International Journal of All Research Education and Scientific Methods (IJARESM),
ISSN: 2455-6211, Volume 13, Issue 12, December-2025, Available online at: www.ijaresm.com

Heart Disease Risk Prediction

Mr. Rajsingh Thakur[1], Mr. Sidram Patil[2], Mr. Moizuddin Siddiqui[3], Mr. Mohsin Khan[4], Ms. S. S. Wagre[5], Ms. N. L. Pariyal[6]

[1,2,3,4]B. Tech Final year Students, Dept. of Computer Science and Engg, MGM's College of Engineering, Nanded, Maharashtra, India
[5,6]Guide, Asst. Prof. (M.E.), Dept. of Computer Science and Engg, MGM's College of Engineering, Nanded, Maharashtra, India

----------**********----------

ABSTRACT

Cardiovascular diseases pose a significant global health burden, highlighting the need for accurate and early risk prediction. Given the increasing volume of healthcare data, machine learning techniques offer effective solutions for predicting heart disease by analyzing lifestyle factors that significantly influence cardiovascular health. This research introduces a stacking ensemble machine learning method, which utilizes multiple conventional models as base learners and a bagging or boosting model as the meta-learner. A voting-based feature selection technique is implemented to pinpoint essential lifestyle predictors. Comparative experiments using both complete and reduced feature sets confirm the method's efficacy, achieving an accuracy of up to 91.94%, thereby supporting its value in heart disease risk prediction. This ensemble model provides superior performance metrics compared to individual classifiers, ensuring reliable clinical utility. The emphasis on lifestyle factors makes the model highly interpretable, offering actionable insights for preventive care planning. The developed framework is deployed as a user-friendly application, aiding clinicians and individuals in proactive risk management.

Keywords: Cardiovascular diseases (CVD), Electronic health records (EHRs), Machine learning (ML), Coronary heart disease (CHD), Electrocardiogram (ECG), Body mass index (BMI), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs).

INTRODUCTION

Cardiovascular diseases (CVD), a collection of disorders affecting the heart and blood vessels—including coronary artery disease, heart failure, and complex arrhythmias—constitute the single most significant source of morbidity and premature mortality worldwide. This pervasive public health crisis extends beyond individual suffering to impose a substantial economic burden on global healthcare systems. The costs associated with long-term specialized treatment, emergency care interventions, and frequent re-admissions for CVD patients are monumental, stressing national budgets and resources.

The increasing ubiquity of electronic health records (EHRs) and large-scale public health datasets has resulted in an Immense volume of cardiovascular data. To effectively utilize this rich resource, traditional statistical methods are often inadequate, as they struggle to model the highly complex, non-linear interactions inherent in human physiology and behavior. Consequently, sophisticated machine learning (ML) methodologies have emerged as powerful data analytics tools, offering enhanced capabilities for pattern recognition and risk modeling compared to conventional approaches.

A critical phase in the ML model development pipeline is feature engineering and selection. Given the high dimensionality of clinical and lifestyle data, employing techniques to select the most relevant attributes is essential. These feature selection methods serve a dual purpose: they dramatically improve the computational efficiency and robustness of the predictive models, and, crucially, they enhance model interpretability by isolating the most influential physiological and lifestyle risk factors. This research is specifically motivated by the need to create a highly accurate and clinically applicable machine learning framework focused on predicting an individual's risk of coronary heart disease (CHD). Our methodology utilizes a comprehensive and diverse set of health-related and lifestyle variables, going beyond standard clinical markers. Ultimately, the predictive outputs of this model are intended to function as critical decision support tools, aiding clinicians in implementing targeted preventive care strategies and contributing to a more data-driven, lifestyle-focused paradigm for cardiovascular disease management.

Machine learning algorithms enhance this process by learning complex patterns that are difficult to detect through traditional clinical assessment alone. Accurate risk prediction enables timely lifestyle modifications and medical interventions. Thus, such systems play a vital role in reducing disease progression and improving long-term cardiovascular outcomes.

