Cardiovascular Disease Prediction Using the Framingham Heart Study and Machine Learning Techniques

Omar Magdy Mostafa
1900884
Department of Computer Engineering
Faculty of Engineering
Ainshams University
Cairo, Egypt

AbdELRahman Yasser
19P1492
Department of Computer Engineering
Faculty of Engineering
Ainshams University
Cairo, Egypt

Mostafa Nasrat
19P4619
Department of Computer Engineering
Faculty of Engineering
Ainshams University
Cairo, Egypt

Sherif Essam
16P8248
Department of Computer Engineering
Faculty of Engineering
Ainshams University
Cairo, Egypt

***Abstract*—This paper explores the application of machine learning to cardiovascular disease (CVD) risk prediction using data from the Framingham Heart Study (FHS). This work outlines different data preprocessing techniques, including imputation, and standardization and utilizes machine learning algorithms such as Logistic Regression, K-Nearest Neighbors (KNN) and Random Forest to classify patients as high or low risk of CVD. We demonstrate the performance of different models, and show the importance of a proper preprocessing pipeline. We find that using Random Forest after applying a proper handling of missing values, outlier, and features scaling showed the better results.**

***Index Terms*—Cardiovascular Disease, Framingham Heart Study, Machine Learning, Imputation, Classification.**

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of morbidity and mortality globally, imposing significant burdens on healthcare systems. Early detection and risk assessment of CVDs are paramount for effective preventive interventions and treatments. Traditional risk assessments for CVD often rely on a limited set of clinical and lifestyle factors, which may not fully capture the complex interplay of variables contributing to the disease development. The Framingham Heart Study (FHS), a landmark longitudinal study initiated in 1948, has provided invaluable insights into the epidemiology and risk factors associated with CVD. This study has collected a vast amount of data on various clinical, lifestyle and genetic factors over multiple generations, making it a valuable resource for researchers aiming to understand CVD pathogenesis. In this paper, we harness the rich dataset of the FHS and employ machine learning techniques to develop robust models for CVD risk prediction, aiming to enhance early detection capabilities and aid in better management of this growing global health concern.

II. Background

Traditional cardiovascular disease (CVD) risk prediction methods often use risk scores derived from logistic regression models that assess the probability of developing CVD within a defined time period, typically 10 years. These scores are based on a limited set of risk factors, such as age, sex, blood pressure, cholesterol levels, smoking status, and presence of diabetes, and can be calculated using lookup tables or more complex equations. However, these methods have limitations because they often fail to capture the complex interplay of various risk factors and may not be able to accurately predict CVD risk for all individuals [1], [2], [3].

Machine learning (ML) has emerged as a promising alternative to traditional risk prediction models, offering the ability to learn complex relationships from large datasets and handle heterogeneous data types. Unlike traditional models, ML algorithms do not make strong assumptions about the underlying data distribution and can automatically identify the most important features that are predictive of CVD risk. ML algorithms, such as logistic regression, K-nearest neighbors (KNN) and random forests (RF), have been utilized to achieve great results in terms of CVD prediction using data from the Framingham Heart Study [5], [4], [6]. These techniques have demonstrated strong performance in various classification tasks for medical prediction and diagnosis.

Other research groups have also used data from the Framingham Heart Study to investigate the applicability of machine learning models for CVD risk prediction using the Framingham dataset. For instance, one research group used Logistic Regression, Naive Bayes, SVM and KNN for identifying the best model [5]. Another research group compared the use of Random Forest and Support Vector Machine to predict CVD risk [6]. These previous works motivate the use of ML for CVD detection to be more accurate than traditional statistical methods, and we aim to further explore the potential of these methods using different preprocessing methods and different classifiers.

III. Methods

A. Data Description

The dataset used in this study was obtained from the Framingham Heart Study (FHS), which is publicly available and consists of a large number of health and lifestyle features from 4240 patients. The features include both categorical variables, such as 'male' and 'currentSmoker', and numerical variables, such as 'age', 'cholesterol', 'blood pressure', and other pertinent clinical variables, which are a total of 16 features including the 'TenYearCHD' variable that refers to the dependent variable. The aim of this study is to build machine learning models that can identify patients with higher likelihood of developing a Ten Year Coronary Heart Disease (CHD). It is noted that there is an issue regarding data imbalance of the dependent variable which will be mentioned later, but initially, the dataset consists of 4240 instances, with each instance representing a patient.

