# Coronary Heart Disease Risk Prediction: A Comparative Study of Multiple Classifiers

Kunal B. Kirtak
*Electronics Engineering Department*
*Ramdeobaba University, Nagpur, India*
kirtakkb@rknec.edu

Saurabh R. Raut
*Electronics Engineering Department*
*Ramdeobaba University, Nagpur, India*
rautsr_1@rknec.edu

***Abstract*—Coronary heart disease (CHD) remains a leading cause of morbidity and mortality worldwide, necessitating effective risk assessment tools for early detection and prevention. This project aims to predict the ten-year risk of developing CHD using machine learning techniques. Utilizing a dataset containing various health indicators, including demographic, behavioral, and clinical factors, we implement multiple classification algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, to identify significant predictors of CHD risk. Through rigorous data preprocessing, feature selection, and hyperparameter tuning, we evaluate model performance based on metrics such as precision, recall, F1-score, and accuracy. The results demonstrate the potential of machine learning models to enhance predictive accuracy and facilitate informed clinical decision-making. This work underscores the importance of integrating advanced analytical techniques in public health initiatives to improve cardiovascular health outcomes.**

## I. Introduction

Coronary heart disease (CHD) is a major public health concern, accounting for a significant proportion of global mortality rates. It is characterized by the narrowing or blockage of coronary arteries, primarily caused by atherosclerosis, which leads to decreased blood flow to the heart muscle. According to the World Health Organization, CHD is responsible for approximately 9 million deaths annually, highlighting the urgent need for effective prevention and early detection strategies.

Traditional risk assessment methods for CHD, such as the Framingham Risk Score, have provided valuable insights into cardiovascular health; however, they often rely on a limited set of variables and may not capture the complex interactions among numerous risk factors. With the advent of machine learning (ML), there is an opportunity to improve risk prediction models by leveraging vast amounts of health data, which include demographic, lifestyle, and clinical variables. Machine learning algorithms can analyze these complex datasets to identify patterns and relationships that may be overlooked by conventional methods.

This project focuses on developing a predictive model for assessing the ten-year risk of CHD using advanced machine learning techniques. By employing various algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, we aim to uncover significant predictors of CHD risk and evaluate the performance of each model. The findings from this study have the potential to inform clinical practices and guide interventions aimed at reducing the burden of coronary heart disease.

The remainder of this report is organized as follows: we will first review the related work in the field of cardiovascular risk prediction, followed by a detailed description of the dataset and features used in the analysis. Subsequently, we will outline the methods employed in our experiments, present the results and discussion of our findings, and conclude with implications for future research and clinical practice.

## II. Related Work

Advancements in machine learning and big data analytics have significantly contributed to healthcare, particularly in the early prediction and management of chronic diseases such as coronary heart disease (CHD). Prior research, such as a study presented at IEEE TENCON 2019, utilized the Framingham Heart Study dataset with 4,240 records, demonstrating the efficacy of machine learning models—specifically Random Forest, Decision Tree, and K-Nearest Neighbors—in CHD risk prediction. Through a robust preprocessing approach that included handling missing data, resampling, standardization, and normalization, Random Forest achieved the highest accuracy at 96.8%, followed closely by K-Nearest Neighbor and Decision Tree. The study highlighted the importance of preprocessing and the use of K-fold cross-validation for model robustness. Building on this work, our study incorporates additional algorithms such as Logistic Regression, Gradient Boosting, and Support Vector Machines, focusing on a broader range of performance metrics, including precision, recall, and F1-score. By enhancing preprocessing steps and employing cross-validation and hyperparameter tuning, we aim to provide a comprehensive analysis and identify optimal models for predicting ten-year CHD risk.

## III. Dataset and Features

The dataset used in this study, CHDdata.csv, comprises records on individuals with various features related to coronary heart disease (CHD) risk. It contains variables reflecting patients' demographic information, lifestyle habits, and clinical measurements, which are commonly associated with CHD risk. The dataset has the following key features:

### A. *Demographics:*

Gender (male), age (age), and education level (education). Lifestyle Factors: Smoking status (currentSmoker), cigarettes smoked per day (cigsPerDay), body mass index (BMI), and alcohol use.