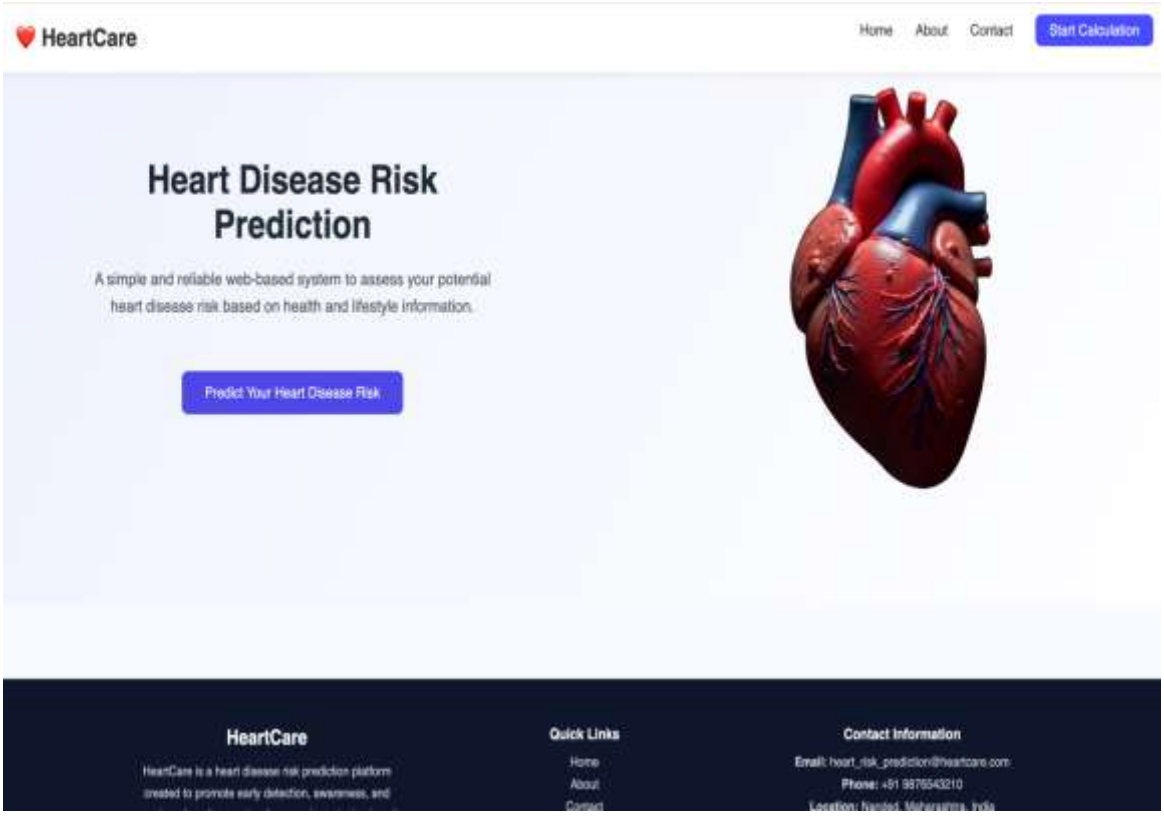

Fig. 1: Home Page

Heart Disease Risk Prediction Framework: Detailed System Architecture

The Heart Disease Risk Prediction framework is a structured, end-to-end system that converts raw health data into actionable medical insights. The system architecture is divided into three major layers, each performing a specific and critical role in accurate risk assessment.

1. Data Input Layer:
The Data Input Layer is responsible for collecting and organizing patient-related information from multiple sources. It includes clinical data obtained from electronic health records such as age, gender, blood pressure, cholesterol levels, blood sugar, ECG readings, and medical history, along with lifestyle-related factors like diet, physical activity, smoking, alcohol consumption, stress, and sleep patterns. This layer ensures that data from different formats is properly integrated and checked for completeness and consistency before being passed to the next stage.

