Riya Jaiswal, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

HEART DISEASE PREDICTION (Xgboost Random Forest , And KNN)

Riya Jaiswal,Simran Sahu ,Prince Pandey

Guide : Vandana Tripathi

Department of Computer Science and Engineering,
MGM's College of Engineering and Technology ,Noida ,India

Abstract: Heart complaint remains one of the leading causes of mortality worldwide, making early discovery pivotal for effective treatment and forestallment. This design focuses on developing a prophetic model to identify the threat of heart complaint in individualities using machine literacy ways. By assaying patient data, including vital health pointers similar as age, blood pressure, cholesterol situations, casket pain type, and other applicable medical attributes, the model aims to classify individualities grounded on their liability of developing heart complaint. colorful bracket algorithms are applied and compared to determine the most accurate approach. The results demonstrate that machine literacy can serve as a dependable tool for aiding healthcare professionals in early opinion, enabling timely intervention, and eventually perfecting patient issues.

Keywords: Heart disease prediction, machine learning, early detection, mortality prevention, patient data analysis, health indicators, blood pressure, cholesterol, chest pain, classification algorithms, risk assessment, predictive model, healthcare support, timely intervention, medical diagnosis.

I. INTRODUCTION

1.1 Preface

Cardiovascular conditions, particularly heart complaint, are among the leading causes of death encyclopedically. Rapid urbanization, sedentary cultures, stress, and unhealthy salutary habits have contributed to the adding frequence of these conditions. Beforehand discovery and timely intervention are critical in reducing the threat of severe complications and perfecting patient issues. With the advancement of technology and the vacuity of large quantities of healthcare data, machine literacy has surfaced as a important tool for prognosticating conditions. By assaying patterns in patient data, prophetic models can help medical professionals in relating high- threat individualities, thereby enabling preventative measures. This design focuses on erecting a heart complaint vaticination system using patient medical records and crucial health pointers, aiming to support accurate, data- driven healthcare opinions and ameliorate overall heart health operation.

1.2 Motivation

Heart complaint is one of the most critical health challenges worldwide, responsible for millions of deaths each time. numerous cases of heart complaint can be averted or managed effectively if detected beforehand. still, traditional individual styles frequently calculate on homemade interpretation of symptoms and medical tests, which can be time- consuming and prone to mortal error. The provocation behind this design is to work ultramodern data analytics and machine literacy ways to prognosticate the liability of heart complaint in individualities directly and efficiently. By assaying patient health data including factors similar as age, blood pressure, cholesterol situations, and life habits this design aims to identify high- threat individualities beforehand,

enabling timely medical intervention. The ultimate thing is to give a dependable decision- support tool for healthcare professionals, reduce the burden of cardiovascular conditions, and ameliorate patient issues.

1.3 Problem Statement

Heart complaint continues to be a leading cause of death worldwide, and early discovery remains a significant challenge due to the complexity of symptoms and the variety of factors impacting cardiac health. Traditional individual approaches frequently depend on homemade assessment, which may lead to delayed or inaccurate identification of high- threat cases. With the growing vacuity of healthcare data, there's a need for an automated, dependable, and data- driven system that can help in prognosticating the liability of heart complaint. This design aims to develop a machine literacy – grounded vaticination model that analyzes crucial medical attributes to classify individualities according to their threat position. The ideal is to support healthcare professionals with a tool that enhances individual delicacy, enables early intervention, and eventually contributes to more patient issues.

1.4 Objectives

- To analyze patient health records and identify key factors that significantly influence the risk of heart disease.
- To develop a machine learning model capable of predicting the likelihood of heart disease using relevant medical and lifestyle attributes.
- To compare multiple classification algorithms and determine the most accurate and efficient model for heart disease prediction.
- To create a user-friendly system that can assist healthcare professionals in making data-driven decisions for early diagnosis and intervention.
- To improve the reliability and speed of prediction by automating risk assessment and reducing the chances of manual diagnostic errors.
- To contribute to preventive healthcare by enabling timely identification of high-risk individuals and promoting early treatment strategies.

1.5 Scope And Limitations

1.5.1 Scope

- The project focuses on using machine learning techniques to predict the likelihood of heart disease based on patient medical data.
- It includes the collection, preprocessing, and analysis of key health attributes such as age, blood pressure, cholesterol levels, blood sugar, chest pain type, and other relevant indicators.
- The system evaluates and compares different classification algorithms to identify the most accurate prediction model.
- The project aims to build a decision-support tool that can assist healthcare practitioners in early diagnosis and risk assessment.
- The scope covers model development, performance evaluation, and implementation of a simple interface for demonstration or practical use.

1.5.2 limitations

- The accuracy of the prediction model depends heavily on the quality, size, and diversity of the dataset used.
- The system cannot replace clinical expertise and should be used only as an aid, not as a final diagnostic tool.
- The model may not perform well for populations or patient groups that differ significantly from the data it was trained on.
- Limited attributes in the dataset may restrict the model's ability to capture all possible factors influencing heart disease.
- External factors such as lifestyle habits, genetic history, and environmental influences may not be fully represented in the dataset.
- Real-time prognostications may bear fresh computational coffers or integration with sanitarium systems, which is beyond the current design compass.

1.6 Assosiation Of The Design

The design is structured as follows:

- Chapter 1: Preface to the design.
- Chapter 2: Literature check related to the design.
- Chapter 3: Accoutrements and styles used for the design.
- Chapter 4: Perpetration details of the design.
- Chapter 5: Deployment phase of the design.
- Chapter 6: Conclusion of the design.
- Chapter 7: Future scope and implict advancements of the design.

II. LITERATURE SURVEY

1.1. Overview

Machine literacy(ML) has become a dominant approach for prognosticating heart complaint because it can discover complex patterns in clinical data that are delicate to capture with rules-grounded styles. Reviews and checks published in the last many times constantly report that classical supervised models (logistic retrogression, decision trees, SVM, arbitrary timber) and ensemble styles (grade boosting, XGBoost) are the most constantly used and frequently achieve competitive delicacy on standard clinical datasets. Recent methodical reviews also emphasize the growing part of careful point selection and mongrel/ensemble channels to raise robustness and interpretability.

2.2. Generally used datasets

The UCI Heart Disease collection — and specifically the Cleveland subset is the de facto standard used in most academic studies. The Cleveland dataset contains clinical and demographic attributes (generally a 14-point subset is used) and 303 cases; because it's small and well understood, numerous papers use it for algorithm comparison and evidence-of-conception work. Other datasets (Hungary, Switzerland, VA Long Beach) appear in combination or as indispensable sources, but most recent experimental work still evaluates models on Cleveland-derived data or Kaggle clones of it.

2.3. Classical ML algorithms and relative studies

A large body of work compares traditional classifiers on the UCI/Cleveland dataset. Common findings are:

- Logistic Retrogression (LR) and Support Vector Machines (SVM) frequently serve as strong nascences because of their simplicity and regularization options.
- Random Forest (RF) and grade boosting / XGBoost constantly achieve advanced prophetic performance because of ensemble averaging and runnig of nonlinearity.
- K-Nearest Neighbours (KNN) and Naive Bayes (NB) appear in early studies but generally underperform compared to tree ensembles. Relative studies across 2019–2025 report that RF and boosting styles are constanly top players, though the "styles" algorithm depends on preprocessing, point selection, and hyperparameter tuning.

2.4. Point engineering and selection

Because clinical datasets are small and may contain identified or noisy attributes, point selection mainly affects model quality. Ways generally applied include:

- Sludge styles (ANOVA, Chi-square, collective information) to rank features snappily.
- Wrapper styles (Recursive point Elimination — RFE) to find compact, high-performing subsets.
- Bedded styles (L1 regularization, tree-grounded significance) which elect features during model training. Recent studies show that combining sludge wrapper strategies or applying RFE before ensemble litracy yields measurable earnings in delicacy and reduces overfitting—especially when datasets are small.

2.5. Deep literacy and signal-grounded approaches

While irregular clinical trait models dominate, deep literacy approaches have gained traction where richer data (ECG signals, imaging) are available. Convolutional Neural Networks (CNNs), LSTMs, and cold-blooded CNN-LSTM channels are used to prize temporal and spectral ECG features automatically, frequently outperforming hand-drafted point channels when large ECG corpora are available. specially, large ECG-grounded threat models trained on hundreds of thousands or millions of traces (rather than small clinical irragular records) can prognosticate unborn cardiac events and long-term mortality — a recent high-profile illustration is a clinical AI ECG threat model being trialed in the NHS, trained on >1 million ECGs and demonstrating promising prognostic performance for unborn heart failure and mortality. These ECG-grounded models point to a reciprocal path: combine structured clinical features with learned signal representations for stronger threat position.

2.6. Ensemble and mongrel strategies

Ensemble designs (mounding, advancing, blending) and mongrel channels (point selection + ensemble + estimation) are extensively explored because they tend to stabilize prognostications across small, noisy datasets. Several 2023–2025 papers report mounding ensembles that combine tree models and SVMs or meta-learners to squeeze redundant performance while controlling overfitting. Hyperparameter optimization (grid hunt, randomized hunt, Bayesian tuning) and nested cross-validation are constantly used together with ensemblees to produce reproducible performance estimates.

2.7. Evaluation practices and reproducibility enterprises

Common evaluation criteria in the literature include delicacy, perfection, recall (perceptivity), particularity, F1-score, and area under the ROC wind (AUC). Because numerous datasets (like Cleveland) are small, authors emphasize k-fold cross-validation, stratified slice, and repeated trials to gain stable estimates. Still, literature reviews note reproducibility challenges: inconsistent preprocessing (class balancing, insinutation), unclear hyperparameter choices, and use of multiple dataset variants make direct comparison across papers delicate. methodical reviews call for standardized reporting (dataset splits, exact preprocessing channels, and law sharing) to enable fair benchmarking.

2.8. Recent trends (2022–2025)

Point-centric optimization: Newer studies concentrate on rigorous point selection and point-birth channels (PCA, MI, RFE) before model training; these constantly ameliorate both delicacy and model interpretability.

XGBoost and grade boosting variants: Numerous relative analyses in 2023–2025 indicate XGBoost or analogous grade boosting machines frequently perform at or above RF and SVM on irregular clinical data.

ECG + clinical emulsin large: Large-scale ECG-trained models (and their NHS trialing) point to clinically practicable vaticination of unborn adverse issues, suggesting signal-grounded models will decreasingly be combined with tabular predictors.

III. METHODOLOGY

3.1 System Armature

The system armature provides a high-position overview of how the entire model operates. A simplified representation of the workflow is illustrated below:

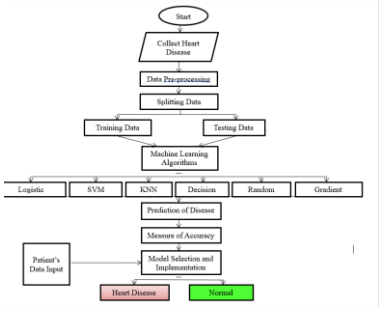


Figure: Proposed System

3.2 Dataset Details

Dataset

The dataset for this study is available at Kaggle. It contains information applicable to prognosticating heart failure and includes the following attributes:

1. Age: Age of the case in times.
2. Sex: coitus of the case(manly or womanish).
3. ChestPainType: Type of casket pain endured — Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP), or Asymptomatic (ASY).
4. Cholesterol: Serum cholesterol position in mg/dl.
5. RestingBP: Resting blood pressure in mmHg.
6. FastingBS: Indicator for dieting blood sugar (1 if >120 mg/dl, 0 else).
7. RestingECG: Results of resting electrocardiogram — Normal, ST-T surge abnormality (ST), or signs of left ventricular hypertrophy (LVH).
8. MaxHR: Maximum heart rate achieved, ranging from 60 to 202 bpm.
9. ExerciseAngina: Presence of exercise-convinced angina (Yes or No).
10. Oldpeak: Numeric value of ST depression convinced by exercise
11. ST_Slope: Slope of peak exercise ST member — Upsloping, Flat, or Downsloping.
12. HeartDisease: Target variable indicating opinion (1 = heart complaint, 0 = normal).

3.3 Machine Learning

Machine literacy involves creating prophetic models that can assign markers or orders to input data grounded on its features.

3.3.1 Supervised Machine Learning

Supervised machine literacy is used to prognostic heart complaint by training models on labeled patient data. Each record contains medical features as age, blood pressure, cholesterol, casket pain type—and an outgrowth indicating whether the case has heart complaint. The model learns patterns from this data and can also prognosticate the threat for new cases.

Common algorithms include Logistic Retrogression, Decision Trees, Random Forest, SVM, KNN, and Grade Boosting, all of which classify cases grounded on their health pointers. After training, models are estimated using delicacy, perfection, recall, and AUC to insure dependable performance.

Supervised literacy helps automate early opinion, supports croakers in decision-timber, and improves early discovery, though its delicacy depends on high-quality data and proper preprocessing.

3.3.2 Unsupervised Machine Learning

Unsupervised machine literacy is used to find retired patterns in patient data without counting on labeled issues. Rather of prognosticating whether a person has heart complaint, these styles group cases grounded on parallels in their health features similar as blood pressure, cholesterol, age, and ECG results.

Ways like K-Means Clustering, Hierarchical Clustering, and PCA help identify threat groups, descry unusual cases biographics, and uncover connections among attributes. Although unsupervised styles cannot give direct judgements, they help croakers explore patterns, member cases, and ameliorate understanding of factors linked to heart complaint.

3.4 Supervised Algorithms

3.4.1 Random Forest

Random Forest is an important supervised machine-learning algorithm extensively used for bracket tasks similar as prognosticating heart complaints. It works by creating a large number of decision trees, where each tree is trained on an arbitrary subset of the dataset and features. During vaticination, every tree gives its own affair, and the final vaticination is made grounded on the maturity vote.

In heart complaint vaticination prediction, Random Forest is effective because it can reuse complex medical features similar as blood pressure, cholesterol situations, age, ECG results, and casket pain type—while reducing the threat of overfitting. It identifies the most influential factors contributing to heart complaint and delivers high delicacy,

stability, and robustness. Its capability to handle missing values, nonlinear connections, and noisy data makes it a dependable model for healthcare operations.

3.4.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple and effective supervised machine-learning algorithm used to classify whether a case is likely to have heart complaint grounded on parallels with once cases. Rather of erecting a model during training, KNN stores all the data and compares a new case's medical features similar as age, blood pressure, cholesterol, casket pain type, and sugar levels—with the k most similar records in the dataset.

The algorithm calculates the distance (usually Euclidean) between the new patient and all existing patients, identifies the closest k neighbors, and assigns the majority class from those neighbors as the prediction. KNN works well for heart disease prediction because it naturally captures patterns in small to medium-sized datasets and adapts to complex decision boundaries. However, it requires proper data scaling and can be slow for large datasets.

3.4.3 Logistic Retrogression

Logistic Retrogression is a extensively used supervised machine- literacy algorithm that helps prognosticate the liability of a double outgrowth — similar as whether a person has heart complaint or not. rather of prognosticating a numerical value, Logistic Retrogression estimates the probability that a case belongs to the " complaint " class grounded on their medical features. In heart complaint vaticination, the algorithm analyzes crucial health pointers suchlike age, cholesterol position, blood pressure, casket pain type, and ECG results. It also applies a logistic(sigmoid) function to convert the input combination into a probability between 0 and 1. still, the model classifies the case as at threat of heart complaint, If the probability exceeds a set threshold. Logistic Retrogression is preferred in medical analysis because it's easy to interpret, computationally effective, and easily shows how each point influences the vaticination. still, it assumes a direct relationship between features and the outgrowth, which may limit performance when dealing with complex patterns.

