Comparative Study on Coronary Heart Disease Prediction Using Five Machine Learning Models

Haikun Guo
College of Science and Engineering, James Cook University, Cairns, Australia

Keywords: Coronary Heart Disease, Machine Learning, Random Forest Classification, Gradient Boosting.

Abstract: In recent times, Coronary Heart Disease (CHD) has emerged as a significant global public health concern, not only due to its mortality rate but also because of the substantial financial burden it places on healthcare systems. Traditional statistical methods for predicting the onset of CHD have limitations in handling multi-dimensional data and often fail to capture complex interactions among different risk factors. This has created a pressing need for a cost-effective cardiac health monitoring system capable of leveraging large-scale, multi-dimensional data for accurate CHD predictions. In this research, the author employs machine learning (ML) techniques to address this discrepancy. The author designs and perform a comparative analysis of five ML classifiers: Logistic Regression (LR), K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), and Gradient Boosting (GB) for robust CHD forecast. These classifiers were rigorously tested on a dataset that includes various risk factors contributing to CHD. Performance metrics were employed to evaluate the effectiveness of each model, including accuracy, specificity, and sensitivity. The results demonstrate that ML classifiers, particularly Random Forest and Gradient Boosting, show high efficacy in predicting CHD, thereby confirming the potential of ML in augmenting cardiac health monitoring systems. This study has far-reaching implications in preventive healthcare, as it offers a pathway to early diagnosis and effective management of CHD.

1 INTRODUCTION

Globally, coronary heart disease (CHD) remains a primary contributor to both mortality and morbidity. The disease manifests through a variety of mechanisms such as arterial plaque buildup, inflammation, and blood clot formation, which can obstruct blood flow to the heart, leading to life-threatening conditions like heart attack or heart failure. Despite medical advancements, accurate early prediction of CHD remains a challenge, hindering preventative and therapeutic measures. While traditional statistical data analysis may struggle to draw inferences from multidimensional parameters, Machine Learning (ML) models are equipped to handle a wide array of variables and effectively discern underlying patterns (Kumar Thakkar et al 2020). Moreover, the advent of modern lifestyle factors, including high cholesterol levels, smoking, and obesity, has only intensified the urgency for effective predictive models for CHD. Therefore, this research aims to address this critical healthcare gap by employing machine learning techniques to estimate the likelihood of CHD based on diverse medical factors.

To achieve a precise and reliable predictive model, the author has implemented five machine learning algorithms: Logistic Regression, K-Nearest Neighbour (K-NN), Random Forest Classifier, Decision Tree Classifier, and Gradient Boosting Classifier. Each of these models was trained and tested on a dataset comprising various features that have been identified as significant contributors to CHD, such as systolic blood pressure (`sysBP`), glucose levels, age, cigarettes smoked per day (`cigsPerDay`), total cholesterol (`totChol`), diastolic blood pressure (`diaBP`), and others. A feature selection method based on chi-square statistics was used to identify the top ten features that have the strongest influence over CHD occurrence. The author then employed hyperparameter tuning methods to fine-tune the most promising classifiers to optimize their performance.

The objective of this research is to evaluate the effectiveness of these predictive models in forecasting CHD outcomes and pinpoint the most precise model. The findings from this research could be pivotal in

Guo, H.
Comparative Study on Coronary Heart Disease Prediction Using Five Machine Learning Models.
DOI: 10.5220/0012800700003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 256-261
ISBN: 978-989-758-705-4

advancing the clinical identification of individuals at high risk of developing CHD, thus facilitating early intervention strategies. This paper offers a comprehensive analysis, aiming to enrich the current understanding in medical data science and set the groundwork for subsequent research and practical strategies to address CHD. In summary, this research is not merely an academic study but a timely intervention in understanding and predicting a disease that has widespread implications for public health.

