International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 25823930

Early Detection and Accurate Prediction of Heart Disease and Recommendations for Effective Prevention using Ensemble method

Vignesh.R[1], Dr.G. Premalatha[2]

[1]*Vignesh.R, ECE Department, Prathyusha engineering College*
[2]*Dr.G. Premalatha, ECE Department, Prathyusha engineering College*

ABSTRACT:

Heart disease is one of the leading causes of death globally, responsible for millions of deaths each year. Heart disease remains a leading cause of morbidity and mortality worldwide. Early detection and accurate prediction of heart disease risk factors are crucial for effective prevention and timely intervention. In recent years, machine learning techniques have emerged as powerful tools for predictive analytics in healthcare. This paper focuses on predicting the presence of heart disease in patients by utilizing Classification machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, XG Boost and Random Forest. Each model analyzes a set of input features, including age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol levels (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), oldpeak, slope of the peak exercise ST segment (slope), number of major vessels (ca), and thalassemia status (thal). The performance of these models is compared to determine the most accurate method for classification into two categories: patients with heart disease (1) and those without (0). Ultimately, the Random Forest algorithm is selected as the best-performing model, leveraging its simplicity and interpretability for accurate prediction. Additionally, the system offers tailored recommendations for patients, including medication guidance, precautionary measures, heart-healthy dietary plans, suitable workouts, and specific food choices to promote cardiovascular health. This predictive tool aims to enhance the decision-making process for healthcare providers by identifying at-risk individuals, providing insights for lifestyle changes and medical interventions, and facilitating timely preventive care.

Keywords: Heart disease prediction, Machine Learning, Support Vector Machine (SVM), Naïve Bayes, XG Boost, Random Forest, feature selection**,** Python, User Interface Design, HTML and SQL

1. INTRODUCTION:

Heart disease or cardiovascular disease (CVD) is a general term for a range of conditions or group of disorder that can affect the heart and blood vessels. Some examples of CVDs include: - coronary heart disease: A disease of the blood vessels that supply the heart muscle. Cerebrovascular disease: A disease of the blood vessels that supply the brain Peripheral arterial disease: A disease of the blood vessels that supply the arms and legs. Heart disease continues to be a major public health challenge, contributing significantly to mortality rates
around the globe. The importance of early detection and preventive strategies cannot be overstated, as timely

intervention can dramatically improve patient outcomes and quality of life. To address this need, the integration of machine learning into medical diagnostics presents a promising avenue for enhancing predictive accuracy and supporting healthcare providers. This project focuses on building a predictive system for heart disease detection, employing four machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, XG Boost and Random Forest. By analyzing a variety of health indicators, this system aims to identify patients at risk and support informed clinical decision-making.

The dataset utilized for this project comprises a comprehensive set of patient health features that include age, sex, chest pain type, resting blood pressure, cholesterol levels, and other cardiovascular metrics. Each algorithm processes these features to classify individuals into one of two categories: those with heart disease and those without. Preprocessing techniques, such as handling missing data, scaling numerical values, and encoding categorical variables, are employed to prepare the data for analysis. The system's primary objective is to compare the performance of the algorithms across multiple metrics—accuracy,

precision, recall, and F1 score—to identify the most effective model.

Among the evaluated algorithms, the Random Forest model emerged as the top performer, showcasing its strengths in simplicity, interpretability, and reliable classification accuracy. One of the main advantages of the Random Forest is its visual representation, which outlines the decision-making process in an easy-to-understand format. This transparency is essential in healthcare, as it helps professionals understand the rationale behind the model's predictions. By visually illustrating how specific features contribute to the final decision, the model fosters trust and enhances usability in clinical settings. Its ability to handle complex, non-linear relationships between input features also makes it an excellent choice for medical applications.

