contribute to the early prediction of heart disease, offering a proactive approach to healthcare and potentially reducing the need for expensive treatments.

II. LITERATURE SURVEY

- This study looks at the prediction of heart disease using a range of methods, such as logistic regression, Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, and K-Nearest Neighbor. These algorithms were assessed using many metrics, including F1-score, precision, accuracy, AUC, and others. Based on the data, Random Forest performs better than the other supervised machine learning algorithms, with an 83.52% prediction rate for heart disease. Based on the experimental data, the Random Forest classifier similarly displays an F1-Score, AUC, and precision score of 84.21%, 88.24%, and 88.89%, respectively.[1]

- Three different disease databases—diabetes, heart disease, and breast cancer—from the UCI repository were subjected to predictive analysis using different classification techniques, and each produced a different set of benefits. Using the p-value test and backward modeling, features were chosen for each dataset. The results of the study highlight the potential of machine learning for early sickness diagnosis even more.. [2]

- The study presents Apache Spark, a stable distributed computing platform perfect for large-scale operations, in the context of a real-time heart illness prediction system. To effectively manage streaming data events, it incorporates machine learning using in-memory algorithms. The streaming processing component and the two main parts of the system are the data storage and display components. In the former, a classification model is applied to data events using Spark MLlib and Spark streaming in order to predict heart illness. In the meantime, the latter stores the significant volume of generated data using Apache Cassandra. [3]

- The study presents Apache Spark, a stable distributed computing platform perfect for large-scale operations, in the context of a real-time heart illness prediction system. To effectively manage streaming data events, it incorporates machine learning using in-memory algorithms. The streaming processing component and the two main parts of the system are the data storage and display components. In the former, a classification model is applied to data events using Spark MLlib and Spark streaming in order to predict heart illness. In the meantime, the latter stores the significant volume of generated data using Apache Cassandra. [4]

- Investigated heart disease, a significant cause of patient mortality. The complexity in diagnosing these conditions arises due to symptoms overlapping with other ailments such as chest pain, shortness of breath, palpitations, and nausea, causing diagnostic challenges. Their research concentrated on using machine learning algorithms to diagnose cardiac disease, highlighting the significance of patient data weighting to increase diagnostic precision. They also introduced a method to determine weight coefficients. Their proposed approach achieved a success rate of 86.90%, analyzing 13 unique patient-derived features.[5]

- This work uses machine learning techniques such Gaussian Naïve Bayes, Random Forest, K-Nearest Neighbor, and Support Vector Machine to construct a Heart Disease Prediction system, focusing on the critical topic of heart disease. Thirteen features are used by the framework, such as age, gender, blood pressure, cholesterol, and obesity, among others. Efficient healthcare decision-making is facilitated by the user-friendly solution, which comprises dataset uploading, algorithm selection, accuracy prediction, and model development. [6]

III. METHODOLOGY

➢ *Data Collection:*
A heart disease dataset is acquired for analysis.

➢ *Preprocessing and Data Loading:*
The system loads the dataset. and undergoes preprocessing using various machine learning techniques. This involves handling missing values, scaling, and encoding categorical variables.

➢ *Data Splitting:*
Training and testing sets of preprocessed data are created. datasets to ensure a reliable evaluation of the machine learning models.

➢ *Model Development:*
Prediction models are constructed using machine learning techniques such as Random Forest, Naïve Bayes, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN).

➢ *Training the Model:*
To help the model identify patterns and correlations in the data, it is trained using the training dataset.

➢ *Model Testing:*
Using the testing dataset, the trained model is put to the test, and accuracy metrics are computed to assess how well each method performs.

➢ *Model Selection:*
The ultimate prediction model is determined by selecting the algorithm with the best accuracy.

➢ *Model Serialization:*
The completed model is stored for later usage and transformed into a binary pickle model.

➢ *Front-End Programming:*
A user-friendly front end is developed using Flask and HTML, providing an interface for users to input various parameters for heart disease prediction.

➢ *User Input and Prediction:*
Users enter pertinent factors into the front end, and the final algorithm uses this data to forecast the risk of heart disease.

➢ *Result Display:*
The predicted output is displayed on the front end, informing the user whether the individual is likely to have heart disease or not.