2. Machine Learning Processing Layer:
The Machine Learning Processing Layer acts as the analytical core of the system where meaningful insights are extracted from the input data. In this layer, preprocessing techniques are applied to handle missing values, reduce noise, normalize data, and encode categorical variables. Feature selection methods are then used to identify the most significant clinical and lifestyle attributes, improving model efficiency and interpretability. An ensemble machine learning model analyzes the selected features to learn complex and non-linear relationships associated with heart disease risk.

3. Output, Decision, and Recommendation Layer:
The Output, Decision, and Recommendation Layer transforms model predictions into practical outcomes for healthcare use. It generates a heart disease risk score and categorizes individuals into different risk levels such as low, moderate, or high risk. The layer also provides clinical decision support by offering data-driven insights to assist healthcare professionals in early diagnosis and intervention. Additionally, it delivers personalized lifestyle and care recommendations to encourage preventive measures and better cardiovascular health management.

International Journal of All Research Education and Scientific Methods (IJARESM),
ISSN: 2455-6211, Volume 13, Issue 12, December-2025, Available online at: www.ijaresm.com

LITERATURE SURVEY

Machine learning techniques have significantly advanced in heart disease prediction, starting with basic statistical and rule-based models. Early methods, such as linear classifiers and tree-based approaches, offered initial insights into clinical risk factors but were limited in their ability to represent complex relationships in medical data. As research progressed, more advanced supervised algorithms—including probabilistic models, margin-based classifiers, and instance-based learners—were introduced to improve prediction performance. These models were typically tested on publicly available cardiovascular datasets, where their effectiveness was heavily influenced by preprocessing and feature selection strategies, leading to variations in reported outcomes. Recent focus has shifted towards ensemble-based learning methods that combine multiple models to achieve more reliable predictions. By capturing nonlinear interactions among risk factors, these methods have generally outperformed single classifiers. However, most existing systems concentrate on simple binary outcomes. Addressing this limitation, the current study proposes an ensemble framework designed to provide a more informative and interpretable assessment of heart disease risk.

METHODOLOGY

The proposed methodology follows a structured machine learning pipeline to predict heart disease risk with reliability and reproducibility. The process involves dataset preparation, preprocessing, feature selection, model implementation, and performance evaluation.

1. Dataset Description

The dataset includes health records from individuals with varied demographic and lifestyle backgrounds. It contains numerical and categorical attributes related to body composition, general health, dietary habits, smoking, and alcohol consumption. The target variable is heart disease risk, which facilitates supervised classification. The full list of data input fields is shown in Fig. 2.

	Category	Field Name	Input Type	Example / Description
2	Personal Profile	Age	Numeric	e.g., 45
3	Personal Profile	Biological Sex	Dropdown	Male / Female / Other
4	Personal Profile	Height (cm)	Numeric	e.g., 175
5	Personal Profile	Weight (kg)	Numeric	e.g., 70
6	Personal Profile	BMI	Auto-calculated	Calculated from height & weight
7	Medical History	Self-Rated Health	Dropdown	Excellent / Good / Fair / Poor
8	Medical History	Last Checkup	Dropdown	Within 1 year / 1–2 years
9	Medical History	Diabetes Status	Dropdown	Yes / No
10	Medical History	Skin Cancer	Checkbox	Yes / No
11	Medical History	Other Cancer	Checkbox	Yes / No
12	Medical History	Depression	Checkbox	Yes / No
13	Medical History	Arthritis	Checkbox	Yes / No
14	Lifestyle Factors	Regular Exercise	Dropdown	Yes / No
15	Lifestyle Factors	Smoking History	Dropdown	Yes / No
16	Lifestyle Factors	Alcohol (Days/Month)	Numeric	e.g., 0–30
17	Lifestyle Factors	Fruit (Servings/Month)	Numeric	Count
18	Lifestyle Factors	Green Veggies (Servings/Month)	Numeric	Count
19	Lifestyle Factors	Fried Food (Servings/Month)	Numeric	Count

Fig. 2: Data Input Fields

2. Data Preprocessing

Preprocessing steps are applied to enhance data quality and learning performance. Missing values are managed using imputation techniques, and categorical features are converted into a numerical format via encoding. Numerical features are normalized for uniform scaling. Class imbalance is addressed using oversampling methods, and outliers are identified and treated to reduce noise.