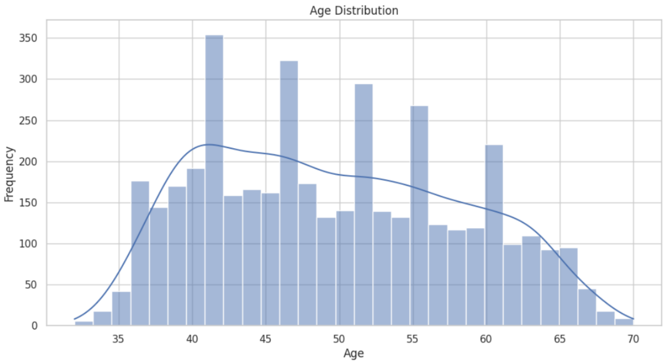

Fig. 1. Feature Distribution Graph

### B. *Medical History:*

Family history of heart disease (family history), history of stroke (prevalentStroke), history of hypertension (prevalentHyp), diabetes (diabetes), and medication for blood pressure (BPMeds).

### C. *Clinical Measurements:*

Total cholesterol (totChol), systolic blood pressure (sysBP), diastolic blood pressure (diaBP), heart rate (heartRate), and blood glucose levels (glucose).

### D. *Outcome Variable:*

The target variable, TenYearCHD, indicates whether an individual developed CHD over the following ten years (1 for CHD and 0 otherwise).

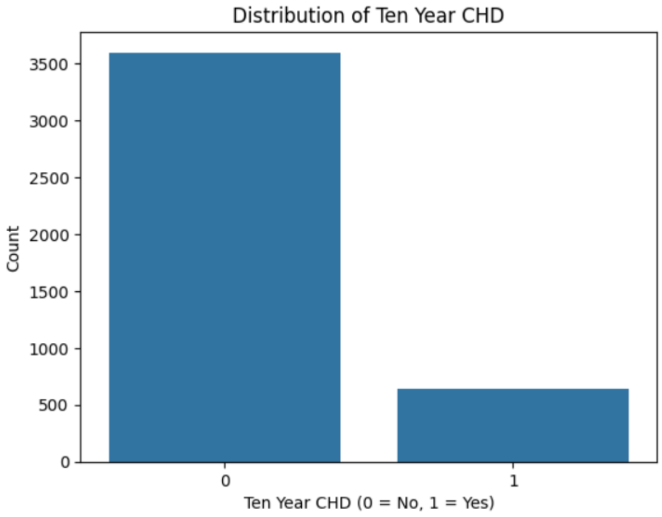

Fig. 2. Distribution Of Ten Years CHD

The dataset, with over 4,200 entries, provides a detailed basis for evaluating the impact of these features on CHD risk prediction. It includes both continuous variables, such as blood pressure and cholesterol levels, and categorical variables like smoking status and family history. This extensive set of features allows for comprehensive analysis and model development to predict CHD risk accurately.

## IV. Methods

In this project, the methodology comprises several key stages, from data preprocessing to model evaluation. Here's an overview of the methods used:

### A. *Data Preprocessing:*

*1) Data Cleaning::* Handled missing values by filling them with median values or by removing rows where appropriate, ensuring completeness in the dataset.

*2) Encoding Categorical Features: :* Transformed categorical features (e.g., male, currentSmoker, prevalentStroke, prevalentHyp, diabetes) into numerical format to make the data compatible with machine learning models.

*3) Feature Scaling: :* Applied standardization and normalization techniques to ensure all features contribute proportionately to the model performance. This transformation brings all numerical features onto a comparable scale, optimizing the training process.

*4) Feature Selection::* Selected relevant features based on domain knowledge and their impact on the target variable (TenYearCHD). Features included in the final model are male, age, cigsPerDay, currentSmoker, prevalentHyp, diabetes,BPMeds, BMI, heartRate, sysBP, totChol, and TenYearCHD .

