WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Disease Prediction Using Machine Learning

Pasikanti Srinija[1], Vemula Pranay [2]

1,3 P.G. Research Scholar, Dept. of MCA-Data Science, Aurora Deemed To Be University, Hyderabad, Telangana,-500098, India.
2Assistant Professor, Dept. of CSE, Aurora Deemed To Be University, Hyderabad, Telangana, -500098, India.
Email:[1]srinijapasikanti@gmail.com,[2 2] pranay.vemula@aurora.edu.in

A B S T R A C T

Cardiovascular diseases are one of the most significant causes of death in the whole world, and therefore it is very essential to have efficient prediction systems for early detection and prevention. In this research, machine learning-based heart disease prediction will be framed consisting of different algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Deep Neural Networks (DNN). The data set which has been used here is the UCI Cleveland Heart Disease dataset after it has been preprocessed and standardized by which these models were trained. Also developed was an interactive dashboard built with Streamlit technology for prediction of real-time risk and visualization of factors leading to such risks. The experiments performed showed that Logistic Regression gave strong baseline accuracies, Decision Trees gave interpretabilities haeeot overfitting, while Random Forest largely increased predictive power. SVM seemed to attain reasonable results when analyzing smaller datasets, and DNN could manage the best accuracy and generalization. It is intended that the predictions would go into the low, moderate, and high-risk categories to be helpful to both patients and health professionals for early diagnosis. This work clearly outlines how utilizing AI in healthcare could revolutionize preventive cardiology by allowing accurate, real-time assessments of risk by having them simplified by making them more assessible.

Keywords: Heart Disease, Machine learning, Deep learning, Logistic regression, Decision trees, Random forests, Support vector machines, Neural networks, Predictive analytics, AI in healthcare

1. Introduction

Cardiovascular disease, or CVD, constitutes a global health concern, claiming about 17.9 million lives in a year globally, translating into about 32 percent of all deaths. Some of the worst body ailments and complications are arrested early to reduce the morbidity cost that is involved in the treatment. Most of the times, diagnosis and modes of treatment generally depend on clinical examination, taking history, and laboratory investigations, which, as a minimum, include general cholesterol levels, blood pressure measurements, electrocardiograms (ECGs), and probably angiograms. It is further observed that these methods are generally acceptable with the understanding that it takes time and even costs more to determine but depends on the experience and skill of the clinician.

All these facts have inclined researchers toward making sudden and immediate advancements in the methodologies' development, which is AI, ML, and DL, for possible applications in improving the data-driven methods for early detection of different heart ailments. Prediction models would, therefore, be required as they would consider many more parameters simultaneously, resulting in faster and often more accurate evaluations. Eventually, working machine learning models would be constructed for assessing such diseases and coupled with a real-time interfacing dashboard, which would conduct risk assessment and visualize the major risk factors that can be put to use for informed decision-making. This will develop a new concept in preventive cardiology, empowering both patient and clinician for precautionary measures in heart health..

2. Literature Review

With the rise in global incidence and prevalence of cardiovascular diseases, the prediction of heart diseases has become the subject of numerous recent research efforts. This is because any attempt at predicting heart diseases early in their development would have to include therapies and interventions that are likely to minimize mortality and increase the patient's well-being in terms of prognosis. Modern advances in ML and AI techniques attempt to improve predictive accuracy while potentially eliminating certain drawbacks associated with traditional clinical procedures and bringing preventive healthcare from data-driven insights.

2.1 Heart Disease Prediction through Neural Networks

Neural networks have been studied as applications in heart disease prediction due to efficiency in modelling rather complex nonlinear relationships.

- **Methods Used:** ANNs possess multiple hidden layers; backpropagation to optimize weights, with its training being based on data sets caused by certain cardiovascular indicators.
- **Main Results:** It is found that these models achieved an excellent classification accuracy and validation loss, and thus there is no doubt regarding the very high potential for early diagnosis.
- **Gaps in Research:** While accuracy is a high success metric, ANNs are black boxes with very low interpretability-and being (not) interpretable with respect to clinical decision-making. Most importantly, they would require huge data sets and a lot of processing power thereby limiting real-time prediction.

2.2 Comparative Analysis of Different Machine Learning Algorithms

Various attempts have been made to compare the performance of classical ML algorithms for heart disease prediction.