B. Data Preprocessing

The dataset underwent several preprocessing steps, as Machine Learning models are not robust to missing or dirty data. In this section, we discuss how we handled these issues.

- **Handling Missing Values:** Missing values were imputed using univariate imputation techniques. For numerical features ('age', 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'), the mean of each column was calculated and used to fill the missing values. For categorical features ('male', 'education', 'currentSmoker', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'TenYearCHD'), the mode (most frequent value) of each column was used to fill the missing values. This approach was chosen due to the simplicity and effectiveness and in accordance to previous research on the dataset.
- **Outlier Removal:** Outliers in the numerical features were identified and removed using the Interquartile Range (IQR) method. The first quartile (Q1), third quartile (Q3), and IQR (Q3-Q1) were calculated for each numerical column. Data points that fall below (Q1 - 1.5 * IQR) or above (Q3 + 1.5 * IQR) were considered outliers and removed from the dataset. This technique will help in removing noisy data that doesn't improve our model's results.
- **Feature Scaling:** After handling the missing values and outlier, we need to scale our numerical features, therefore, MinMaxScaler was used to scale all the numerical features. MinMaxScaler scales the features to a range between 0 and 1, which can be helpful for models sensitive to the scale of features. This transformation enhances the stability and performance of the models.

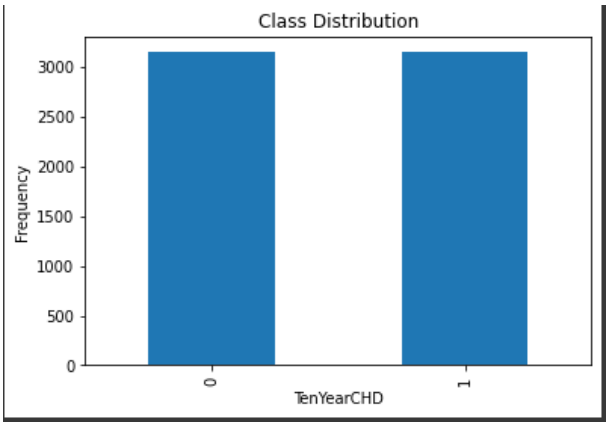

Fig. 1. Distribution of the target variable (*TenYearCHD*) before balancing.

The dataset had imbalanced classes of the dependent variable, *TenYearCHD*, so to overcome this issue and improve the model's ability to predict both low and high risk patients, the Synthetic Minority Oversampling Technique (SMOTE) was applied using a random state of 42, generating the synthetic samples to equalize the classes before training the model. Figure 1 shows the distribution of the target variable before applying SMOTE.

C. Model Development

Three different classification models were implemented for predicting the CVD risk:

- **Logistic Regression:** A simple and effective model that models the probability of a binary outcome (TenYearCHD) using a logistic function. It is a linear model for binary classification, and has shown good performance for this task. The model is trained using the L-BFGS solver and limited to maximum iterations of 1000.
- **K-Nearest Neighbors (KNN):** A non-parametric model that classifies a data point based on the majority class among its k-nearest neighbors in the feature space. KNN was implemented with three neighbors as its parameter (K=3). For the model, we will measure its performance in terms of its ability to separate different class types for cardiovascular disease prediction.

- **Random Forest:** An ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy and robustness, making this classifier more robust to outliers and complex data. Our model is a randomized forest with balanced class weights to make sure that under-represented class has a proper contribution, to avoid the model's bias towards the majority class.

D. Model Evaluation

For model evaluation, the preprocessed dataset was split into two subsets: a training set (80%) used to train the models and a testing set (20%) used to evaluate the performance of the trained models. The split was done randomly, using a random state of 42 to ensure reproducibility.

- **Cross Validation:** Cross-validation was applied using 5-fold cross-validation to estimate the generalization performance of the models on unseen data using the training set. The training data set was partitioned into 5-folds and we trained each model on 4 folds and then tested it on the remaining fold. The average of those 5 scores were calculated and reported for each model. We will use the average of these cross validation scores as a performance metric of our models on unseen data.
- **Accuracy:** After applying the cross validation we trained our model on the entire training data and we calculated model accuracy on the test set. Accuracy measures the percentage of correctly classified instances out of all instances, for example, a value such as 0.85 indicates 85% of the predictions on the test set were accurate.
- **Confusion Matrix:** Confusion matrix was utilized to visualize the performance of models by presenting a breakdown of the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.
- **Classification Report:** Precision, recall and F1 scores will be reported to provide a comprehensive assessment of the classification performance and class-wise predictive ability of each of the models.