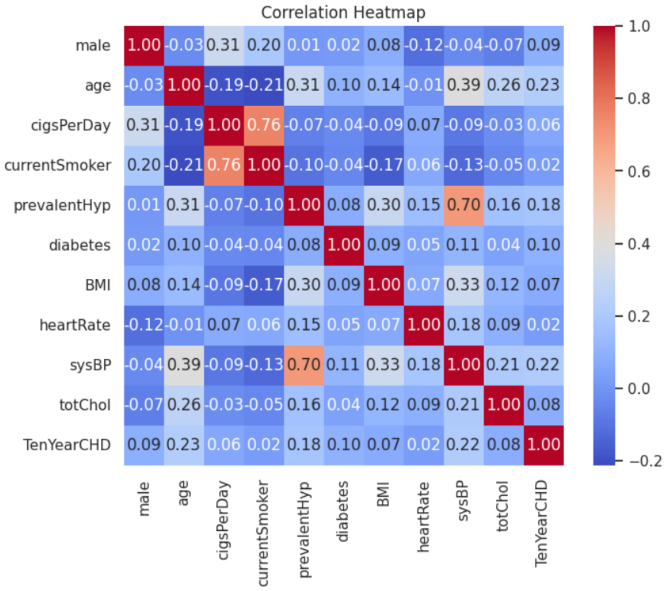

Fig. 3. Correlation Matrix for Selected Features

*5) Train-Test Split::* The dataset was split into training and testing sets, allowing for an 80-20% division. This split enables the model to learn patterns from the training data and generalize on unseen test data, helping evaluate its real-world performance.

### *B. Model Selection and Training:*

Implemented three machine learning models for CHD risk prediction:

*1) Logistic Regression::* Used as a baseline model due to its interpretability and effectiveness in binary classification tasks.

*2) Random Forest Classifier::* Leveraged for its ability to handle feature interactions and its robustness to overfitting, given its ensemble structure.

*3) Gradient Boosting Classifier::* Applied to improve accuracy by building an ensemble of weak learners and iteratively correcting errors in predictions.

### *C. Model Evaluation and Comparison:*

Evaluated each model using metrics like accuracy, precision, recall, and F1-score. These metrics provided insights into the predictive performance and reliability of each model.

### *D. Hyperparameter Tuning:*

Performed grid search-based tuning for Random Forest and Gradient Boosting models to optimize their parameters and achieve better accuracy and predictive power.

### *E. K-Fold Cross-Validation:*

Conducted k-fold cross-validation to ensure the robustness of the models. This technique helps mitigate the effects of data variance by splitting the data into multiple folds and iteratively training on different combinations of training and validation sets.

By implementing this structured approach, we developed models that not only predict the ten-year CHD risk but also highlight the importance of preprocessing and tuning in achieving optimal model performance.

## V. Experiments, Results, and Discussion

The experiments in this study aimed to evaluate the performance of different machine learning models—Logistic Regression, Random Forest, and Gradient Boosting—on the task of predicting the ten-year risk of coronary heart disease (CHD). Each model was assessed on metrics including accuracy, precision, recall, and F1-score to measure its effectiveness in correctly classifying patients at risk.

### *A. Logistic Regression:*

The Logistic Regression model achieved an accuracy of 66%. It showed strong precision for non-CHD cases (88%) but struggled with recall for CHD-positive cases, achieving only 22% precision and a 30% F1-score for this group. This suggests that while the model can identify non-CHD patients reasonably well, it has difficulty detecting true positive CHD cases, leading to a low recall rate of 48% for the CHD-positive class.