2 RELATED WORK

Ramya G. Franklin and B. Muthukumar introduced a detailed framework for analyzing cardiovascular conditions through advanced analytic approaches (Franklin and Muthukumar 2020). Their multi-staged approach not only aimed at effective early diagnosis by examining various risk parameters but also ensured data security through Advanced Encryption Standard (AES). A. Lakshmanarao, A. Srisaila, and T. Srinivasa Ravi Kiran addressed the pressing global issue of cardiovascular diseases. The authors introduced an ensemble classifier model specifically designed for heart disease prediction, utilizing two different datasets from Kaggle and UCI for validation (Lakshmanarao et al 2021). The study suggested that the ensemble model notably outperformed existing solutions. Priyanka Gupta and D.D. Seth focused on the crucial task of early detection of Cardiovascular Diseases (CVDs) (Gupta and Seth 2022). To this end, the authors explored the efficacy of various Machine Learning classifiers. Finally, the research developed a system aimed to streamline medical care by saving physicians' time and reducing treatment costs. Meghavi Rana, Mohammad Zia Ur Rehman, and Srishti Jain leveraged the burgeoning amount of medical data to utilize artificial intelligence techniques for analyzing respiratory conditions (Rana et al 2022). The authors noted the importance of their work for medical practitioners and researchers seeking to predict heart disease based on a patient's age.

Ignatious K Pious, K Antony Kumar, Y.Cephas Soulwin, and E.Nipun Reddy addressed the pressing issue of heart disease (Pious et al 2022). The authors concluded that early diagnosis was crucial in managing heart disease, which had been exacerbated by today's sedentary lifestyles and stress. Reldean Williams, Thokozani Shongwe, Ali N. Hasan, and Vikash Rameshar focused on the critical global health issue of heart diseases (Williams et al 2021). The paper underscored the vital role of early prediction in enabling preventative measures and suggested that incorporating additional variables like family history could further improve model performance. A. Ordonez focused on leveraging association rules for forecasting cardiovascular disorders (Ordonez 2006). The paper addressed two primary challenges: the generation of an excessive number of medically irrelevant rules and the lack of validation on an independent test set. The study confirmed that the use of search constraints and validation techniques significantly reduced the number of irrelevant or poorly generalizing rules, providing a set of high-accuracy predictive rules.

Xiaoming Yuan, Jiahui Chen, Kuan Zhang, Yuan Wu, and Tingting Yang addressed the limitations of existing heart disease prediction models, which often only determined the presence of disease but not its severity (Yuan et al 2022). The authors confirmed that their Bagging-Fuzzy-GBDT model demonstrated outstanding accuracy and consistency in not only detecting the presence but also determining the severity of heart disease. Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava tackled the critical challenge of predicting cardiovascular diseases, a leading cause of death globally (Mohan et al 2019). They proposed a novel machine learning-based model that emphasized feature selection to improve prediction accuracy. Azam Mehmood Qadri, Ali Raza, Kashif Munir, and Mubarak S. Almutairi introduced a novel feature engineering approach to select the most significant patient health parameters (Qadri et al 2023). The study validated the performance of all applied methods through cross-validation.

3 METHODS

The methodology for predicting the onset of coronary heart disease (CHD) in this research paper employs a comprehensive analysis of the Framingham Heart Study dataset using five machine learning models. The experimental steps begin with data collection, then working on data preprocessing, exploratory data analysis (EDA), feature selection, model training and evaluation, and finally, comparative analysis. In the data collection phase, the Framingham dataset is imported, which consists of a variety of variables such as age, sex, cholesterol levels, and smoking status, among others, each contributing differently to CHD risks. The data preprocessing step involves handling missing values, either by imputation or deletion, and normalization or scaling of variables if needed. This ensures that the dataset is fit for further analysis.

The next stage is EDA, where the author gains initial insights into the dataset through statistical summaries and visualizations. Single-variable analysis includes evaluating the distribution of each variable, while dual-variable analysis aims to reveal any correlation or causal relationship between two variables. Multivariable analysis focuses on visualizing high-dimensional relationships. One specific objective during EDA is to identify the variables that appear to have a significant impact on CHD risks, such as the correlation between smoking and CHD, the influence of age, or the relevance of blood pressure and cholesterol levels.

Feature selection follows EDA. Based on insights from the Exploratory Data Analysis (EDA), The author will leverage both automated techniques, such as Recursive Feature Elimination (RFE), and expert opinions to select the key variables for our prediction models. The author have opted for five machine learning models for this study. The Author performed hyperparameter tuning on both the Random Forest and Gradient Boosting models using Randomized Search Cross-Validation. This process optimized several parameters, including tree count, max depth, and minimum samples required for splits and leaves. The author assessed the models using various performance metrics. Additionally, a confusion matrix for each model highlighted the false positives and negatives, offering a deeper understanding of the model's clinical relevance.

Feature selection is pivotal in the methodology. Aside from automated algorithms like Recursive Feature Elimination (RFE), expert judgment is solicited to ensure that the selected features are not just statistically significant but also clinically relevant. This involves consulting healthcare professionals experienced in cardiology and referencing peer-reviewed literature that identifies important risk factors for CHD. By combining algorithmic feature selection and expert opinion, the author aims for a more holistic, clinically-applicable model.

For evaluation, several metrics such as accuracy, recall, F1-score, and the ROC Curve are computed for each model. For model evaluation, it's worth noting that the F1-score was especially considered because it balances the trade-off between precision and recall—an essential attribute in healthcare where both false positives and false negatives can have significant clinical implications. Special attention is given to the interpretability of each model, considering that the models not only need to make accurate predictions but should also be interpretable for healthcare professionals who might use this tool in decision-making. Once individual evaluations are completed, a comparative analysis is performed to rank the models based on their performance metrics. The models are also evaluated for their computational efficiency, which is crucial in real-time applications.

4 EXPERIMENTS

4.1 Data Collection

The initial phase of this research focused on the collection of data from the Framingham Heart Study dataset. The dataset was imported into a Python environment using the Pandas library, serving as the cornerstone for all subsequent analyses. The dataset is comprehensive, encompassing multiple variables that contribute to coronary heart disease (CHD) risk such as age, sex, cholesterol levels, blood pressure, and smoking status. Given its wide usage and credibility in prior research, the Framingham dataset was deemed to be an optimal choice for the analysis.

4.2 Data Preprocessing

Upon importing the data, it underwent a rigorous preprocessing routine to prepare it for machine learning analysis. Missing values were handled through mode imputation, while variables with different scales were normalized using Min-Max scaling techniques, implemented using Python's Scikit-learn library. The aim was to produce a dataset devoid of anomalies that could interfere with how accurate machine learning models are. Through the utilization of technologies such as data visualization, this study conducted outlier detection on the data. While outliers in other numerical columns were found to be essential for the model, the author removed the outliers from totChol and sysBP to enhance the accuracy of the model.

4.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to provide an initial understanding of the dataset's characteristics and distributions. The analysis starts with a univariate examination of each variable in the dataset, both categorical and numerical. For categorical variables, the focus is on the distribution of different categories within each variable. For numerical variables, distribution plots are generated to look at the spread of the data. This sets the stage for bivariate and multivariate analyses. In the study the authors found that :

4.3.1 Univariate Analysis

Some features like BPmeds, prevalentStroke, and diabetes are highly imbalanced in terms of their distribution. Categorical features are mostly binary, but education has four levels. In terms of numerical features, cigsPerDay has a very uneven distribution.

4.3.2 Bivariate Analysis

There was no evident correlation between educational level and the daily cigarette consumption (cigsPerDay). According to the dataset, males showed a slightly higher risk for coronary heart disease, as Fig. 1. The age group of 38-46 has more current smokers.

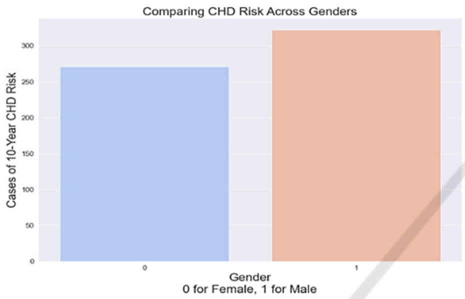

Figure 1: Comparision about CHD Risk Across Genders Plot (Picture credit: Original).

4.3.3 Multivariate Analysis

A minor relationship was found between totChol and glucose. cigsPerDay appears to have almost no relationship with age. The systolic and diastolic blood pressure (sysBP and diaBP) are plotted with respect to whether a person is a current smoker and their gender, and show some relationships there.

4.4 Class Imbalance and Feature selection

This part addresses the issue of class imbalance in a dataset, specifically focusing on the variable TenYearCHD, which presumably represents the likelihood of developing a cardiac condition over the next 10 years. The dataset initially has an unequal number of positive and negative cases, making it difficult for machine learning algorithms to generalize well.

The first step involves separating the positive and negative samples into two different DataFrames. Next, oversampling is performed on the positive samples using the resample function to match the number of negative samples.

After oversampling, the new positive samples are concatenated back with the original negative samples to create a balanced dataset. This new DataFrame is then checked to verify that the number of positive and negative cases is now equal, making the classes balanced.

Finally, the research includes data visualization to confirm this balance. It uses a bar chart to display the number of occurrences for each class and a pie chart to show the percentage distribution of the classes. Fig. 2 and Fig. 3 are bar charts before and after balancing the positive and negative samples respectively.

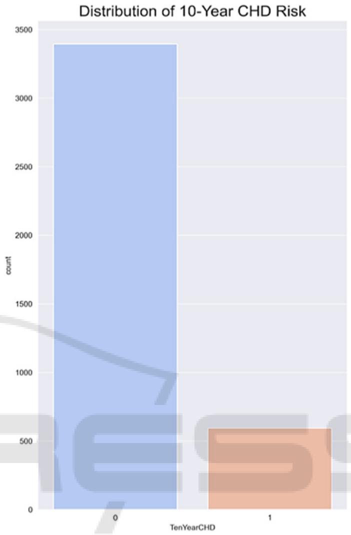

Figure 2: Allocation of Positive and Negative Samples Prior to Equilibrium (Picture credit: Original).

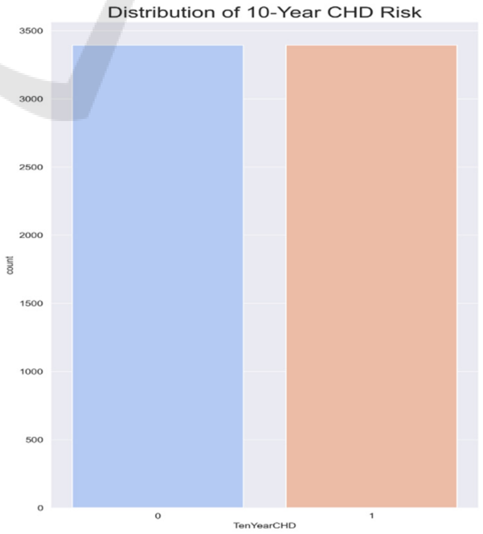

Figure 3: Balanced Distribution of Positive and Negative Samples(Picture credit: Original).

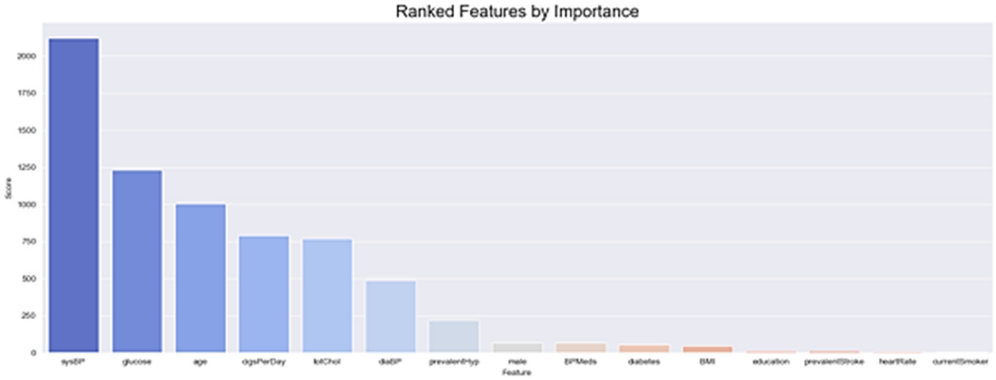

Figure 4: Feature Score Ranking Chart(Picture credit: Original).

4.5 Feature Selection

The feature selection process identifies which variables are most relevant to the target variable, TenYearCHD. Using a chi-squared test through the SelectKBest function, the top 10 features with the highest chi-squared scores were selected. This step is crucial for model performance, as irrelevant or less important features can negatively impact the model's efficiency and predictive power. Fig. 4 shows the feature ordering map.

4.6 Data Splitting and Scaling

The dataset was subsequently divided into training and testing subsets for the purposes of model training and assessment. The train_test_split function was used, allocating 60% of the data to the training set and 40% to the testing set. Lastly, the features were scaled using Min-Max scaling, transforming them to fall within the [0, 1] range.

4.7 Model Training

Five machine learning models were trained on the selected features. Hyperparameter tuning was carried out for Random Forest and Gradient Boosting using Randomized Search Cross-Validation. Various hyperparameters like the number of trees, maximum depth, and minimum samples for splitting and leaves were optimized to improve the models' performances. Table. 1 presents the accuracy of the five machine learning models on the validation set, and table. 2 records the prediction accuracy of the five training models.

Table 1: Comparative Accuracy of Five Machine Learning Models on the Validation Set.

Model	Accuracy
Logistic Regression Model	66.58%
Decision Tree Model	87.29%
K-Nearest Neighbour Model	89.81%
Random Forest Model	95.83%
Gradient Boosting Model	95.47%

Table 2: Five Model Training Accuracy Tables.

Model	Accuracy
Logistic Regression Model	68.35%
Decision Tree Model	87.32%
K-Nearest Neighbour Model	89.84%
Random Forest Model	95.87%
Gradient Boosting Model	95.57%

To thoroughly examine the efficacy of the five machine learning models, the author employed both a validation set and a training set. The accuracy metrics obtained from these datasets provide a comprehensive understanding of each model's predictive capabilities and potential for overfitting.

Table 1 illustrates the comparative accuracy rates of the models when tested on the validation set. The Random Forest model exhibited the top accuracy, registering a score of 95.83%, with the Gradient Boosting model trailing slightly behind at 95.47%. The K-Nearest Neighbour model displayed a notable performance, reaching an accuracy of 89.81%. On the other hand, the Logistic Regression model lagged, recording the least accuracy of 66.58%, suggesting its suboptimal fit for this particular challenge. Table 2 displays the accuracy percentages of the models derived from their respective training sets. The training accuracy follows a similar trend to the validation set. The Random Forest model once more topped the ensemble, achieving a training accuracy of 95.87%, whereas the Logistic Regression model trailed with an accuracy of 68.35%.

By comparing the performance metrics from both tables, it becomes evident that the Random Forest and Gradient Boosting models show a promising capability for robust heart disease prediction, with only marginal discrepancies between training and validation accuracies, thus indicating minimal overfitting.

5 CONCLUSION

In this research, the author examined the efficacy of five machine learning classifiers—Logistic Regression, K-Nearest Neighbour, Random Forest, Decision Tree, and Gradient Boosting—in predicting Coronary Heart Disease (CHD) risk. Utilizing a feature-selected dataset, the models were assessed for both accuracy and interpretability. Random Forest and Gradient Boosting emerged as the most accurate classifiers, recording accuracies of 95.87% and 95.57%, respectively. Despite promising outcomes, the study's limitation lies in its reliance on a single dataset. Future research will aim to validate these models on more diverse datasets to improve predictive accuracy and generalizability, ultimately aiding in the reduction of CHD's societal and economic impact。

REFERENCES

H. Kumar Thakkar, H. Shukla, and S. Patil, “A Comparative Analysis of Machine Learning Classifiers for Robust Heart Disease Prediction,” in 2020 IEEE 17th India Council International Conf. (INDICON), 2020, pp.1-6

R. G. Franklin and B. Muthukumar, “Survey of Heart Disease Prediction and Identification using Machine Learning Approaches,” in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 553-557.

A. Lakshmanarao, A. Srisaila, and T. Srinivasa. R. Kiran, “Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques,” in 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 994-998.

P. Gupta and D. D. Seth, “Comparative Analysis of Machine Learning Classifiers for Accurate and Early Detection of Heart Disease,” in 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2022, pp. 1-5.

M. Rana, M. Z. Ur Rehman, and S. Jain, “Comparative Study of Supervised Machine Learning Methods for Prediction of Heart Disease,” in 2022 IEEE VLSI Device Circuit and System (VLSI DCS), 2022, pp. 295-299.

I. K. Pious, K. Antony Kumar, Y. Cephas. Soulwin, and E. Nipun. Reddy, “Heart Disease Prediction Using Machine Learning Algorithms,” in 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2022, pp. 1-6.

R. Williams, T. Shongwe, A. N. Hasan, and V. Rameshar, “Heart Disease Prediction using Machine Learning Techniques,” in 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 2021, pp. 118-123.

C. Ordonez, “Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction,” IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 2, pp. 334–343, 2006.

X. Yuan, J. Chen, K. Zhang, Y. Wu, and T. Yang, “A Stable AI-Based Binary and Multiple Class Heart Disease Prediction Model for IoMT,” IEEE Transactions on Industrial Informatics, vol. 18, no. 3, pp. 2032–2040, 2022.

S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” IEEE Access, vol. 7, pp. 81542–81554, 2019.

A. Qadri, A. Raza, K. Munir, and M. Almutairi, “Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning,” IEEE Access, vol. 11, pp. 56214–56224, 2023.