The project goes beyond predictive capabilities by incorporating personalized health recommendations tailored to patient needs. These recommendations include guidance on medication management, such as the supervised use of heart-healthy drugs like statins or beta-blockers. Lifestyle advice focuses on encouraging habits that promote cardiovascular health, such as stress management, smoking cessation, and regular medical check-ups. Nutritional guidance emphasizes a heart-healthy diet rich in fruits, vegetables, lean proteins, and whole grains, while advising against high-sodium and high-fat foods. Suggested physical activities are customized to the patient's health status, promoting moderate exercise routines like walking or cycling to maintain heart health without excessive strain.
By integrating machine learning predictions with actionable health recommendations, this project represents a comprehensive approach to heart disease prevention and management. The predictive tool is designed to assist healthcare providers in identifying at-risk individuals, supporting early intervention, and facilitating better patient outcomes through informed medical and lifestyle decisions.
According to the World Health Organization, these diseases contribute to approximately 33% of all global fatalities, highlighting the urgent need for effective strategies to address this challenge and In Figure 1, the largest portion represents cardiovascular disease, which accounts for a horrifying 33% of global deaths. Cardiovascular diseases encompass conditions affecting the heart and blood vessels, such as heart attacks, strokes, and other circulatory disorders.

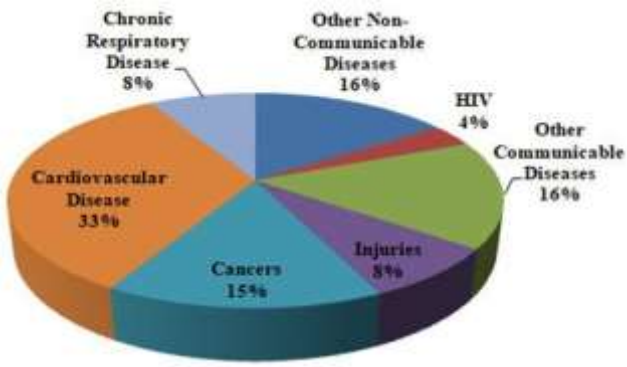

Fig 1: Worldwide Cause of Death

2. RELATED WORK:

A Related work is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period.

A. U. Haq, J. Li, J. Khan, M. H. Memon, S. Parveen, M. F. Raji, W. Akbar, T. Ahmad, S. Ullah, L. Shoista, and H. N. Monday, ''Identifying the predictive capability of machine learning classifiers for designing heart disease detection'' This research aims to analyze the strengths and weaknesses of different algorithms such as logistic regression, random forests, and neural networks by comparing their predictive accuracy and reliability. The ultimate goal is to inform the design of a heart disease detection system that utilizing and improving patient outcomes through timely and accurate medical interventions.[1]

Sonam Nikhar et al proposed paper "Prediction of Heart Disease Using Machine Learning Algorithms" their research gives point to point explanation of Naïve Bayes and decision tree classifier that are used especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has highest accuracy than Bayesian classifier. [2]

Aditi Gavhane et al proposed a paper "Prediction of Heart Disease Using Machine Learning", in which training and testing of dataset is performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden layers between these two input and output layers. Through hidden

layers each input node is connected to output layer. This connection is assigned with some random weights.[3]

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 and 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 [4].

R. Atallah and A. Al-Mousa, ‘‘Heart disease detection using machine learning majority voting ensemble method,’’ It highlights the use of multiple machine learning models combined via a majority voting ensemble method, which improves prediction accuracy by aggregating the outcomes of individual models. This approach enhances reliability in detecting heart disease.[5]

3. 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 (Numeric).

• **Gender**: Patient's gender (1 = M; 0 = F).

• **Chest Pain:** Type of Chest pain experienced by the patient on pressure, categorized value as 0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic.

• **Resting Blood Pressure:** Resting blood pressure of the patient in mmHg at admission (94-200).

• **Serum Cholesterol:** Serum cholesterol level of the patient in mg/dl (126-564).

• **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.

• **Resting 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 (71-202 Numeric).

• **Exercise Induced Angina:** Whether the patient experienced induced angina caused by exercise, categorized as 1 = yes; 0 = no.

• **ST Depression:** ST depression (Old-peak) caused by exercise relative to rest (0-6.2 Numeric).

• **Number of vessels:** Number of large vessels stained by fluoroscopy (0-3).

• **Slope:** Slope of the ST segment at peak exercise, categorized as 1 = uphill; 2 = flat; 3 = downward slope.