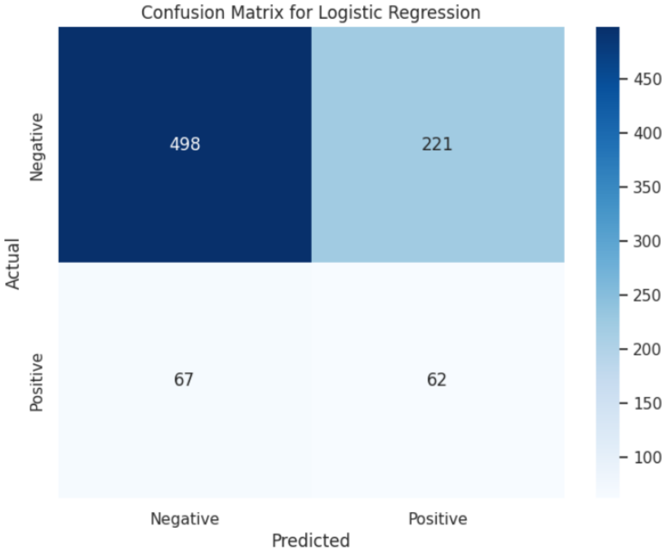

Fig. 4. Confusion Matrix for Logistic Regression

### *B. Random Forest:*

Random Forest performed better overall with an accuracy of 77%. It demonstrated balanced precision and recall for non-CHD cases, with both metrics around 87%, indicating strong performance in identifying patients without CHD. For CHD-positive cases, however, precision and recall remained low (25% and 26%, respectively). Nevertheless, the model outperformed Logistic Regression on accuracy and weighted F1-score, making it more reliable for detecting CHD in the dataset.

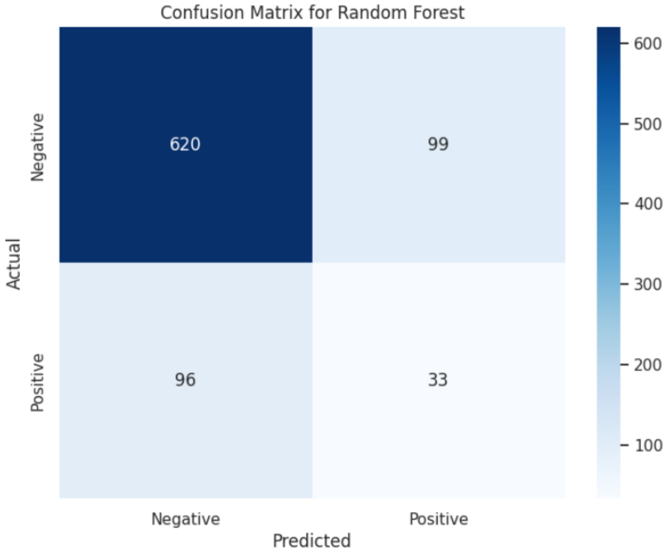

Fig. 5. Confusion Matrix For Random Forest

### *C. Gradient Boosting:*

Gradient Boosting also achieved moderate success, with a slightly lower accuracy of 72% compared to Random Forest. It had a recall of 36% for CHD cases, indicating an improvement in capturing true positives over the Logistic Regression model. However, its precision remained low at 23%, suggesting that while Gradient Boosting has better sensitivity for CHD cases than Logistic Regression, it may still produce false positives.

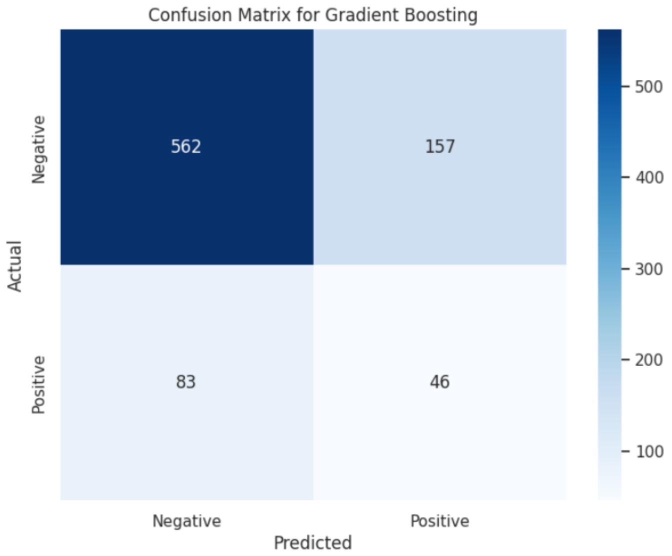

Fig. 6. Confusion Matrix for Gradient Boosting

### D. Hyperparameter Tuning:

Further tuning was conducted on the Random Forest model to optimize its performance. Using grid search, the best parameters were identified as ’max_depth’: None, ’max_features’: ’sqrt’, ’min_samples_leaf’: 1, ’min_samples_split’: 2, ’n_estimators’: 100. With these settings, the tuned Random Forest model maintained a similar accuracy of 77% and exhibited balanced performance across metrics for non-CHD cases, though improvements for CHD cases were limited.