IV. Experiments and Results

A. Environment Setup

The environment setup used for our experiments involves a Python 3.9.12 environment, utilizing the pandas library for data handling, NumPy for numerical computations, scikit-learn for machine learning algorithms, imbalanced-learn for handling imbalanced data, and matplotlib for visual representations. The data preprocessing stage includes using FSL for segmentation, registration and skull stripping. The code was written and executed in Google Colab.

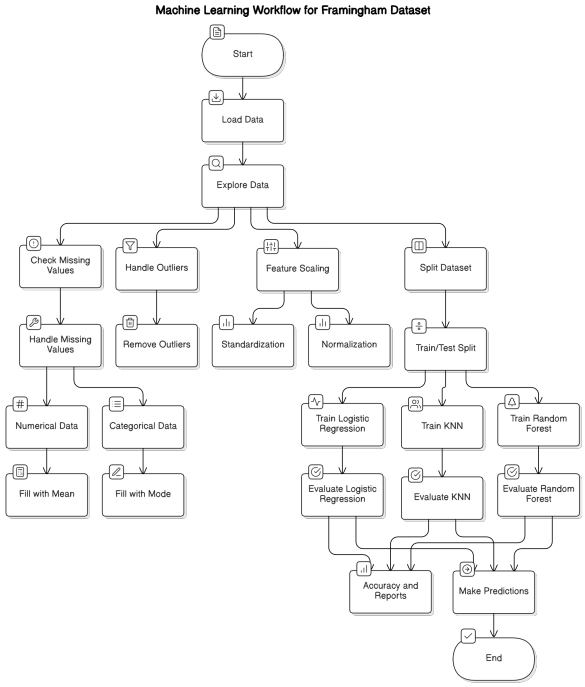

Fig. 2. High-level Overview of Workflow

B. Results

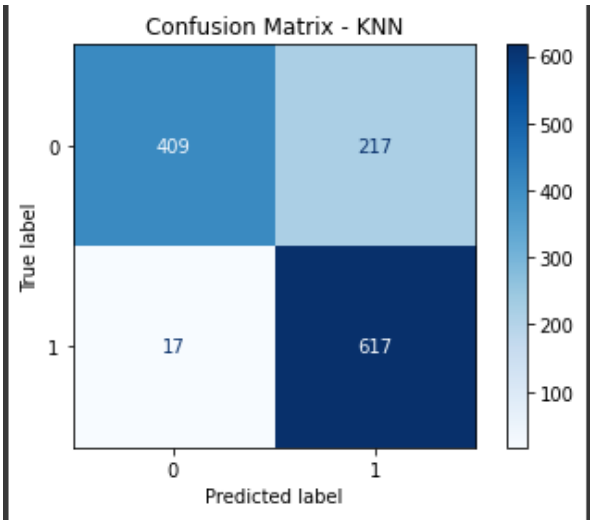

Fig. 3. Confusion Matrix for K-Nearest Neighbors (KNN).

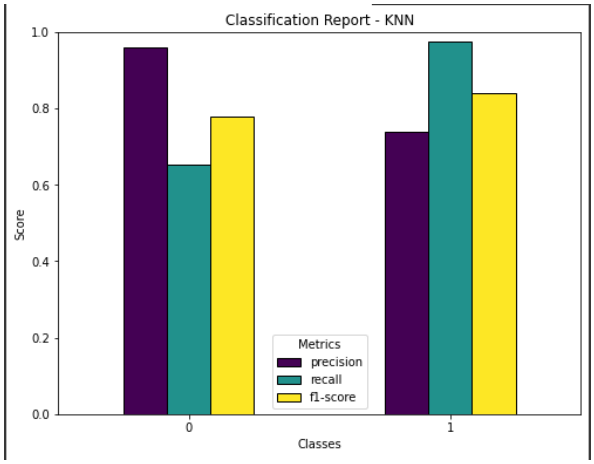

Fig. 4. Classification Report for K-Nearest Neighbors (KNN).