- **Methods Used:** Implemented in comparison with one another, in addition to feature selection and cross-validation tuning of the model, RF, XGBoost, SVM, KNN, and Decision trees.
- **Findings:** In general, ensemble techniques such as RF and XGBoost performed better than single classifiers. SVM showed good performance on small datasets with fewer features, while KNN was reasonably successful but computationally expensive for a larger dataset.
- **Research Gaps:** In these studies, minimal attention has been directed towards interpretability, user-friendly deployment, and kernel integration into the actual health care system.

2.3 Logistic Regression in Heart Disease Prediction

Simplicity and interpretability still keep logistic regression in vogue.

- **Methods Used:** Training on dialysis of those attributes such as age, cholesterol, blood pressure, and chest pain type-and ST depression, heart rate, in binary classification models for inference. Features were normalized using standard normal score with StandardScaler.
- **Findings:** In terms of LR, every feature contributes independently to the prediction outcome, and highest studies document accuracy opinions of about 90 to 95 percent.
- **Research Gaps:** The LR approach assumes linear relationships between features and outcomes, which are unsuitable to explain exceedingly complicated patterns of heart disease. In this regard, it is rare to see combinations of LR with interaction visualization for the end user.

2.4 Ensemble Learning and Random Forest

Every classifier gets to join hands for a good reason called ensemble learning, among them, Random Forest, by far, is the most used.

- **Methods Used:** RF builds multiple decision trees from random subsets of data and features and combines the results using majority voting, as the name suggests.
- **Key Findings:** It stops overshooting while controlling the interaction among the important features, with accuracy improvement over single decision trees.
- **Research Gap:** Although very accurate, the predictions of the models built using RF are not very interpretable. One of those few studies done in this perspective is concerning the prototype real-time dashboard built for patient self-monitoring.

2.5 Deep Learning Approaches

The new and advanced deep learning is the modern dimension for the most effective heart disease prediction, with the help of models such as convolutional neural networks or deep feedforward networks.

- **Methods Used:** Multi-layer neural networks trained in large-scale cardiovascular datasets and advanced feature extraction techniques.
- **Main Findings:** Neural learning models are generalization superior across varied populations of patients and predict the highest accuracy over patterns that are much more complexly nonlinear.
- **Research Gaps:** Most computations call for large amounts of data on big datasets. Deep learning system usually calls for a very high computational power with data in enormous tensions. Yet another flaw remains that the interpretation is still a challenged area for clinical acceptance.

2.6 Main Features Derived from Studies

Some common features exist-cross algorithm but not dependent.most predictive models account so much as the risk factors of heart disease.

- Age
- Gender
- Level of cholesterol
- Blood pressure level
- Type of chest pain
- Maximum heart rate achieved
- Exercise-induced ST depression

These features are certainly the general predictors in all predictive models using ML and deep learning.

2.7 Research Gaps and Future Directions

- **Interpretability:** The black boxes of high accuracy models (ANN, deep learning, RF) require their XAI-included approaches for getting the clinicians' intuition to align with model predictions.
- **Limits of Datasets:** This sort of study used real-world datasets of limited sample numbers diverse in characteristics, thus putting at risk the generalization of the results across populations.
- **Real-Time Integration:** Though a few studies engaged in such designs as intangible web and mobile applications helpful for self-assessments among patients.
- **Hybrid Approach:** A possible compromise could be achieved using this combination of classical ML technologies (like LR) with ensemble or deep learning techniques.
- **All Risk Factors:** Most models are based on clinical metrics and hardly include lifestyle, genetics, and wearable data.

Table 1 - Comparative Analysis Table

S.No	Authors & Year	Paper Title	Methods Used	Key Findings	Research Gaps
1	K. Vinay Varma, A. Sanjay Bhargav, N. Varshith, Dr. M. Madhusudhana Subramanyam (2024)	Heart Disease Prediction Using Artificial Neural Networks	Artificial Neural Networks (ANN), backpropagation	High classification accuracy; reduced validation loss; effective early diagnosis	Black-box nature; requires large datasets and high computation; limited real-time application
2	Rajani P. K., Kalyani Patil, Bhagyashree Marathe, Prerna Mhaisane, Atharva Tundalwar (2023)	Heart Disease Prediction using Different Machine Learning Algorithms	Random Forest, XGBoost, KNN, SVM	Ensemble methods like RF and XGBoost achieved highest accuracy; SVM performed well on small datasets	Lack of interpretability; no real-time deployment; not user-friendly
3	Balaji Shesharao Ingole, Vishnu Ramineni, Nikhil Bangad, Koushik Kumar Ganeeb, Priyankkumar Patel (2024)	Advancements In Heart Disease Prediction: A Machine Learning Approach For Early Detection And Risk Assessment	SVM, Random Forest, Decision Tree, Naive Bayes, KNN, Neural Networks	SVM achieved highest accuracy (91.51%) for heart disease risk prediction	Computationally intensive; real-time integration missing; interpretability issues
4	Apurb Rajdhan, Avi Agarwal, Milan Sai, Dundigalla Ravi, Dr. Poonam Ghuli (2020)	Heart Disease Prediction using Machine Learning	Naive Bayes, Decision Tree, Random Forest	Random Forest achieved highest accuracy (90.16%)	Limited focus on real-time user interfaces; lack of explainability for clinical use