3. Feature Selection

Feature selection is conducted to reduce dimensionality and identify the significant predictors of heart disease risk. Multiple selection techniques are combined to remove irrelevant attributes and improve model efficiency. Important lifestyle-related features, such as BMI, body weight, smoking habits, dietary patterns, and overall health status, are retained.

4. Machine Learning Models

A variety of machine learning models are implemented, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, AdaBoost, XGBoost, and Multilayer Perceptron (MLP). Each model is trained independently. A stacking ensemble approach is also used to combine predictions from multiple models for enhanced accuracy and generalization.

5. Model Evaluation

Model performance is assessed using accuracy, precision, recall, F1-score, and confusion matrix analysis. These metrics offer a detailed evaluation of predictive reliability, particularly for medical decision-making. Results from individual models are compared with the ensemble method to determine overall effectiveness in heart disease risk prediction.

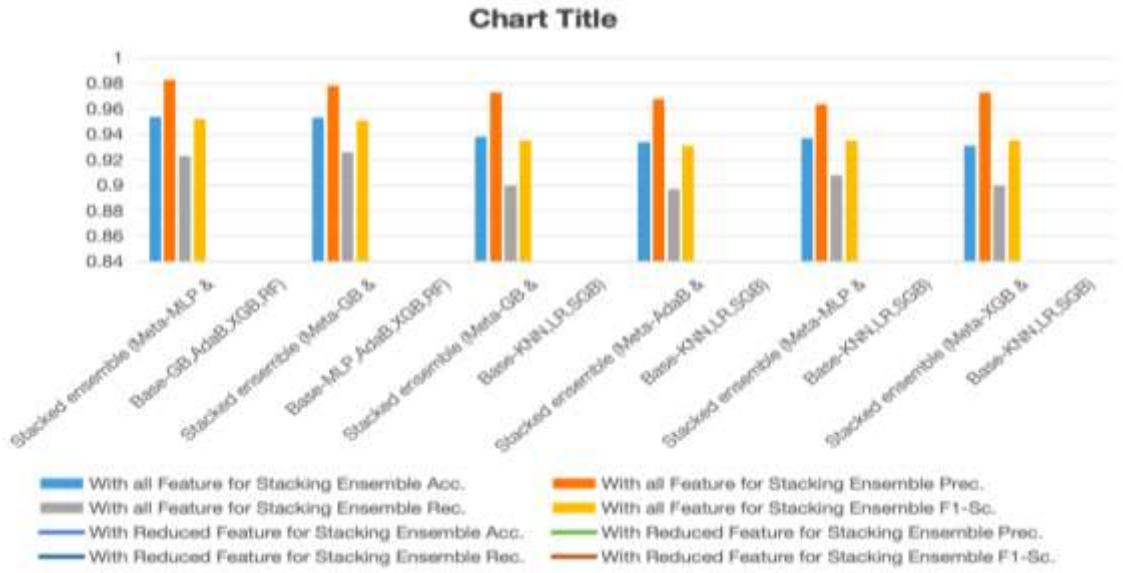

Fig. 3.1: Performance comparison of stacking ensemble models for heart disease risk prediction