ISSN No:-2456-2165

IV. SYSTEM DESIGN

➢ *System Architecture*

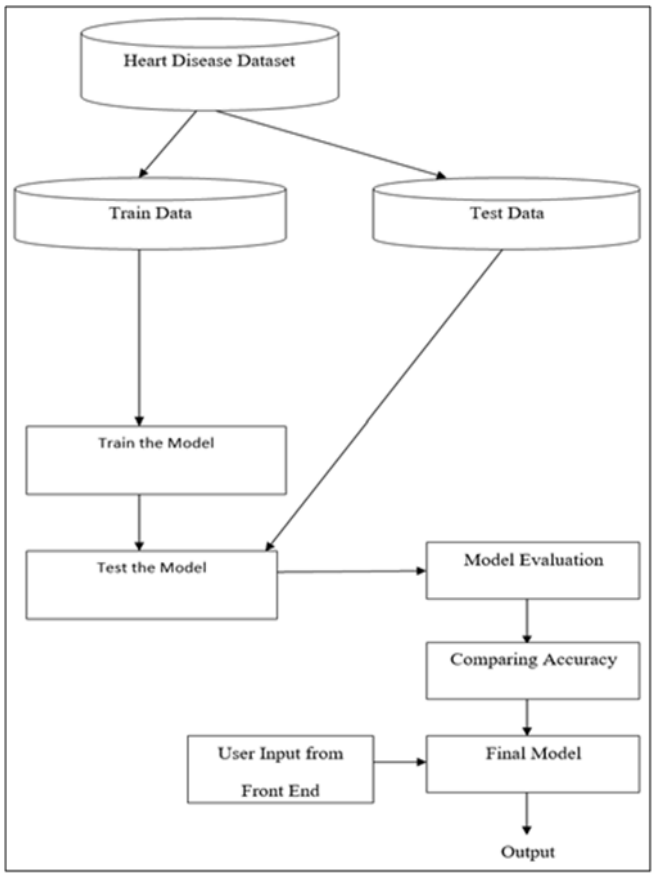

Fig 1 System Architecture

➢ *Heart Disease Dataset:*

A dataset containing heart disease-related information is obtained for analysis.

➢ *Data Preprocessing:*

The dataset undergoes preprocessing to remove errors like duplicates and missing numbers and to clean up and interpret the data.

➢ *Data Splitting:*

The preprocessed data is divided into training and testing datasets to facilitate the building and evaluation of models.

➢ *Model Building:*

Prediction models are constructed using machine learning techniques such as XGB Classifier, Random Forest Classifier, KNN, SVC, Naïve Bayes, Decision Tree Classifier, and Logistic Regression.

➢ *Training the Model:*

The selected machine Preprocessed training data is used to train the learning model so that it can identify patterns and correlations.

➢ *Testing Models:*

Using the testing dataset, the trained model is evaluated, and accuracy is determined. for each machine learning algorithm.

➢ *Model Selection:*

The algorithm demonstrating the highest accuracy is chosen as the final prediction model.

➢ *User Input and Prediction:*

In the front end, users input various parameters for heart disease prediction, and the finalized algorithm predicts the likelihood of heart disease.

➢ *Data Flow Diagram*

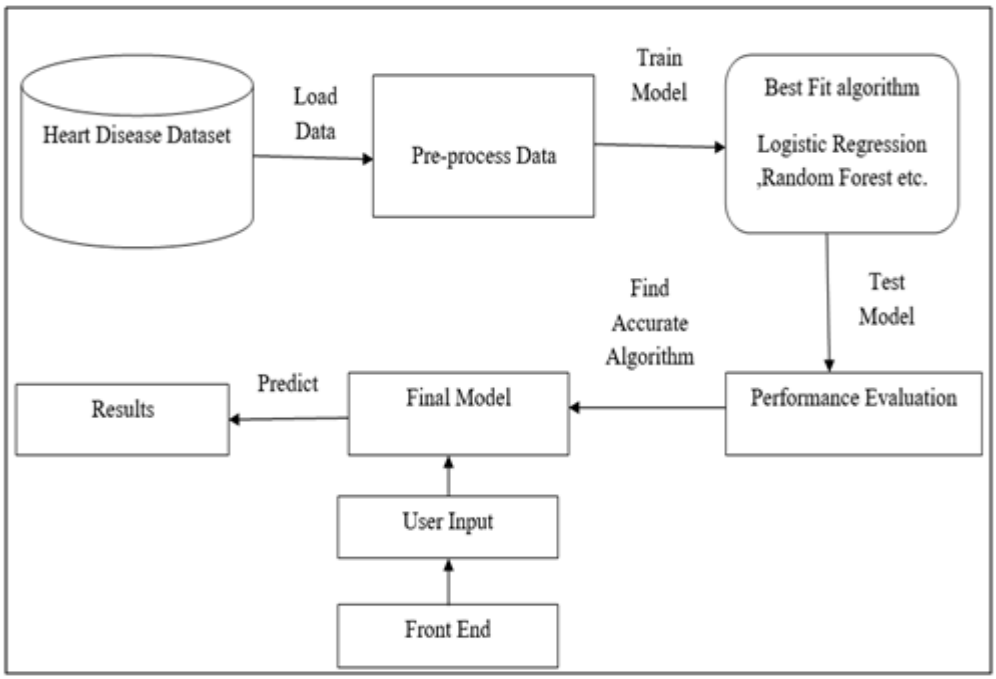

Fig 2 Data Flow Diagram

- *Heart Disease Dataset:*

A dataset containing heart disease-related information is taken and loaded into the system.

- *Data Preprocessing:*

The data undergoes preprocessing, which improves accuracy by removing duplicates and addressing missing values.

- *Data Splitting:*

Training and testing datasets are created using the preprocessed data.

- *Model Building:*

Prediction models are constructed using machine learning techniques as KNN, SVC, Naïve Bayes, Random Forest Classifier

- *Getting the Model Ready:*

The preprocessed training data is used to train the model.

- *Testing Models:*

For every machine learning approach, accuracy measures are computed and the trained model is evaluated.

- *Model Choice:*

The final model for predicting heart disease is determined by selecting the algorithm that has the best accuracy.

- *User Contribution and Forecast:*

The finished algorithm uses the data that users supply through the front end to forecast the possibility of heart disease.

V. ALGORITHMS

➢ *Random Forest Classifier:*

The supervised learning technique known as the Random Forest algorithm works well for applications involving both regression and classification, earning its reputation as a highly flexible and user-friendly approach. Comprising an ensemble of decision trees, this method emphasizes robustness by incorporating numerous trees within its framework. Here's an overview of how it operates: Random Forest generates decision trees on randomly selected data samples. Each tree provides predictions, and the algorithm selects the best solution via voting among these predictions.This method also offers insights into feature importance, aiding in understanding the significance of different attributes.Technically categorized as an ensemble method utilizing a divide-and-conquer strategy, Random Forests build decision tree classifiers on randomized subsets of the dataset. Attribute selection metrics such as information gain, gain ratio, or Gini index are used in the construction of each tree. These trees operate independently, with classification problems resolved through voting among the trees and regression tasks obtaining the average output as the final result.Noteworthy advantages include: High accuracy and robustness due to the involvement of multiple decision trees. Mitigation of overfitting concerns by averaging predictions, balancing out biases. Applicability to both classification and regression problems. Capability to handle missing values, employing methods like replacing continuous variables with median values or utilizing proximity-weighted averages.In essence, the Random Forest classifier constructs decision trees from randomly chosen subsets of the training data, forming an assembly of decision trees whose collective votes determine the final prediction.

➢ *Naive Bayes Classifier :*

The naive Bayes classifier, a generative model used in classification, held prominence in machine learning applications before the emergence of user-friendly deep learning libraries. Despite its straightforwardness, this classifier demonstrated commendable performance across various applications. This probabilistic machine learning model revolves around the fundamental principles of the Bayes theorem.

➢ *Bayes Theorem:*

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

Utilizing Bayes theorem involves determining the probability of event A occurring given the occurrence of event B. In this context, B stands as the evidence, while A represents the hypothesis. The "naive" assumption underlying this theorem presumes independence among predictors or features, implying that the presence of one feature doesn't impact another.

➢ *Naive Bayes Classifier Types:*

- *Multinomial Naive Bayes:*

Primarily employed in document classification tasks, this classifier assesses whether a document belongs to specific categories like sports, politics, or technology. Predictors utilized by the classifier correspond to word frequencies present within the document.

- *Bernoulli Naive Bayes:*

Similar in essence to the multinomial variant, Bernoulli Naive Bayes operates with predictors represented as boolean variables. These variables evaluate parameters based on binary values—yes or no—signifying the occurrence of a word within the text.

➢ *Support Vecotr Machine (SVM)*

SVM is a well-liked supervised learning method that may be applied to both regression and classification issues. Despite having machine learning categorization as its main use, support vector machines (SVM) are also used to find the best line or decision boundary in n-dimensional spaces. This boundary, also referred to as a hyperplane, is essential for classifying newly discovered data points into the appropriate groups. SVM determines this hyperplane based on extreme points known as support vectors, pivotal in its classification. The algorithm's types include: Linear SVM:

Ideal for linearly separable data, where a single straight line effectively categorizes the dataset into two classes.
Non-linear SVM is specifically designed for datasets that are not linearly separable, addressing scenarios where a straight line is inadequate for categorizing the data.

Hyperplane, pivotal in SVM, represents the best decision boundary for classifying data points. Its dimensions correspond to the dataset's features; for instance, in a two-feature dataset, the hyperplane is a straight line, while in a three-feature dataset, it takes the form of a two-dimensional plane. The data points closest to the support vectors are the hyperplane, exert significant influence on its placement, hence the name "Support Vector. "To handle non-linear separations, SVM employs kernels, employing a technique known as the kernel trick. Kernels transform input data spaces, converting non-separable problems into separable ones by introducing higher dimensions. The radial basis function (RBF), polynomial, and linear variants are among the variations; each has a specific function in classification tasks.

Advantages of SVM lie in its accuracy, swift prediction, and memory efficiency compared to algorithms like Naïve Bayes. It excels when clear separation margins and high-dimensional spaces are prevalent in the dataset.

- *K-Nearest Neighbors (KNN) :*

Notably, a straightforward supervised machine learning method that can be used to both regression and classification problems is the k-nearest neighbors (KNN) algorithm. Unlike unsupervised algorithms, it relies on labeled input data to comprehend patterns and generate suitable outputs when presented with new, unlabeled data.Based on the premise that similar entities exist in close proximity, KNN hinges on the concept of proximity or similarity, utilizing mathematical calculations, such as distance metrics, to gauge resemblance between data points.

Advantages of the KNN algorithm encompass its simplicity, ease of implementation, and versatility. Unlike models requiring complex tuning or assumptions, KNN operates without the need to construct a specific model or fine-tune numerous parameters. Its adaptability extends to various applications, including classification, regression, and search tasks.

VI. RESULT AND DISCUSSION

The prediction and detection of heart disease pose critical challenges for healthcare professionals. In addressing the need for early identification and intervention, this initiative proposes a system leveraging machine learning techniques and algorithms, including KNN, SVC, Random Forest, and Naïve Bayes. The objective is to forecast heart disease based on user-input parameters, creating a user-friendly tool for timely diagnosis.The implemented project has yielded promising outcomes, achieving a notable accuracy of 81.97%, with the Random Forest Classifier playing a particularly effective role. This high accuracy underscores the efficacy of the proposed machine learning model in predicting heart disease.Timely identification of heart disease is imperative for swift medical intervention, and the system's ability to predict with precision is a positive advancement for healthcare outcomes. The incorporation of multiple algorithms enhances adaptability and robustness in addressing diverse data and scenarios.

VII. CONCLUSION

In conclusion, the developed system showcases the potential of machine learning in the medical field, specifically for heart disease prediction. The high accuracy obtained emphasizes the significance of incorporating such technologies in healthcare systems to enhance early diagnosis and improve patient outcomes

Table 1 The High Accuracy Obtained

SI.NO	Models	Accuracy
1	KNN	65.57%
2	GaussianNB	80.33%
3	SVC	83.61%
4	Random Forest	81.97%

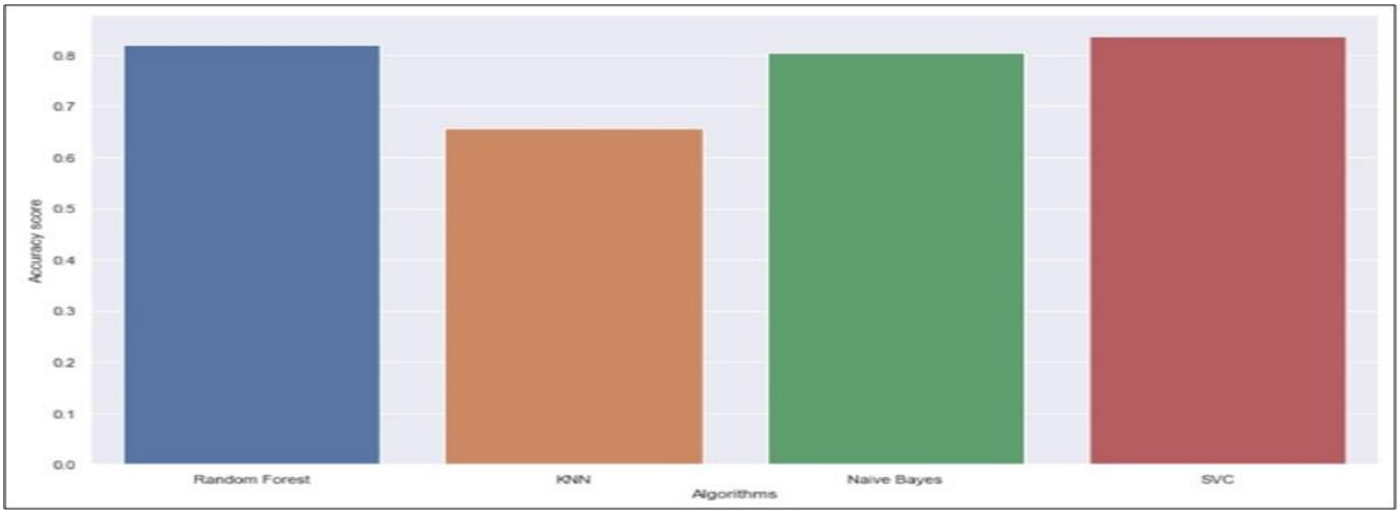

Fig 3 The High Accuracy Obtained

REFERENCES

[1]. "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," P. Sujatha and K. Mahalakshmi, 2020. Sujatha and Mahalakshmi presented at the IEEE International Conference for Innovation in Technology (INOCON) in 2020, a performance evaluation of supervised machine learning techniques for heart disease prediction. The objective of the research is to evaluate the prognostic potential of many algorithms for cardiac conditions and ascertain the comparative advantages and disadvantages of every algorithm.

[2]. P. S. Kohli and S. Arora (2018) "Application of Machine Learning in Disease Prediction" During their 2018 presentation at the 4th International Conference on Computing Communication and Automation (ICCCA), Kohli and Arora talked about the use of machine learning to the prediction of illness. The project investigates the potential of machine learning algorithms for illness prediction, with a focus on how these methods could benefit healthcare technologies.

[3]. Kamalmi and A. Ed-Daoudy (2019): "Big Data Approach to Real-time Machine Learning for Early Heart Disease Detection" Ed-Daoudy and Maalmi presented a research in 2019 at the International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS). The goal of the project is to increase the accuracy of real-time machine learning systems for the early detection of cardiac problems by utilizing big data approaches.

[4]. T. S. R. Kiran, A. Srisaila, and A. Lakshmanarao (2021): "Feature Selection and Ensemble Learning Techniques for Heart Disease Prediction" Lakshmanarao et al. presented a paper at the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) in 2021. Their study on heart disease prediction includes feature selection and ensemble learning approaches to improve overall prediction accuracy.

[5]. S. Güney and A. Erdoğan (2020): "Heart Disease Prediction by Using Machine Learning Algorithms" Erdoğan and Güney's 2020 presentation at the 28th Signal Processing and Communications Applications Conference (SIU) enhanced the prediction of cardiac illness. The study examines how well various machine learning algorithms can predict cardiac disease.

[6]. "Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques" was the study done in 2020 by S. Farzana and D. Veeraiah. In 2020, at the 5th International Conference on Computing, Communication, and Security (ICCCS), Farzana and Veeraiah presented their research findings. Their research focuses on dynamic cardiac illness prediction, utilizing multi-machine learning approaches to increase precision and flexibility. Reword without utilizing any copied material.