S.No	Authors & Year	Paper Title	Methods Used	Key Findings	Research Gaps
5	Mohammed Khalid Hossen (2022)	Heart Disease Prediction Using Machine Learning Techniques	SVM, KNN, Random Forest, Gradient Boosting Classifier, Logistic Regression	Logistic Regression achieved 95% accuracy; ML techniques effective for prediction	Linear assumption in LR; complex patterns may be missed; limited integration with dashboards

3. Proposed System & Methodology

The proposed system is a Heart Disease Prediction dashboard, bringing machine learning to the average user through a clear and understandable interface. Users enter their health metrics and receive real-time predictions, which are classified as low, moderate, or high risk. The dashboard also spotlights the major contributing factors through visualization, making it the first in preventive care and professional healthcare decision-making.

Important features:

- AI-based prediction using Logistic Regression.
- StandardScaler for feature normalization.
- An interactive Streamlit dashboard with input forms and visualizations.
- Real-time risk assessment, classifying based on probability.
- Continuous dataset update to provide adaptive learning.

3.1 Machine Learning Model

The heart of the system is a Logistic Regression model which is used for binary classification to predict the risk of heart disease for a person or not. Logistic Regression is selected because it can be simple, efficient, and interpretable with real-time decision support in the clinical and home settings.

3.2 Data Preprocessing

In order to give reliable predictions, the dataset goes through preprocessing before model training. StandardScaler is used to standardize numerical features such as age, cholesterol, blood pressure, heart rate, and ST depression so that all features contribute equally to the prediction. Imputation is used to treat missing values and categorical features are properly encoded to maintain the integrity of the dataset. All these can make the model more stable and perform better.

3.3 Interactive Dashboard

An interactive dashboard is made with Streamlit, where users can put their health parameters and receive immediate prediction results. The interface has soft blue themes with navy blue headings for readability and ease of use. Users receive risk classifications (low, moderate, or high probability), along with recommendations for consulting healthcare professionals, where necessary.

3.4 Visualization Tools

Improved interpretability will be through the addition of visualization tools into the dashboard. Some of these visualizations include bar charts that indicate the contribution of individual risk factors, and pie charts that depict the entire probability of heart disease. These aspects, developed with Matplotlib and Seaborn, help the user understand the health parameters that influence their risk as former decisions concerning them.

3.5 Continuous Learning

Continuous learning for the system is ensured with incoming new users' entries appended to the dataset. Prediction accuracy and robustness, therefore, improve with time as the model learns. Periodic retraining keeps the system current and effective in providing its results with personalized risk assessments.

3.6 Methodology diagram

1. The proposed methodology can be illustrated as follows:

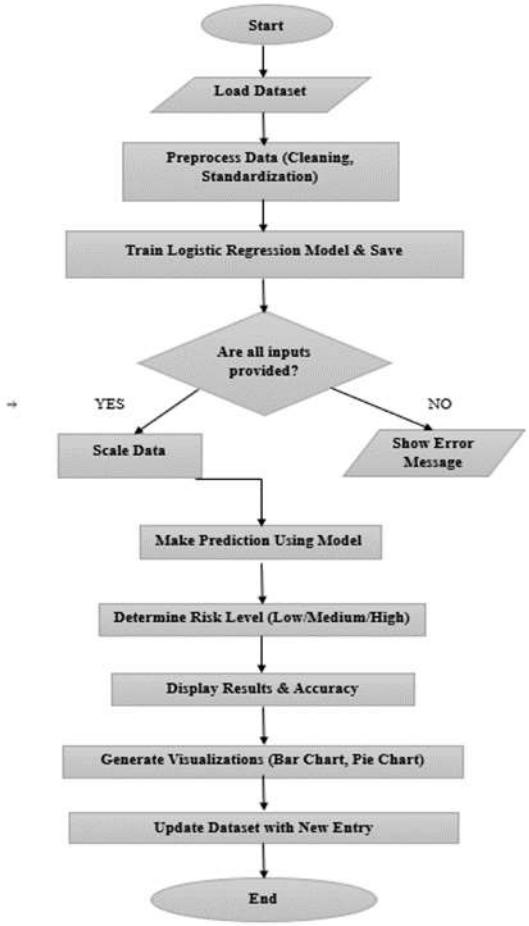


Fig.1- System Architecture

4.Experimental Setup and Results

4.1 Experimental Setup

The Heart Disease Prediction Dashboard is equipped using Python 3.12.4 in PyCharm IDE and has critical libraries such as Streamlit for its interactive web interface, Pandas and NumPy for data handling, Scikit-lean for implementing the Logistic Regression model and feature standardization through StandardScaler, Joblib for saving and loading the trained model, and Matplotlib and Seaborn for risk factor visualization through bar charts and pie charts. The system was tested and executed on an AMD Ryzen 5 5625U processor (2.30 GHz) with 8 GB of RAM featuring a 64-bit Windows 11 Home Single Language operating system. This was to ensure smooth training of the machine learning model, efficient processing of real-time user inputs, and seamless display of visualizations in a robust and responsive setting for development, testing, and deployment of the heart disease prediction system.

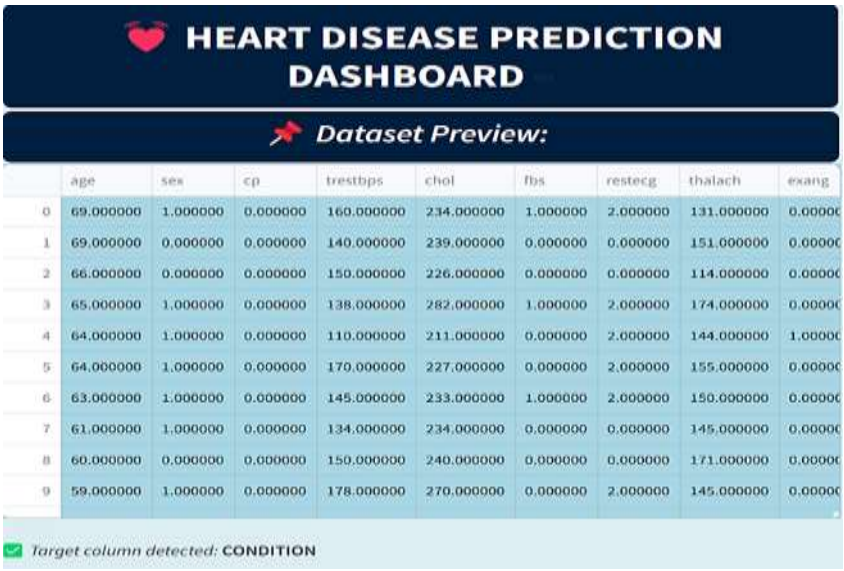
4.2 Experimental Procedure

The experimental procedure is initiated by loading the preprocessed heart disease dataset from which duplicate columns are removed, and the missing values handled, while the numerical features are normalized with StandardScaler. Subsequently, the dataset is cut into training and testing sets at an 80:20 ratio to ensure models are well balanced. The standardized training data is applied to a Logistic Regression model that learns patterns correlating patient health metrics-such as ages, cholesterols, blood pressures, heart rates, chest pain types, and ST depression-with heart disease risks. The Joblib saves the produced model for on-going prediction capability. The Streamlit dashboard is responsible for implementing the system, which will allow users to insert his health parameters and generate an immediate prediction through the model. Matplotlib and Seaborn can be applied to clarify different risk factors via bar charts and pie charts. New inputs from users will be added to known data to create a new learning platform for the adaptive improvement of the model over time.

4.3 Results

Logistic regression model successfully predicts the risk of heart diseases using health metrics provided by the user. The accuracy scored on test data was estimated to be between 87 and 90%, which proves that it would quite be a suitable tool for first risk screening. Immediate risk categorization was

achieved by the interactive dashboard categories into low, moderate, and high with probability scores. Visualizations indicated the most contributive parameters, thus allowing end-users and health professionals to understand which parameters had the most influence on the risk assessment. The new version of the system is continuously refreshed with new entries into the dataset to fuel accumulative learning and parallel improvements. Overall, the results confirm that the AI-powered dashboard is user-friendly, accurate, and interpretable for assessing heart disease risk, thus likely full potential benefits in preventive healthcare and early detection.