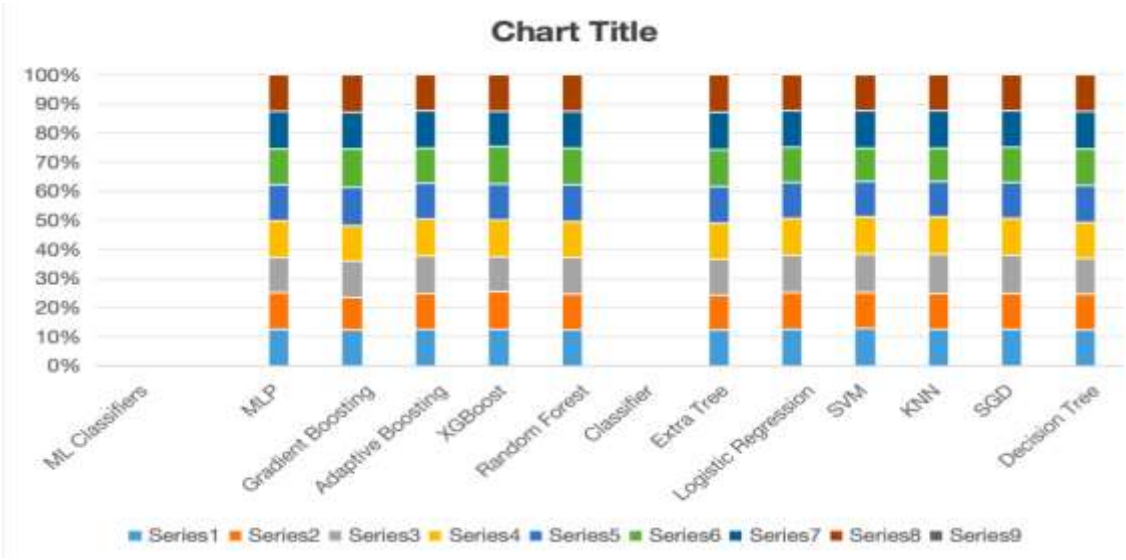

Fig. 3.2: Comparative performance analysis of machine learning classifiers for heart disease risk prediction

IMPLEMENTATION

1. Development Tools and Libraries

The system implementation utilized standard data science libraries for efficient processing and analysis. Pandas and NumPy were used for data manipulation and numerical computations, while visualization tools supported the analysis of data distribution, class imbalance, and feature relationships. Machine learning models were developed and evaluated using Scikit-learn, with XGBoost incorporated to support optimized gradient-boosted ensemble learning.

2. Data Preparation and Splitting

After preprocessing, the dataset was divided into training and testing subsets to evaluate model generalization. The training set was used for model development, and the testing set assessed performance on unseen data. Stratified sampling was applied when necessary to maintain balanced class distributions across both subsets.

3. Model Training and Optimization

Multiple machine learning classifiers were trained separately using the same training dataset to ensure a fair comparison. Hyper-parameter values were optimized using systematic search strategies, and cross-validation was employed to enhance model robustness and reduce overfitting.

Table-I: Comparative Performance Analysis of Different ML Models With and Without Reduced Feature Selection for Stacking Ensemble

ML Classifiers	With all Feature for Stacking Ensemble				With Reduced Feature for Stacking Ensemble			
	Acc.	Prec.	Rec.	F1-Sc.	Acc.	Prec.	Rec.	F1-Sc.
Stacked ensemble (Meta-MLP & Base-GB,AdaB,XGB,RF)	0.954	0.983	0.923	0.952	0.962	0.988	0.935	0.961
Stacked ensemble (Meta-GB & Base-MLP,AdaB,XGB,RF)	0.953	0.978	0.926	0.951	0.961	0.985	0.936	0.960
Stacked ensemble (Meta-GB & Base-KNN,LR,SGB)	0.938	0.973	0.900	0.935	0.953	0.966	0.939	0.952
Stacked ensemble (Meta-AdaB & Base-KNN,LR,SGB)	0.934	0.968	0.897	0.931	0.949	0.958	0.939	0.949
Stacked ensemble (Meta-MLP & Base-KNN,LR,SGB)	0.937	0.964	0.908	0.935	0.952	0.975	0.928	0.951
Stacked ensemble (Meta-XGB & Base-KNN,LR,SGB)	0.931	0.973	0.900	0.935	0.953	0.969	0.936	0.952

Table-II: Comparative Performance Analysis of Different ML Models With and Without Reduced Feature Set