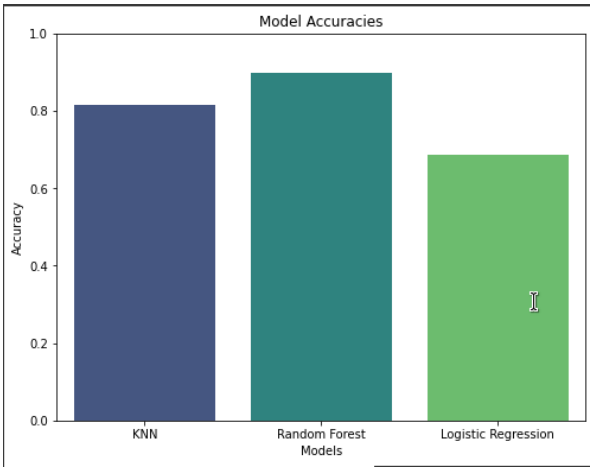

Fig. 7. Comparison of Model Accuracies for KNN, Logistic Regression, and Random Forest.

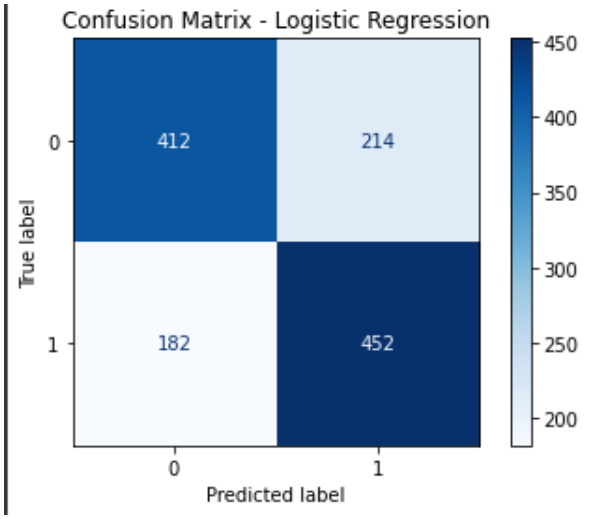

Fig. 5. Confusion Matrix for Logistic Regression (LR).

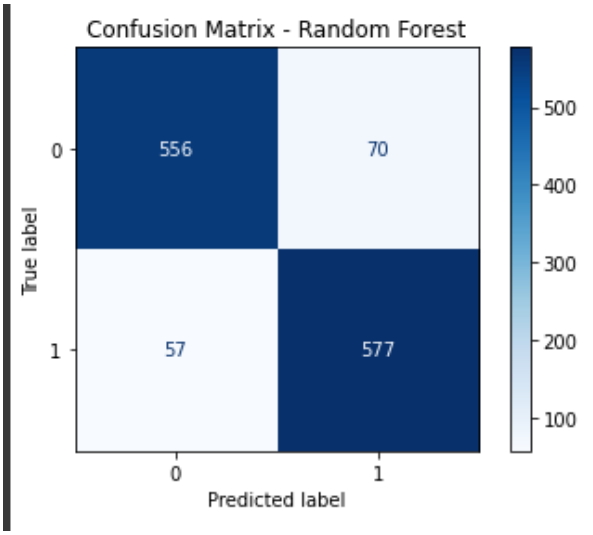

Fig. 8. Confusion Matrix for Random Forest (RF).

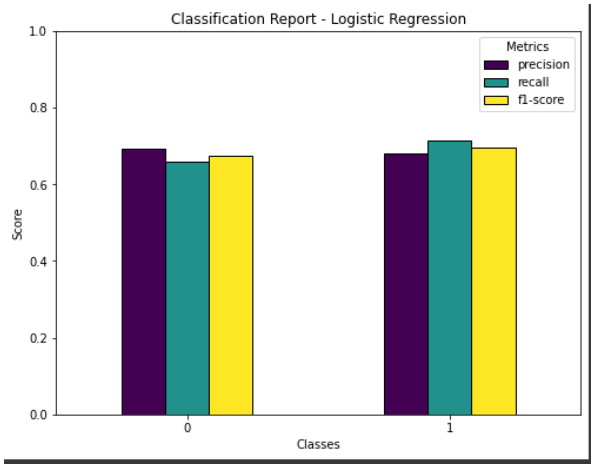

Fig. 6. Classification Report for Logistic Regression (LR).