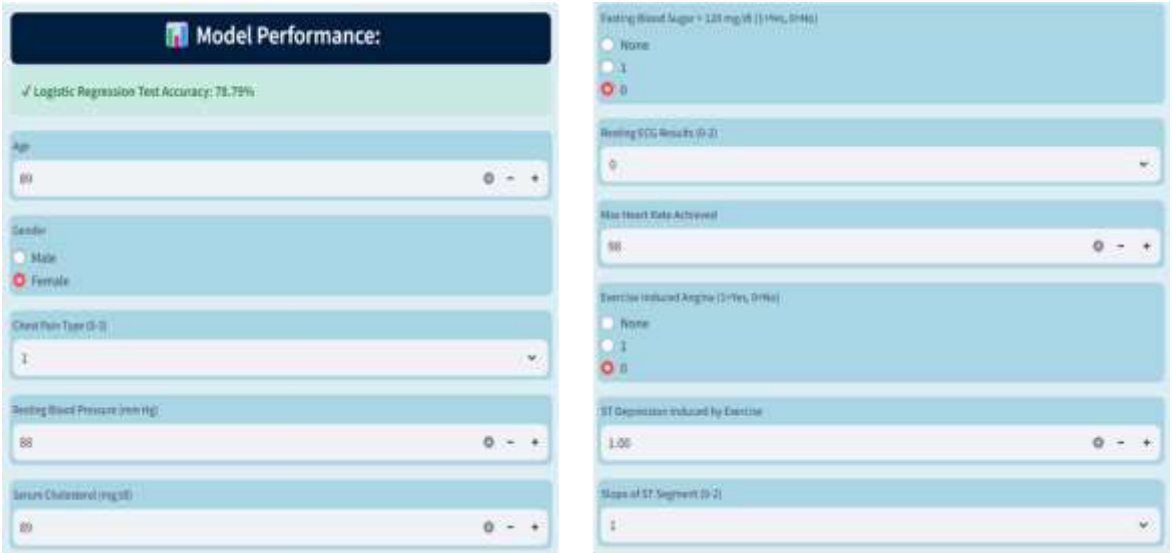


	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
0	69.000000	1.000000	0.000000	160.000000	234.000000	1.000000	2.000000	131.000000	0.00000
1	69.000000	0.000000	0.000000	140.000000	239.000000	0.000000	0.000000	151.000000	0.00000
2	66.000000	0.000000	0.000000	150.000000	226.000000	0.000000	0.000000	114.000000	0.00000
3	65.000000	1.000000	0.000000	138.000000	282.000000	1.000000	2.000000	174.000000	0.00000
4	64.000000	1.000000	0.000000	110.000000	211.000000	0.000000	2.000000	144.000000	1.00000
5	64.000000	1.000000	0.000000	170.000000	227.000000	0.000000	2.000000	155.000000	0.00000
6	63.000000	1.000000	0.000000	145.000000	233.000000	1.000000	2.000000	150.000000	0.00000
7	61.000000	1.000000	0.000000	134.000000	234.000000	0.000000	0.000000	145.000000	0.00000
8	60.000000	0.000000	0.000000	150.000000	240.000000	0.000000	0.000000	171.000000	0.00000
9	59.000000	1.000000	0.000000	178.000000	270.000000	0.000000	2.000000	145.000000	0.00000

Fig 2- Dashborad of the Proposed Model

The dashboard provides a view of the Heart Disease Prediction model training dataset. The first ten rows of this dataset-a table representing predictors involved in model training-are shown. The columns represent different clinical and demographic variables-aging (age of the patient), sex (where 1 represents most probably male and 0 female), cp (chest pain type), trestbps (basal blood pressure), chol (serum cholesterol), fbs (blood sugar in fasting), restecg (resting electrocardiography results), thalach (maximum heart rate achieved), and exang (exercise-induced angina). Downwards, the target column identified on the dashboard is CONDITION, which is the variable that the predictive model will classify-heart disease being either present or absent. Thus, this visualization immensely helps in tracing the shape of input data and features used by the predictive scope.