ML Classifiers	With All Feature Set				With Reduced Feature Set			
	Acc.	Prec.	Rec.	F1-Sc.	Acc.	Prec.	Rec.	F1-Sc.
MLP	0.847	0.857	0.833	0.845	0.851	0.833	0.878	0.855
Gradient Boosting	0.884	0.819	0.891	0.885	0.933	0.954	0.910	0.932
Adaptive Boosting	0.828	0.812	0.852	0.832	0.815	0.795	0.848	0.821
XGBoost	0.939	0.969	0.908	0.938	0.939	0.969	0.906	0.937
Random Forest Classifier	0.932	0.930	0.934	0.932	0.952	0.962	0.942	0.952
Extra Tree	0.918	0.908	0.930	0.919	0.957	0.946	0.968	0.958
Logistic Regression	0.813	0.804	0.828	0.816	0.788	0.771	0.819	0.794
SVM	0.829	0.816	0.849	0.832	0.786	0.759	0.837	0.796
KNN	0.849	0.809	0.913	0.858	0.803	0.763	0.879	0.817
SGD	0.813	0.795	0.843	0.818	0.788	0.771	0.818	0.794
Decision Tree	0.890	0.878	0.905	0.891	0.914	0.905	0.924	0.915

4. Ensemble and Stacking Strategy

An ensemble learning approach was implemented to improve predictive performance. Predictions from multiple base learners were combined using a stacking framework, where a meta-classifier was trained to optimally merge the base model outputs, leading to improved accuracy and reduced variance.

5. Performance Evaluation

Model performance was assessed using standard metrics including accuracy, precision, recall, and F1-score. Confusion matrix analysis was conducted to examine classification errors in detail. Experimental results showed that ensemble and stacked models consistently outperformed individual classifiers in heart disease risk prediction. Experimental observations confirmed that ensemble and stacking-based approaches achieved superior predictive performance, demonstrating improved generalization, reduced variance, and higher stability compared to single classifiers.

International Journal of All Research Education and Scientific Methods (IJARESM),
ISSN: 2455-6211, Volume 13, Issue 12, December-2025, Available online at: www.ijaresm.com