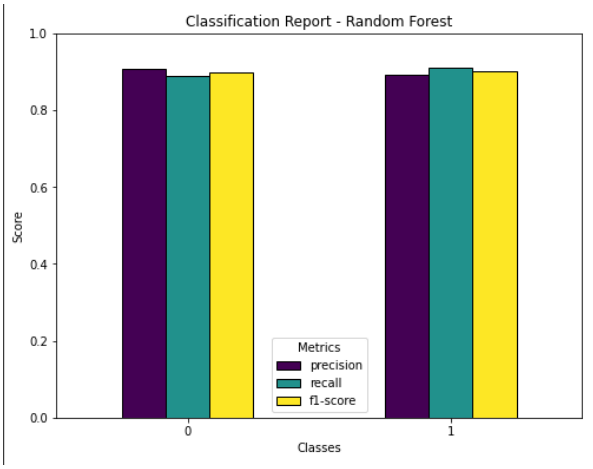

Fig. 9. Classification Report for Random Forest (RF).

C. Feature Importance

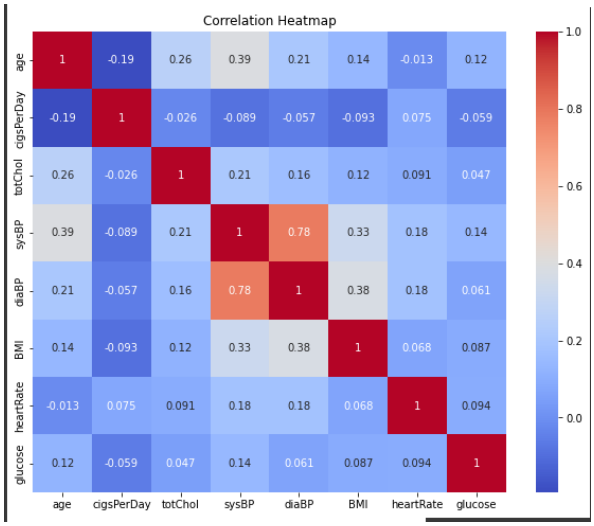

Fig. 10. Correlation Heatmap of Features

Examining the feature importance within the Random Forest model, as described by Chen et al. (2020) [6], reveals that certain features like systolic blood pressure, age, and total cholesterol levels are consistently the most significant predictors of CVD risk. This is in alignment with previous studies on the FHS that have highlighted the importance of these variables as per figure 10.

D. Impact of Data Preprocessing

As emphasized by Li et al. (2018) [5], data preprocessing, particularly handling missing values, is a critical step. The decision to fill missing values with mean and mode reflects a pragmatic approach to preserving the information in the dataset rather than discarding observations completely. While this approach may introduce some bias, it maintains the dataset's size, which is valuable for training machine learning models effectively.

V. Discussion

The results presented in figure 7 indicates that the Random Forest (RF) model exhibits the highest performance across all measured metrics, achieving cross-validation and testing accuracy of 0.87 and 0.89 respectively, with a well-balanced precision, recall, and F1-score for both classes as shown in the confusion matrices in Figure 5, 3, and 8. This demonstrates a robust ability to predict both high and low risk patients. The strong performance of the RF is attributed to its ability to handle the complex interactions of features and its inherent resistance to overfitting due to the ensemble learning methodology. Logistic Regression, although a simpler model, provided a good baseline performance, with an accuracy of around 0.68. It is a more interpretable model but was not able to capture the complex relationships in the data. The KNN model has shown comparable results to the logistic regression, achieving an accuracy score of 0.81. However, its performance is more sensitive to the choice of K parameter and the distance metric, and that is why its performance isn't as strong as RF.

The metrics of precision and recall were balanced and high in RF compared to other models, which suggests that the use of SMOTE prior to model training greatly improved the performance of the model, and the model didn't favor one class over another. The real world implications of our findings are that Machine learning models can be integrated into medical settings to help improve CVD prediction, that can aid in early detection and personalized treatment. The performance metrics showcase the superiority of Random Forest for this task, with its ability to capture the different complex correlations and its robustness.