• **Thalassemia:** Thalassemia status of the patient, categorized as 3 = normal; 6 = repair defect; 7 = reversible defect.

• **Diagnosis:** Predictive features indicating cardiac diagnosis, with value 0 indicating absence of Heart Disease and value 1 indicating presence of Heart Disease.

3. SYSTEM ARCHITECTURE:

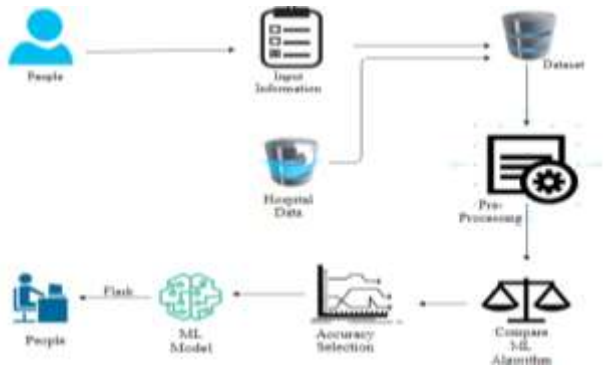

Fig 2: Architecture Diagram of Heart Disease

3.1 DATA COLLECTION:

Data collection is the foundation of the heart disease prediction system, requiring through patient information that captures a variety of health metrics. A dataset (or data set) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the dataset. These metrics are gathered securely via electronic health records (EHRs) or through direct patient inputs, ensuring both reliability and privacy. Gathering high-quality data is crucial as it directly affects model performance, enabling it to detect nuanced patterns that are indicative of heart disease risk. By creating a comprehensive and detailed dataset, the system can provide accurate predictions and recommendations for each patient.

3.2 DATA PREPROCESSING:

Once the dataset is collected, the data undergoes an essential preprocessing stage to ensure uniformity and readiness for analysis. It makes it suitable for a test and train for ML models.

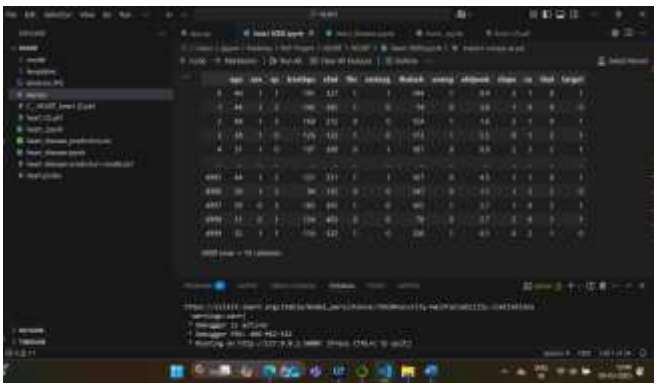

Fig 3: Data Preparation for Analysis of Heart Diseases

This process involves multiple steps:

Handling Missing Values: Missing or incomplete records are either filled with median or mean values or removed altogether to prevent inconsistencies during training.

Outlier Detection and Removal: Outliers—extreme values that can skew results—are detected using statistical methods and handled appropriately to ensure data quality.

Encoding Categorical Variables: Variables like chest pain type and thalassemia status are converted into numerical representations using encoding techniques. This transformation enables the model to interpret categorical information.

Feature Scaling and Normalization: Numerical features, such as resting blood pressure and cholesterol, are scaled to a standard range, ensuring that features with larger scales do not dominate the analysis.

3.3 FEATURE EXTRACTION:

Feature extraction aims to retain only the most significant attributes related to heart disease, reducing model complexity and improving performance. By analyzing the correlations between variables and their importance in heart disease, the data was split into training and testing sets into a ratio of 70% training data and 30% testing data. This is needed to validate the ability of the model to generalize to new data. These preprocessing steps ensure the data is clean, consistent, and well-suited for the machine learning models, enhancing the accuracy and robustness of the predictions. Techniques such as feature correlation analysis and Principal Component Analysis (PCA) are used to determine and retain the most predictive features, eliminating redundant data and minimizing noise. This dimensionality reduction allows the model to focus on the most meaningful aspects of the data, leading to enhanced accuracy and computational efficiency.