### E. Model Comparison:

Among the models, Random Forest exhibited the highest mean cross-validation F1-score (0.875), followed by Gradient Boosting and Logistic Regression. Despite its stronger performance on average, the Random Forest model’s recall and precision for CHD cases indicate room for improvement, particularly in identifying at-risk patients accurately.

In summary, Random Forest emerged as the best-performing model based on accuracy and F1-score, demonstrating that an ensemble approach can better capture patterns in CHD data compared to Logistic Regression and Gradient Boosting. The overall results suggest that while machine learning models can assist in identifying patients at risk for CHD, further work is necessary to improve sensitivity for CHD cases, potentially by exploring additional features or advanced modeling techniques.

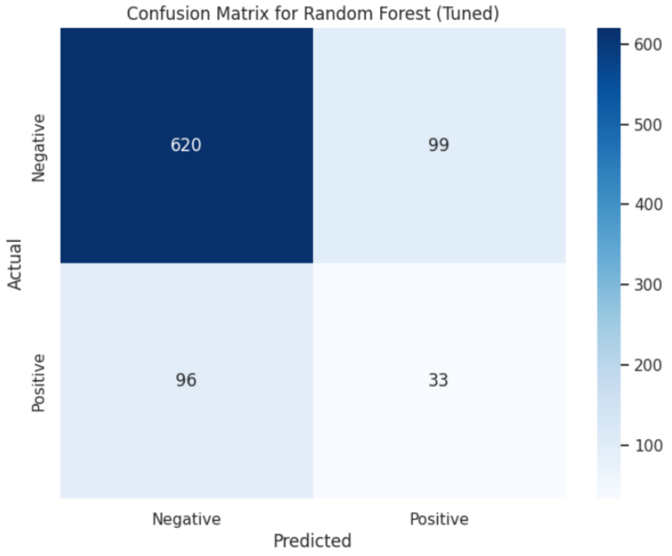

Fig. 7. Confusion Matrix For Random Forest (Tuned)

## VI. Conclusion and Future Work

This study successfully developed a machine learning approach to predict the ten-year risk of coronary heart disease (CHD) using the Framingham Heart Study dataset. Among the evaluated models—Logistic Regression, Random Forest, and Gradient Boosting—Random Forest demonstrated the best overall performance, achieving the highest accuracy and F1-score. This suggests that ensemble-based models may be more effective for CHD prediction, as they handle complex patterns and feature interactions better than simpler linear models. Despite these positive results, the models exhibited limited recall for CHD-positive cases, indicating room for further improvement in sensitivity.

Future work will focus on enhancing the detection of CHD-positive patients by exploring advanced modeling techniques such as deep learning and feature engineering to capture more complex relationships in the data. Additionally, expanding the dataset to include more diverse demographics and additional clinical features may help increase the model’s robustness and generalizability. With further refinement, the proposed approach could contribute to early CHD detection, enabling timely intervention and better patient outcomes.

## VII. Appendices

The following sections provide supplementary information relevant to the project, including detailed explanations of algorithms, preprocessing techniques, and evaluation metrics.

### A. Appendix A: Algorithms Overview

Logistic Regression: A statistical model that uses a logistic function to model a binary dependent variable. It provides a linear approach for classification, suited for datasets with linearly separable classes.

*1) Random Forest::* An ensemble learning method based on decision trees, which aggregates multiple trees to improve accuracy and control overfitting. It randomly selects subsets of features and observations for each tree, enhancing robustness.

*2) Gradient Boosting::* An iterative algorithm that builds weak learners sequentially, where each learner tries to correct the errors of the previous ones. This approach is effective in reducing bias and variance, often yielding high prediction accuracy.