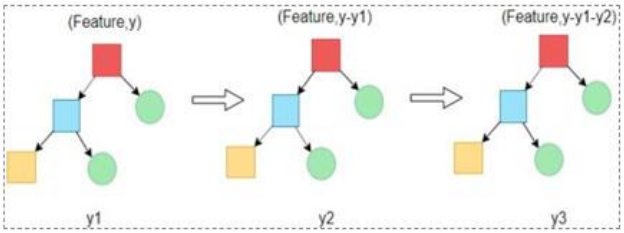


Figure: XG Boost

IV. Implementation

4.1 Existing System

The being system for detecting heart complaint substantially depends on traditional medical tests and homemade evaluation by croakers . Health parameters similar as blood pressure, cholesterol, ECG results, and casket pain details are reviewed collectively, without any automated tool to combine and dissect them. opinion frequently depends on a croaker 's experience, which may lead to detainments, mortal crimes, and inconsistent results. The current system focuses on relating heart complaint after symptoms appear, rather than prognosticating the threat beforehand.

4.2 Proposed System

The proposed system automates heart complaint vaticination using machine literacy. Case data similar as age, blood pressure, cholesterol, casket pain type, and ECG results are reused and anatomized by algorithms like Logistic Retrogression, KNN, or Random Forest. The system evaluates all features together, identifies retired patterns, and provides real- time threat prognostications. This approach improves delicacy, reduces mortal error, supports early opinion, and offers harmonious, data- driven decision support for healthcare providers.

Data Collection

- Data Preprocessing
- Point Selection

- Model Evaluation

4.2.1. Data Collection

Patient data is collected from clinical sources or intimately available datasets(e.g., UCI Heart Disease dataset). Each record contains crucial health features similar as age, blood pressure, cholesterol, casket pain type, ECG results, and a marker indicating the presence or absence of heart complaint. High- quality and different data is essential for erecting an accurate prophetic model.

4.2.2. Data Preprocessing

Raw data is gutted to remove duplicates, handle missing values, and correct inconsistencies. Categorical variables are decoded into numerical formats, and point scaling(normalization or standardization) is applied to insure all features contribute inversely. Preprocessing prepares the dataset for effective model training.

4.2.3. Point Selection

Point selection identifies the most important attributes that impact heart complaint vaticination. ways similar as correlation analysis, collective information, or recursive point elimination(RFE) are used to reduce noise and ameliorate model performance. opting applicable features also enhances interpretability and reduces computational cost.

4.2.4. Model Evaluation

Named features are used to train machine literacy models like Logistic Retrogression, K- Nearest Neighbors(KNN), Random Forest, or Decision Trees. The dataset is resolve into training and testing sets, and models are estimated using criteria similar as delicacy, perfection, recall, F1- score, and ROC- AUC. The best- performing model is chosen for vaticination and deployment.

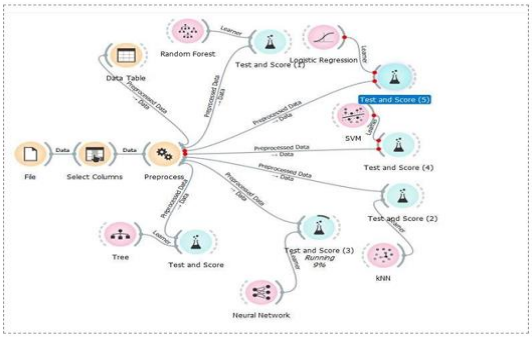


Figure: Connection of widgets in Orange

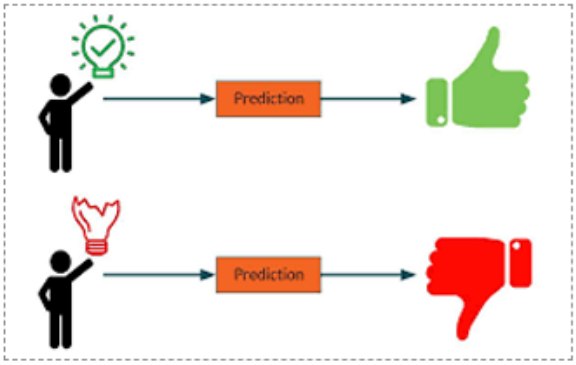


Figure: Prediction of Disease

In this study, we estimated the performance of several machine learning algorithms for heart complaint vaticination and recorded their separate rigor :

- XGBoost: 99.03% accuracy
- Random Forest: 96% accuracy
- K-Nearest Neighbors (KNN): 88.31% accuracy

V. DEPLOYMENT

5.1 Hardware Platform

The heart complaint vaticination system can be stationed on standard tackle platforms suitable for machine- literacy operations. The tackle conditions depend on the size of the dataset, complexity of the model, and real- time vaticination needs.