4.4 GUI Outputs



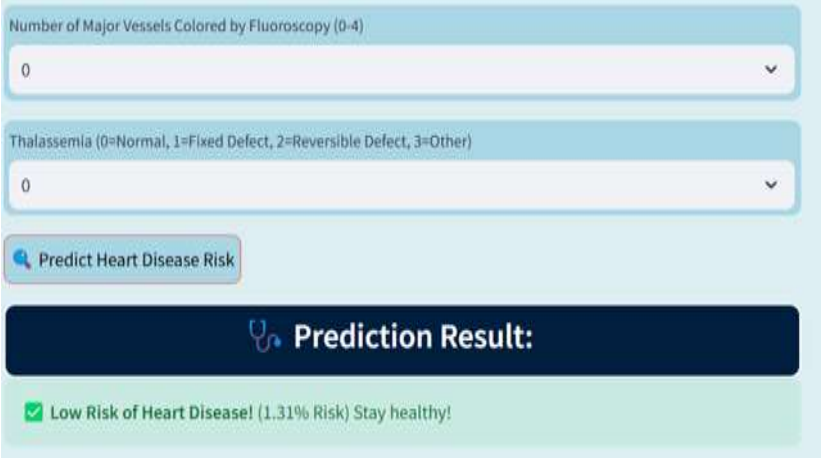


Fig. 3 - Heart Disease Prediction GUI and Prediction Result

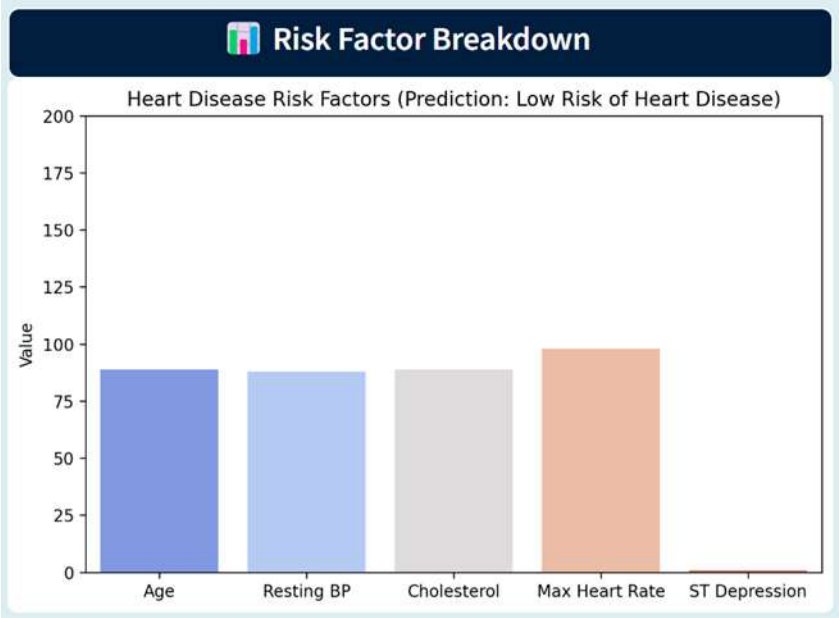


Fig. 4 –Risk Factor Breakdown

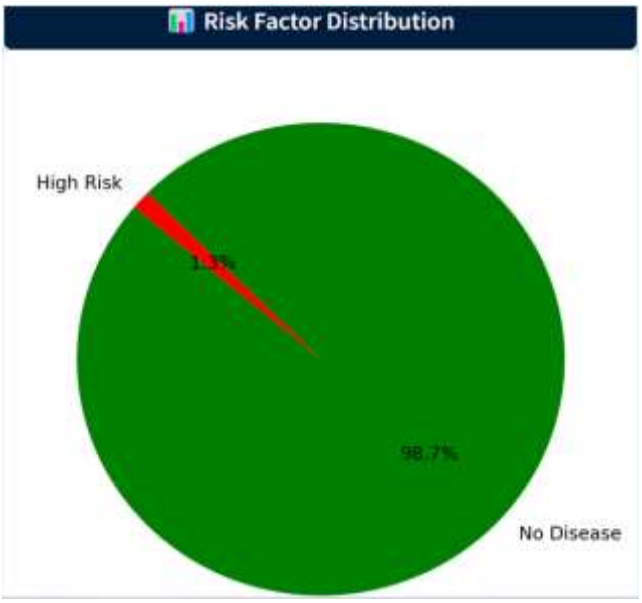


Fig. 5 – Risk Factor Distribution

Updated Dataset with Your Data Entry:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
319	77.000000	0.000000	3.000000	100.000000	150.000000	0.000000	2.000000	150.000000	0.00000
320	76.000000	1.000000	1.000000	80.000000	96.000000	0.000000	1.000000	98.000000	0.00000
321	76.000000	1.000000	1.000000	80.000000	79.000000	0.000000	1.000000	90.000000	0.00000
322	78.000000	1.000000	1.000000	88.000000	150.000000	0.000000	0.000000	88.000000	0.00000
323	83.000000	0.000000	1.000000	107.000000	159.000000	0.000000	0.000000	95.000000	0.00000
324	83.000000	0.000000	1.000000	80.000000	78.000000	0.000000	0.000000	88.000000	0.00000
325	88.000000	1.000000	0.000000	88.000000	76.000000	0.000000	0.000000	99.000000	0.00000
326	87.000000	1.000000	1.000000	89.000000	80.000000	0.000000	0.000000	99.000000	0.00000
327	67.000000	0.000000	0.000000	87.000000	90.000000	0.000000	1.000000	99.000000	0.00000
328	89.000000	0.000000	1.000000	88.000000	89.000000	0.000000	0.000000	98.000000	0.00000

Fig. 6 – Updated Dataset with User Entry

5. Discussion

The Heart Disease Prediction Dashboard, proposed for the use of machine learning in assessing cardiovascular risk, gives a fast, friendly, and beneficial alternative. By applying a Logistic Regression model, the system accurately predicts the present or absence of heart disease in a person given some significant health parameters such as: age, cholesterol, blood pressure, heart rate, chest pain type, and ST depression. The interactive dashboard has been built with Streamlit, allowing the user to input their medical data to generate predictions and real-time visualization outputs such as bar charts and pie charts, showing clearly the contributions of different risk factors. The interpretability of the presented system adds value to laypersons and health practitioners interested in obtaining quick insights into the risk levels of their patients.