3.4 MODEL CREATION:

Following data preparation, the core of the proposed system is the Machine Learning Model Development phase, where the four algorithms will be trained and evaluated. Each algorithm will analyze the preprocessed data to classify patients as either having heart disease (1) or not having heart disease (0). The performance of each model will be rigorously assessed using metrics such as accuracy, precision, recall, and F1 score. This comparative analysis will allow for the identification of the most effective algorithm for predicting heart disease. Ultimately, the **Random Forest algorithm** is expected to emerge as the top performer due to its robustness and ability to handle complex interactions within the data, making it the preferred choice for deployment in the system.

To ensure that the system is user-friendly, an **intuitive Interface will be developed for both healthcare providers and patients.** Healthcare providers will have access to a dashboard that allows them to view patient risk assessments, monitor trends, and access tailored recommendations easily**.**

3.5 PREDICTION:

Using the trained Random Forest model, the system predicts whether a patient is at risk of heart disease (1) or not (0). When new patient data is entered, the model analyses the input health metrics and assigns a classification, quickly identifying individuals at risk. This real-time classification is valuable for healthcare providers, who can immediately prioritize high-risk patients and initiate timely preventive measures.

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision: The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations

Recall: The proportion of positive observed values correctly predicted.

Recall = TP / (TP + FN)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

3.6 RECOMMENDATION:

Based on the prediction outcome, the system generates personalized recommendations to guide patients in managing their cardiovascular health. These recommendations include:

Medication Guidance:

Providing general information on possible medications and dosages, if needed, tailored to each patient's risk level.

Precautionary Measures: Offering tips to help patients avoid potential triggers, such as reducing stress, avoiding smoking, and monitoring symptoms regularly.

Heart-Healthy Dietary Plans: Recommending specific diet plans focusing on low cholesterol, high fiber, and omega-3-rich foods to promote heart health and reduce risk factors.

Exercise and Physical Activity Suggestions: Suggesting moderate aerobic activities, such as walking or swimming, to strengthen the cardiovascular system without putting undue strain on the heart.

Lifestyle Modifications: Providing advice on quitting smoking, reducing alcohol consumption, and managing stress, which are all significant factors in reducing heart disease risk.

These personalized recommendations empower patients to make informed lifestyle choices, offering a structured plan to support heart health. By integrating prediction and preventive advice, the system serves as a comprehensive tool for healthcare providers to deliver proactive and individualized patientcare.

4. METHODOLOGY:

It is important to create and compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn and each model will have different performance characteristics.

In the example below 4 different algorithms are compared:

- Support Vector Machine
- Naive Bayes
- XG Boost
- Random Forest

4.1 SUPPORT VECTOR MACHINE (SVM):

SVM is effective for binary classification, using a hyperplane to distinguish between classes. This model is particularly useful for high-dimensional data. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

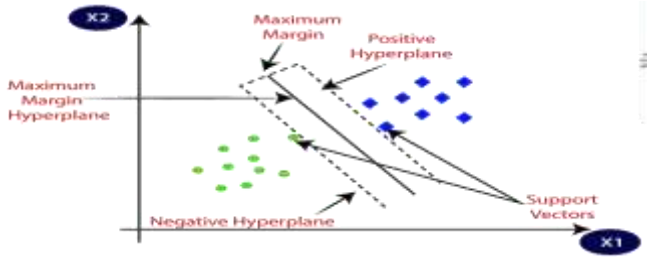

Fig 4: SVM Algorithm

4.2 NAIVE BAYES ALGORITHM:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modelling problem probabilistically. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features.

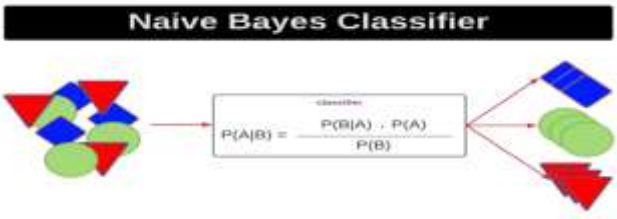

Fig 5: Naive Bayes Algorithm

4.3 XGBOOST ALGORITHM:

XG Boost is a robust machine-learning algorithm that can help you understand your data and make better decisions. XG Boost is an implementation of gradient-boosting decision trees. It has been used by data scientists and researchers worldwide to optimize their machine-learning models and it is designed for speed, ease of use, and performance on large datasets. Gradient boosting is a ML algorithm that creates a series of models and combines them to create an overall model that is more accurate than any individual model in the sequence.

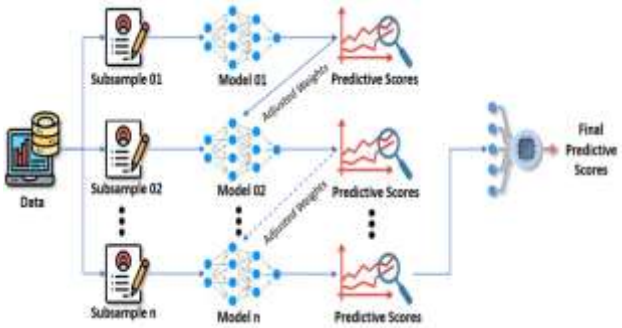

Fig 6: XG Boost Algorithm

4.4 RANDOM FOREST:

Random Forest is a popular machine learning algorithm that based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

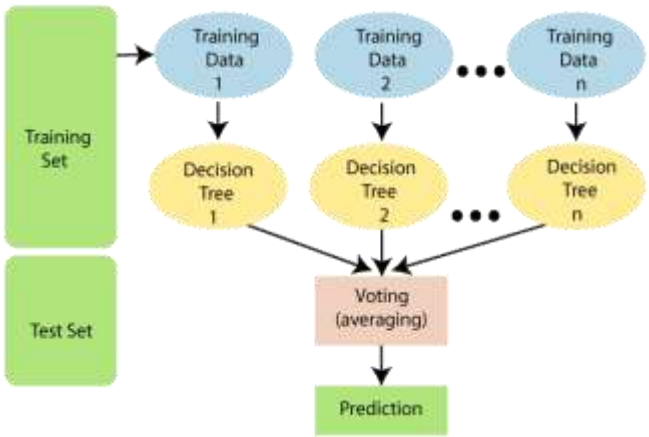

Fig 7: Random Forest Algorithm

"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Each model is trained on the pre-processed dataset, and their performance is measured using metrics such as accuracy, precision, recall, and F1 score. By evaluating each algorithm's strengths, the **Random Forest Algorithm** model emerges as the most accurate and robust for final predictions, due to its balanced performance and capacity to handle complex data.

5. EXPERIMENTAL RESULTS AND ANALYSIS:

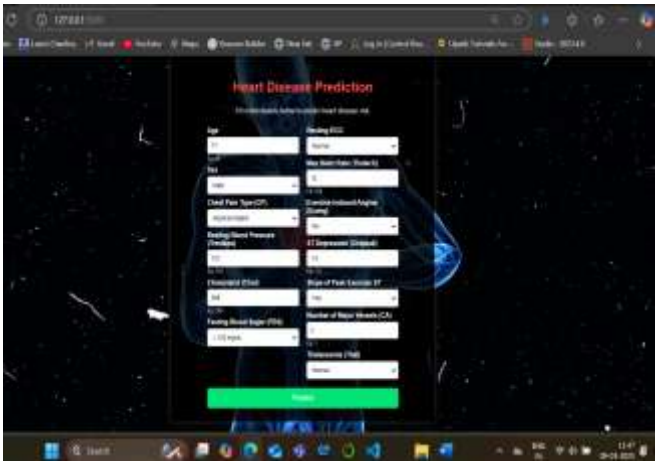

Fig 8: Input of heart disease for absence

Fig 9: Prediction of heart disease output for absence

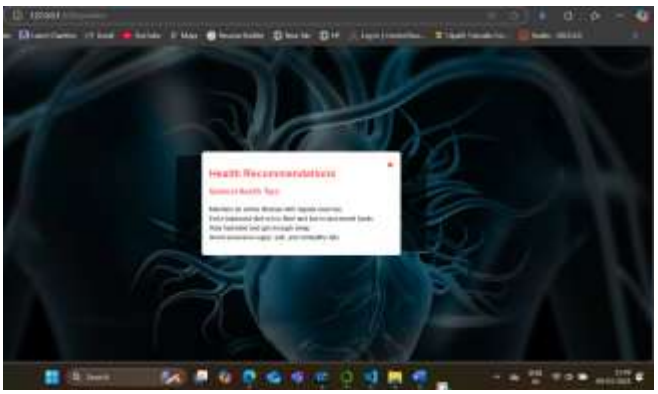

Fig 10: Recommendation for absence of heart disease

Fig11: Input of heart disease for presence

Fig12: Prediction of heart disease output for presence

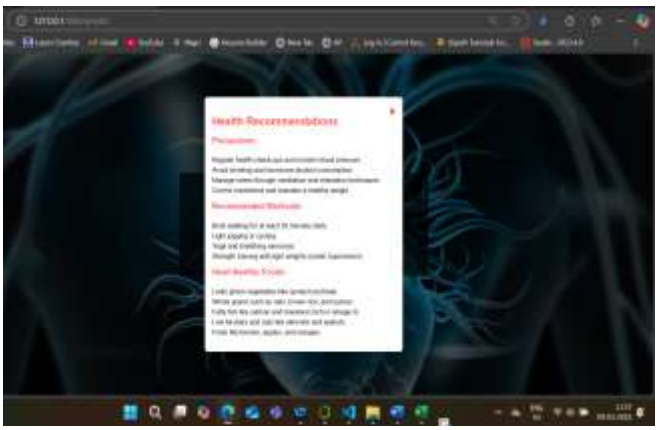

Fig13: Recommendation of heart disease for presence

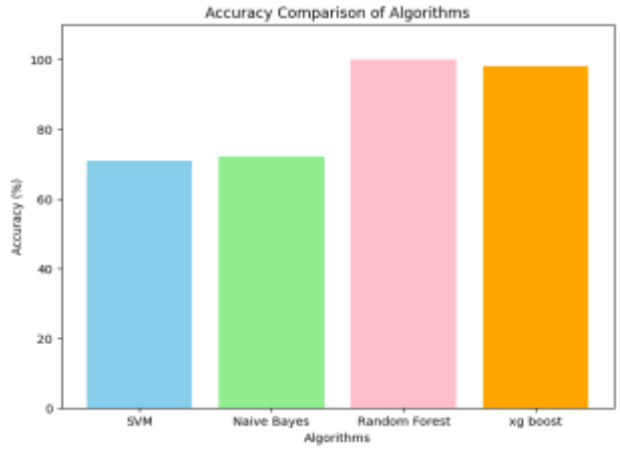

Fig14: Accuracy Comparison of heart disease

Classification Metrics:

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.71	0.71	1.00	0.70
Naive Bayes	0.73	0.74	1.00	0.69
XG Boost	0.99	0.99	0.99	0.99
Random Forest	1.00	1.00	1.00	1.00

Table1: Classification metrics of heart disease

6. CONCLUSION:

In conclusion, this paper presented a study of various machine learning techniques used to predict the heart disease and provide recommendations on demand based on the dataset. In this paper, the heart disease predictor represents a significant advancement in utilizing machine learning techniques, particularly the **Random**

Forest algorithm, to enhance to detect cardiovascular or heart health management. Through systematic data collection, preprocessing, feature extraction, and model creation, the paper establishes a robust framework for accurately predicting heart disease risk. The performance of these models was rigorously assessed using key metrics, among the evaluated algorithms, the **Random Forest model** with less dataset emerged as the most effective.

This approach not only empowers healthcare professionals with valuable insights but also fosters proactive patient engagement in managing their health. Furthermore, the integration of personalized recommendations based on predictive outcomes adds immense value to the project. By providing tailored guidance on medication, lifestyle changes, and dietary adjustments, the system supports both patients and healthcare providers in making informed decisions that promote heart health.