1. particular Computers(PC/ Laptop)

A standard PC or laptop with amulti-core processor(Intel i5/ i7 or original) and at least 8 GB RAM can efficiently run models like Logistic Retrogression, KNN, or Random Forest.

2. Garçon Platforms

For hospitals or larger- scale deployments, a garçon with advanced RAM(16 – 32 GB) and amulti-core processor is recommended.

Can handle multiple contemporaneous prognostications and larger datasets.

3. Software- Hardware Integration

- The system can be stationed using Python- grounded fabrics like Flask or Streamlit, which interact with the tackle platform to give a stoner-friendly interface.
- Druggies can input medical parameters, and the tackle executes the model to give real- time prognostications. The proposed system is flexible and can be stationed on PCs or waiters depending on scale and conditions, furnishing real- time heart complaint threat prognostications to support clinical opinions.

5.2 Software Platform and Libraries

The software terrain specifies the tools and platforms needed to develop and run the prophetic system. The setup for this design includes

- Processor Intel Core i5 or advanced
- Development Tools Jupyter Notebook, Visual Studio Code, Anaconda
- Programming Language
- Libraries and fabrics Flask, pandas, NumPy, scikit- learn, XGBoost, fix, and other applicable Python packages.

This software setup ensures smooth prosecution of machine literacy workflows and facilitates model development, evaluation, and deployment.

5.3 Visualization of Results

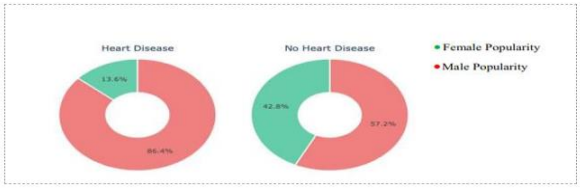
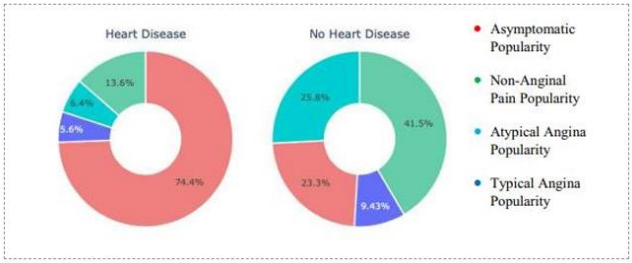


Figure: Shows the presence of heart attack based on Gender

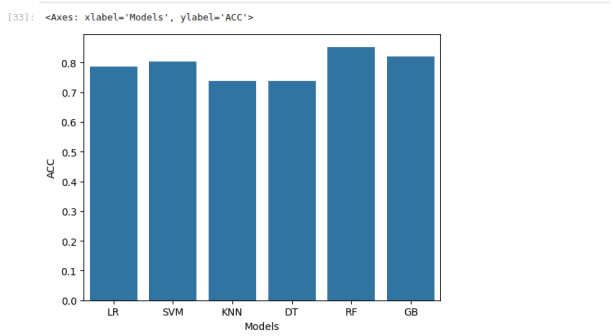
Figure: Shows the presence of a heart attack based on chest pain



5.3 Deployment Process

The deployment process involves making the heart complaint vaticination system functional and accessible for real- time use by croakers and cases. First, the best- performing machine literacy model is perfected and saved using serialization ways similar as Pickle or Joblib, allowing it to be reused without retraining. Next, a stoner-friendly interface is developed, where druggies can input patient details like age, blood pressure, cholesterol, casket pain type, and ECG results. The system processes this data through the trained model to induce prognostications and display threat chances. The stationed system is tested with sample data to insure delicacy, trustability, and smooth operation. also, it can be periodically streamlined with new patient data to retrain the model and ameliorate performance.