A. Historical Context of CVD

As described by Mahmood et al. (2014) [4], the understanding of cardiovascular diseases has evolved significantly over the last century. In the early 20th century, infectious diseases were the leading causes of death, with CVDs playing a relatively minor role. However, by the mid-20th century, chronic non-communicable diseases, including CVDs, emerged as the dominant causes of mortality. This shift was accompanied by a gradual increase in public health concern regarding the prevention and treatment of CVDs, leading to the establishment of large-scale epidemiological studies like the FHS, which has played a pivotal role in identifying modifiable risk factors and understanding the natural history of CVD.

B. Focus on Specific CVD Subtypes

The references also emphasize the importance of distinguishing between different CVD subtypes, highlighting how each subtype presents with unique risk factors, underlying mechanisms, and clinical trajectories. While the broader category of "CVD" is useful in epidemiological studies, finer distinctions (e.g., coronary heart disease, stroke, heart failure, and atrial fibrillation) are important for targeted prevention and intervention strategies. For instance, research stemming from the FHS has shown the distinct role of systolic blood pressure as a leading risk factor for stroke [4], emphasizing the need for separate consideration of different measures of blood pressure control in CVD risk management.

VI. Conclusion

This study explored the application of machine learning to predict CVD using data from the Framingham Heart Study. We show that the Random Forest classifier can be a robust model for predicting CHD using the FHS dataset, when comparing the cross validation scores and the testing accuracy, along with the precision, recall and F1 scores. All the models were

evaluated with proper preprocessing steps including handling missing values and outliers. With its superior performance and great ability to handle complex data, the random forest model has a great potential for future use in CVD prediction.

VII. Future Work

Future work will focus on enhancing the model performance, firstly by exploring different feature selection techniques to identify the most relevant features. Also, we can use different approaches to handle class imbalance problems to boost the performance even more. Furthermore, we will investigate different types of neural network architectures (such as attention based mechanisms) to enhance model performance. It's also worth mentioning A significant bottleneck in the current methodology is the limited size of the FHS dataset, which restricts the model's ability to generalize effectively. To mitigate this, we plan to explore generative models, such as Generative Adversarial Networks (GANs) or transformer-based architectures, to synthesize realistic and representative data that mirrors the original dataset's distribution. This approach can help expand the dataset artificially, enabling better generalization and robustness in the model.

VIII. Acknowledgments

The authors acknowledge the use of data from the Framingham Heart Study and acknowledge that all resources were publicly available and was used in research purposes.

References

[1] W. B. Kannel, T. R. Dawber, A. Kagan, N. Revotskie, and J. Stokes, "Factors of risk in the development of coronary heart disease—six year follow-up experience," *Annals of internal medicine*, vol. 55, no. 1, pp. 33–50, 1961.

[2] T. R. Dawber, F. E. Moore, and G. V. Mann, "Coronary heart disease in the framingham study," *American journal of public health and the nations health*, vol. 47, no. 4-Pt-2, pp. 4–24, 1957.

[3] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[4] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective," *The Lancet*, vol. 383, no. 9921, pp. 999-1008, 2014.

[5] J. Li, J. Xie, C. Xu, and Q. Zhou, "Cardiovascular risk prediction based on framingham heart study data using logistic regression model with feature selection," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2156–2159.

[6] H. Chen, Y. Zeng, Y. Hu, G. Pan, G. Chen, W. Li, S. Wang, Y. Liu, H. Li, and J. Huang, "A comparative study of machine learning methods on predicting the risk of ten year coronary heart disease," *International journal of medical informatics*, vol. 137, p. 104079, 2020.

[7] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.

[10] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[11] H. Ishwaran and U. B. Kogalur, "Random Survival Forests for R," *R News*, vol. 7, no. 2, pp. 25–31, 2008.

[12] A. Y. Ng, "Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance," in *Proceedings of the 21st International Conference on Machine Learning*, 2004, pp. 78–85.

[13] D. G. Kleinbaum and M. Klein, *Logistic Regression: A Self-Learning Text*, 3rd ed. Springer, 2010.

[14] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Wiley, 2019.

[15] S. Van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. CRC Press, 2018.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[17] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[19] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[20] R. Caruana and A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 161–168.

[21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.