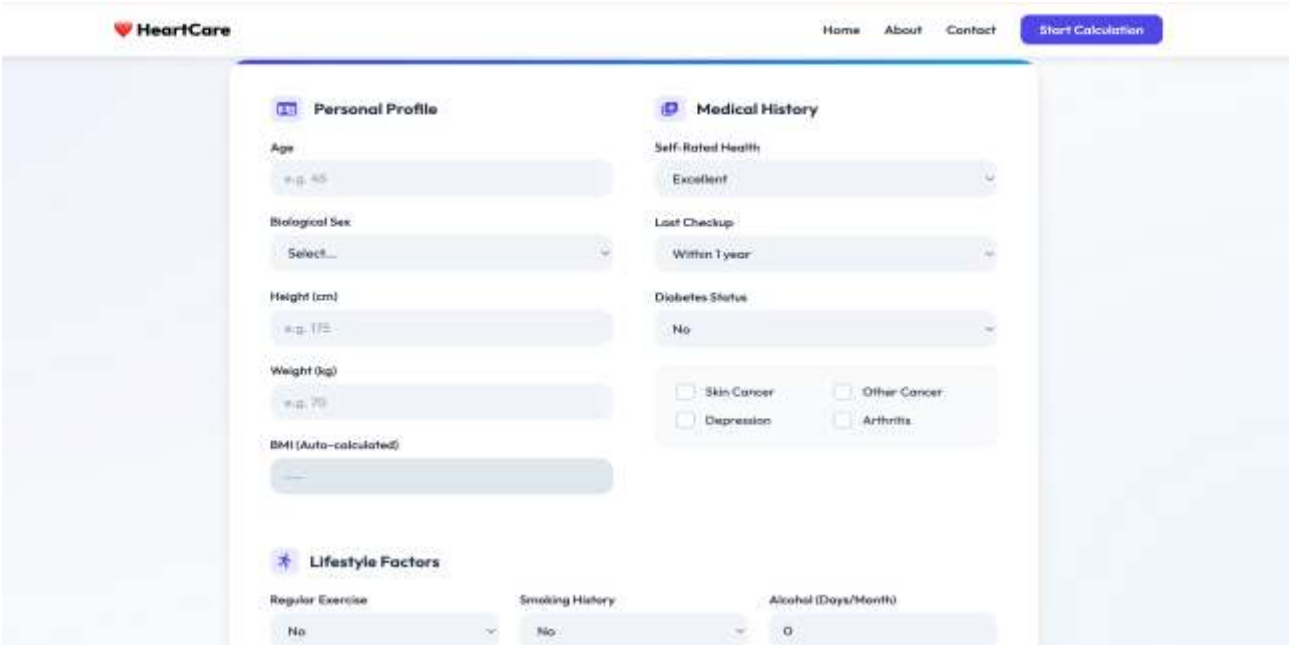

Fig. 4: Heart Disease Risk Prediction Web Page

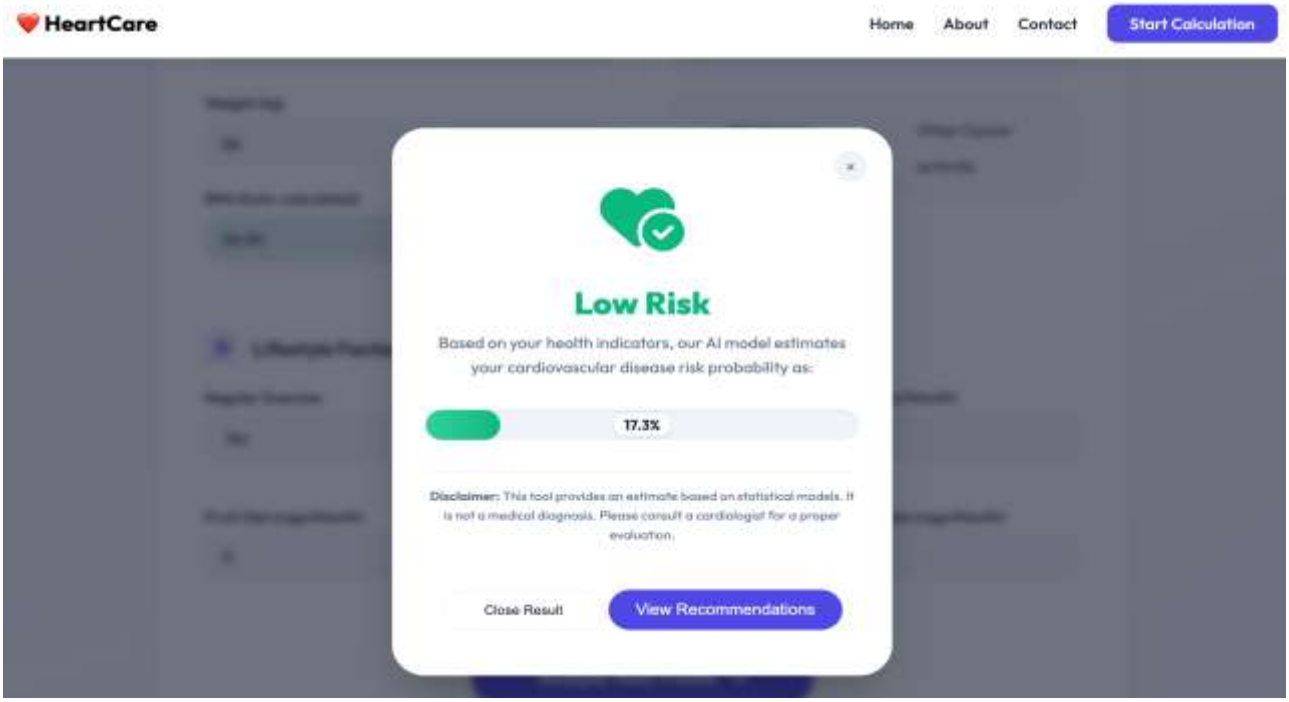

Fig. 5: Result of Predicted Risk

CHALLENGES AND FUTURE SCOPE

While the current system performs well, future research will focus on enhancing its accuracy, adaptability, and clinical relevance by expanding data sources and improving model intelligence.