5.4 Result



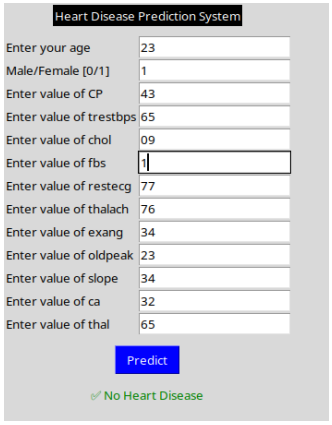


Figure: Prediction & Model

VI. CONCLUSION

Heart complaint vaticination using machine literacy provides a important, data- driven approach to identify individualities at threat of cardiovascular problems. By assaying crucial health pointers similar as age, blood pressure, cholesterol situations, casket pain type, and ECG results, the system can prognosticate the liability of heart complaint with high delicacy. The proposed system improves upon traditional individual styles by automating data analysis, reducing mortal error, and furnishing harmonious, real- time prognostications. It supports early discovery, which is critical for timely treatment and preventative care. also, visualization of results makes prognostications interpretable, allowing croakers and cases to understand the factors contributing to threat. Overall, the design demonstrates that integrating machine literacy into healthcare can enhance individual effectiveness, give decision support for clinicians, and contribute to better case issues. With farther development, similar systems have the eventuality to come essential tools in preventative cardiology and substantiated healthcare.

VII. FUTURE SCOPE

The heart complaint vaticination system can be enhanced and expanded in several ways to increase its utility, delicacy, and connection in real- world healthcare settings

Integration with Real- Time Health Monitoring

The system can be linked with wearable bias or IoT- enabled health observers to continuously collect data similar as heart rate, blood pressure, and ECG signals.

Real- time monitoring enables early discovery and timely cautions for cases at threat.

Addition of Larger and Different Datasets

Expanding the dataset to include cases from different demographics and geographic regions can ameliorate model conception and delicacy.

Multi-center datasets can help the system prisoner a wider variety of heart complaint patterns.

Objectification of Advanced Algorithms

Deep literacy models, similar as Convolutional Neural Networks(CNNs) for ECG signals or mongrel models combining irregular and signal data, can ameliorate prophetic performance.

Ensemble and mounding styles can further enhance trustability and delicacy.

Mobile and Cloud Deployment

Developing mobile operations or pall- grounded platforms can make the system fluently accessible to croakers and cases, indeed in remote areas.

pall deployment also allows nonstop updates and scalability for multiple druggies.

Resolvable AI and Interpretability

Integrating resolvable AI ways can help croakers understand why the system predicts high or low threat, adding trust and relinquishment.

point criterion, SHAP values, and rule- grounded explanations can make the prognostications more transparent.

Beforehand Intervention and Preventive Healthcare

unborn systems can give individualized recommendations for life changes, diet, or medical checks grounded on the prognosticated threat.

Prophetic perceptivity can shift healthcare from reactive treatment to preventative care.

REFERENCES

1. Suad Alhojely," A scalable summarization system using robust NLP", 2020 International Conference on Computational Science and Computational Intelligence(CSCI), 978-1-7281-7624-6/ 20/$ 31.00 © 2020 IEEE.
2. Breck Baldwin and Thomas S. Morton,"Multi-Document Summarization using judgment birth for stoner query", Appl. Sci. 2022, 12(9), 4479.
3. Single Document Text Summarization Algorithm using semantic similarity, International Journal of Computer operations(0975 – 8887) Volume 17 – No. 2, March 2011.
4. A. Siddharthan, A. Nenkova, and K. McKeown. Syntactic simplification for perfecting content selection inmulti-document summarization. In Proc. of COLING, 2004.
5. L. Vanderwende, H. Suzuki, and C. Brockett. Microsoft Research at DUC2006 Task- concentrated summarization with judgment simplification and verbal expansion. In Proc. of DUC, 2006.
6. Jaya Jayashree Jagadeesh," judgment birth Grounded Single Document Summarization", Composition, January 2005, Research Gate.
7. Logan Lebanoff † Kaiqiang Song † F, Scoring judgment Singletons and dyads for Abstractive Summarization, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, runners 2175 – 2189