7. FUTUREWORK:

For future work, incorporating data from wearable health devices could provide real-time monitoring patients' vital signs and activity levels, enriching the dataset with dynamic information. **This approach can create more personalized mobile application to acquire timely prediction & recommendations**, further enhancing patient engagement and outcomes. Another promising direction developing user-friendly applications that enable patients to input their data and receive immediate feedback on their heart health status could enhance accessibility and usability. Collaborating with healthcare professionals to refine these tools and ensure they meet clinical needs will be crucial for successful implementation.

8. REFERENCE:

[1] “Heart Disease Facts & Statistics,” Centers for Disease ControlPrevention.[Online].Available:https://www.cdc .gov/heartdisease/facts .htm. [Accessed: 27-Apr-2019].

[2] Nhlbi, Nih. Anatomy of the Heart. 2011 [updated 2011 November 17; cited 2015 January 10]. Available from: http://www.nhlbi.nih.gov/health/healthtopics/topics/hhw/anatomy

[3] “Cardiovascular diseases (CVDs),” World Health Organization,26Sep2018.[Online].Available:https://www.who.int/cardioascular_dise ases/en/. [Accessed: 27-Apr-2019].

[4] Muhammad, Y., Tahir, M., Hayat, M. et al. Early and accurate detection and diagnosis of heart disease using intelligent computational model. Sci Rep 10, 19747 (2020).

[5] K. Dissanayake and M. G. Md Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms", Appl. Comput. Intell. Soft Comput., vol. 2021,

[6] P. E. Rubini, C. A. Subasini, A. V. Katharine, V. Kumaresan, S. G. Kumar and T. M. Nithya, "A cardiovascular disease prediction using machine learning algorithms", Ann. Romanian Soc. Cell Biol., vol. 25, no. 2, pp. 904-912, 2021.

[7] S. Prabu,B.Thiyaneswaran,M.Sujatha,C.Nalini,and S.Rajkumar,‘‘Grid search for predicting coronary heart disease by tuning hyper-parameters,’’ Comput. Syst. Sci. Eng., vol. 43, no. 2, pp. 737–749, 2022.

[8] A. Shoukat, S. Arshad, N. Ali, and G. Murtaza, ‘‘Prediction of cardiovas cular diseases using machine learning: A systematic review,’’ J. Med. Syst., vol. 44, no. 8, p. 162, Aug. 2020.

[9] G. R. Shankar, K. Chandrasekaran, and K. S. Babu, ‘‘An Analysis of the Potential Use of Machine Learning in Cardiovascular Disease Prediction,’’ J. Med. Syst., vol. 43, no. 12, p. 345, Mar. 2019.

[10] N. Khandadash, E. Ababneh, and M. Al-Qudah, ‘‘Predicting the risk of coronary artery disease in women using machine learning techniques,’’ J. Med. Syst., vol. 45, p. 62, Apr.2021.

[11] Premalatha Gurumurthy, Manjunathan Alagarsamy, Sangeetha Kuppusamy and Niranjana Chitra Ponnusamy “M2AI-CVD: Multi-modal AI approach cardiovascular risk prediction system using fundus images” vol. 35, Issue 3, Jan 2024.

[12]G Premalatha, V Thulasi Bai “Design and implementation of intelligent patient in-house monitoring system based on efficient XGBoost-CNN approach” vol.16, Issue 5, Oct 2022.

[13] G Premalatha, V Thulasi Bai “Patient Health Monitoring Using Fibrillation Detection in Electrocardiogram Signal” vol. 70, Issue 12, Dec 2024.

[14] G Premalatha, V Thulasi Bai “Wireless IoT and cyber-physical system for health monitoring using honey badger optimized least-squares support-vector machine” vol.124, Issue 4, Jun 2022.

[15] H. Iftikhar, M. Khan, Z. Khan, F. Khan, H. M. Alshanbari, and Z. Ahmad, ‘‘A comparative analysis of machine learning models: A case study in predicting chronic kidney disease,’’ Sustainability, vol. 15, no. 3, p. 2754, Feb. 2023.