1. Advanced ML Models: Adopting deep learning techniques such as RNNs and LSTMs will allow the analysis of sequential data to detect complex temporal dependencies in long-term behavioral and physiological changes.
2. EHR Integration: Final implementation will aim for direct integration with Electronic Health Record (EHR) systems and tele-medicine platforms, allowing for automated, seamless risk assessment within standard clinical workflows. These future steps are crucial for transforming the framework into a comprehensive, adaptive, and trusted cardiovascular risk prediction system.

APPLICATIONS

The proposed heart disease risk prediction framework offers wide-ranging utility across the healthcare ecosystem, integrating machine learning with lifestyle data to enable informed decisions and efficient resource use.

1. Early Identification: The system excels at detecting individuals at high risk of heart disease by analyzing behavioral and health indicators before symptom onset. This facilitates timely consultations and lifestyle modifications, curtailing disease progression.

2. Clinical Decision Support: It serves as an intelligent tool for clinicians, providing consistent, data-driven risk estimates that complement medical judgment, reduce diagnostic variability, and support evidence-based treatment plans.

3. Preventive Planning: Healthcare bodies and policymakers can utilize the identified risk patterns to design targeted preventive strategies, including nutrition counseling and fitness programs, aimed at reducing heart disease incidence.

4. Disease Management: The framework can be incorporated into long-term management platforms, providing individuals with continuous feedback on their predicted risk to encourage and track the effectiveness of healthier habits.

5. Risk Assessment: Insurance companies can leverage the system's predictive analytics on lifestyle factors to formulate accurate premium structures and incentivize policyholders to adopt healthier behaviors.

CONCLUSION

This research presents a comprehensive machine learning framework for predicting heart disease risk using lifestyle-related factors. Through effective preprocessing, feature selection, and ensemble learning, the proposed system achieves reliable and interpretable results. The findings demonstrate that lifestyle attributes are strong predictors of cardiovascular risk, emphasizing the importance of preventive healthcare strategies. The study highlights the role of machine learning as a valuable tool in modern healthcare decision support systems.

REFERENCES

[1]. Elias Dritsas and Maria Trigka. 2023. Efficient data-driven machine learning models for cardiovascular diseases risk prediction. (2023), 1161.
[2]. Weiting Huang, Tan Wei Ying, Woon Loong Calvin Chin, Lohendran Baskaran, Ong Eng Hock Marcus, Khung Keong Yeo, and Ng See Kiong. 2022. Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction. Scientific Reports 12, 1 (2022), 1033.
[3]. Zhenzhen Du, Yujie Yang, Jing Zheng, Qi Li, Denan Lin, Ye Li, Jianping Fan, Wen Cheng, Xie-Hui Chen, Yunpeng Cai, et al. 2020. Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: model development and performance evaluation (2020).
[4]. Luis Rolando Guarneros-Nolasco, Nancy Aracely Cruz-Ramos, Giner Alor- Hernández, Lisbeth Rodriguez-Mazahua, and José Luis Sánchez-Cervantes. 2021. Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms (2021), 2537.
[5]. Xi He, B Rajeswari Matam, Srikanth Bellary, Goutam Ghosh, and Amit K Chat-topadhyay. 2020. CHD risk minimization through lifestyle control: machine learning gateway (2020), 4090.
[6]. Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath. 2021. Heart disease prediction using machine learning algorithms. Vol. 1022. IOP Publishing, 012072.
[7]. Meghana Padmanabhan, Pengyu Yuan, Govind Chada, and Hien Van Nguyen. 2019. Physician-friendly machine learning: A case study with cardiovascular disease risk prediction (2019).
[8]. M. Abdar et al., "A new ensemble learning approach for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 141219-141236, 2020.
[9]. S. D. Reddy, B. P. Reddy, and P. Suresh, "Heart disease prediction using machine learning algorithms," in *Proc. Int. Conf. Computing, Communication and Automation (ICCCA)*, IEEE, 2019, pp. 153-157.