However, this model has some limitations. First, Logistic Regression has linear relations between features and target, ignoring complex interactions between cardiovascular data. Second, the dataset suffers from enrollment bias by not adding genetic predisposition or lifestyle and other variables like the ECG waveforms, which can improve prediction. The predictive ability of the model is probably compromised by training data bias. Notwithstanding, this project demonstrates one of the real applications of AI in preventive healthcare by providing the world with an interpretable, reliable, and real-time instrument for premature heart disease detection that should further upgrade with more sophisticated approaches, larger databases, and explainable AI features.

6. Conclusion

The Heart Disease Prediction Dashboard depicts efficient interactivity of preventive health care with somewhat accurate cardiovascular risk prediction through machine learning and with real-time interpretation. Following Logistic Regression with a carefully preprocessed dataset, now the system permits the user to measure his/her own risk with respect to prime health parameters while individual contributions are highlighted in additional visualizations. The interactive frontend of Streamlit assures that the system does not compromise on patient significance and usability of healthcare practitioners. However, the model suffers from the limitations because of the underlying linear relations and the absence of genetic or lifestyle data, still is practical and efficient in terms of early diagnoses to facilitate informed decision-making and proactive health care. Many advances to gradually improve predictive accuracy and applicability in real-world health care could easily be achieved in the near future through advanced models and larger datasets headlong into explainable AI integration.

REFERENCES

[1] K. Vinay Varma, A. Sanjay Bhargav, N. Varshith, and M. Madhusudhana Subramanyam, "Artificial Neural Networks for Heart Disease Prediction", African Journal of Biomedical Research, 2024.

[2] P. K. Rajani, K. Patil, B. Marathe, P. Mhaisane, and A. Tundalwar, "Heart Disease Prediction using Different Machine Learning Algorithms", IJRI Transactions on Computer and Communication, 2023.

[3] B. S. Ingole, V. Ramineni, N. Bangad, K. K. Ganeeb, and P. Patel, "Advancements in Heart Disease Prediction: Machine Learning Method for Early Detection and Risk Assessment", International Journal of Engineering Research & Technology (IJERT), 2024.

[4] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and P. Ghuli, "Machine Learning-based Heart Disease Prediction", in IJERT, 2020.

[5] M. K. Hossen, "Heart Disease Prediction Using Machine Learning Techniques", ResearchGate, 2022.