### *B. Appendix B: Data Preprocessing Steps*

Handling Missing Values: Missing values were filled using the median or mode of each feature column to ensure the dataset remained complete for model training.

*1) Standardization and Normalization::* Features were scaled to have a mean of zero and standard deviation of one, improving the convergence of gradient-based algorithms.

*2) Resampling::* Techniques like undersampling or oversampling could be applied in future iterations to address class imbalance and improve recall on CHD-positive cases.

### *C. Appendix C: Evaluation Metrics*

*1) Precision::* Indicates the accuracy of positive predictions, calculated as TP / (TP + FP).

*2) Recall::* Measures the ability to identify all relevant instances, calculated as TP / (TP + FN).

*3) F1-Score::* The harmonic mean of precision and recall, balancing the trade-off between them.

*4) Accuracy::* The proportion of correct predictions out of all predictions, useful for overall performance but limited in imbalanced datasets.

### *D. Appendix D: Hyperparameter Tuning*

The following hyperparameters were tuned using Grid Search on the Random Forest model:

*1) max_depth::* Controls the depth of each tree to reduce overfitting.

*2) max_features::* Defines the number of features to consider at each split, optimized for better performance.

*3) min_samples_leaf and min_samples_split::* Specifies the minimum number of samples required to split an internal node or be a leaf node, balancing depth and computational efficiency.

*4) n_estimators::* The number of trees in the forest, tuned to improve accuracy and robustness.

### *E. Appendix E:Confusion Matrix and Classification Reports*

Detailed confusion matrices and classification reports for each model are available in the supplementary materials, providing insight into model performance on true positive, false positive, false negative, and true negative predictions. These metrics helped identify strengths and limitations in each model's ability to predict CHD.

These appendices provide a comprehensive view of the methodologies and technical details underpinning the study, enabling further reproducibility and extension of the project.

## VIII. Contributions

### *A. Kunal B Kirtak:*

Led data preprocessing, including handling missing values, normalization, and feature selection. Conducted the implementation and fine-tuning of the Logistic Regression and Random Forest models. Contributed to writing the report sections for Dataset and Features, Methods, and Results.

### *B. Saurabh R Raut:*

Focused on the Gradient Boosting model implementation and hyperparameter tuning. Conducted extensive performance evaluations and prepared the classification reports and confusion matrices. Authored the sections on Related Work, Experiments/Results/Discussion, and Conclusion/Future Work.

Both members collaboratively worked on the project structure, reviewed the report, and participated in the interpretation and discussion of results. Together, they designed the overall approach and contributed equally to the analysis and findings presented in this study.

## References

[1] Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. The Lancet, 383(9921), 999-1008. https://doi.org/10.1016/S0140-6736(13)61752-3

[2] Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837-1847. https://doi.org/10.1161/01.CIR.97.18.1837

[3] Kannel, W. B., & McGee, D. (1979). Diabetes and cardiovascular risk factors: the Framingham study. Circulation, 59(1), 8-13. https://doi.org/10.1161/01.CIR.59.1.8

[4] Akella, A., & Gibert, K. (2018). Machine learning for health informatics in cardiovascular diseases: a survey. International Journal of Healthcare Management, 11(2), 106-115. https://doi.org/10.1080/20479700.2018.1435464

[5] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE, 12(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

[6] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[7] Amin, M. S., Eng, H. K., & Akter, M. (2019). Coronary heart disease prediction using data mining techniques with effective feature selection. Journal of Computer and Communications, 7(1), 88-96. https://doi.org/10.4236/jcc.2019.71009

[8] Aslam, M. S., & Mubeen, M. S. (2019). Prediction of coronary heart disease using machine learning algorithms: comparative analysis based on accuracy and cross-validation. Journal of Biomedical Engineering and Medical Imaging, 6(2), 7-14. https://doi.org/10.14738/jbemi.62.6094

[9] Framingham Heart Study. (n.d.). The Framingham Heart Study. Retrieved from https://framinghamheartstudy.org

[10] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer.