Coronary Heart Disease Prediction: A Comparative Study of Machine Learning Algorithms

Ahmad Hammoud *, Ayman Karaki *, Reza Tafreshi, Shameel Abdulla, and Md Wahid

Mechanical Engineering Department, Texas A&M University at Qatar, Doha, Qatar
Email: ahmad.hammoud@qatar.tamu.edu (A.H.); ayman.karaki@qatar.tamu.edu (A.K.);
reza.tafreshi@qatar.tamu.edu (R.Y.); Shameel.abdulla@qatar.tamu.edu (S.A.); md.wahid@qatar.tamu.edu (M.W.)
*Corresponding author

***Abstract*—Efforts to enhance the precision of heart disease detection methods are crucial in reducing the expensive healthcare expenses associated with the diagnostic processes. Extracting patterns from medical data can unlock associations to improve heart disease diagnosis techniques. This study aims to construct an efficient machine learning model to act as a reliable component of the medical decision support system. Seven different machine learning models were investigated including Logistic Regression, Support Vector Classifier, K-Nearest Neighbor (KNN), Random Forest, Decision Tree, Naïve Bayes, and Gradient Boosting Classifier, which are comprehensively explored for heart disease classification. Hyperparameter optimization for these algorithms involves three techniques: Grid Search, Random Search, and Bayes Search. The assessment of each model's performance incorporates measuring specificity, sensitivity, and F1-scores, leveraging the dataset with 12 attributes and 1189 observations from three medical clinics (Cleveland, Statlog, Hungary). Feature selection methods, including the wrapper method, embedded method Chi-Sqaured, and variance analysis, are deployed to identify highly correlated features, ultimately reducing the data's dimensionality to 7 features. The evaluation process employs 10-fold cross-validation, demonstrating that the Random Forest Model achieves the highest average accuracy at 92.85%, surpassing the previously reported 86.9%. Additionally, 10-fold cross-validation ensures the models' reliability and resilience to data imbalance. Ensemble-based methods reaffirm the Random Forest's superior performance in diagnosing heart diseases, boasting an accuracy of 94.96%. In sum, this developed model exhibits reliability in heart disease classification and presents a promising solution for medical applications, to effectively mitigate diagnostic costs and time constraints.**

***Keywords*—applied machine learning, coronary heart disease, random forest**

I. Introduction

Cardiovascular Diseases (CVD) are highlyprevalent in the world population, responsible for one-third of the total deaths per year, of which 7.5 million deaths are attributed to Coronary Heart Diseases (CHD). Approximately 1.8 million of these deaths aresudden and linked with Acute Coronary Syndrome (ACS) [1]. Clinically, at least 16.1% of heart failure patients are misdiagnosed [2]. The diagnosis of CHD is often done through a coronary calcium scan utilizing X-rays for the arteries [3]. For the detection of CHD, CT Coronary Angiography (CTCA) has a sensitivity and specificity of 89% and 96%, respectively [4]. Machine Learning (ML) is deeply integrated into the medical field, especially inhealthcare applications, where it aids in improving the diagnostic process. ML algorithms use data to learn complex and non-linear patterns relating to the features by minimizing the error between the predicted and actual outcomes.

In the areas of medical application, the integration of Machine Learning (ML) has significantly augmented diagnostic procedures. ML algorithms leverage data to discern intricate and nonlinear patterns in the features, minimizing the disparity between predicted and actual outcomes. However, existing methods suffer from limitations, in particular in areas of subtle cardiac irregularities, which necessitates a more reliable and refined approach to diagnosis. Thus, this study attempts at bridging this gap by proposing an advanced machine learning model for precise and timely CHD detection. By utilizing seven different ML algorithms and optimizing their performance, we aim to offer a superior solution that can effectively reduce the frequency of misdiagnosis and streamline the diagnostic process for the patients.

II. Literature Review

Various ML-algorithms were developed to diagnose heart diseases [5–11]. The difference between the papers was that researchers applied various data mining techniques such as association rules technique, Clustering, and classification algorithms to extract the most prominent parameters for predicting heart disease at good accuracy. In Ref. [5], five main machine learning techniques were implemented to determine which method produces the highest accuracy for this type of data. The implemented classifiers were Naïve Bayes, Decision Tree, Discriminant, Random Forest, and Support Vector Machine. The two datasets used were the Cleveland Clinic Foundation

Manuscript received June 12, 2023; revised August 3, 2023; accepted September 29, 2023; published January 3, 2024.

dataset and the Statlog dataset [12]. Both datasets were processed to produce 14 attributes to reduce the number of variables. The results showed that all classification algorithms were able to accurately classify CHD patients; notably, decision tree was able to outperform all other datasets with an accuracy of 98%, which was followed by Random Forest at 93%. No overfitting analysis was performed on the decision tree model which perturb the reliability of the results considering the number of observations (573).

In Ref. [6], the Cleveland dataset [13] and Statlog dataset were also combined and used as one dataset. The paper used seven different machine learning algorithms to compare the algorithms based on the accuracy of each. The algorithms were Logistic Regression, Support Vector Machine, Deep Neural Network, Decision Tree, Naïve Bayes, Random Forest, and K-Nearest Neighbor. The performance of each was evaluated using the same datasets, and the highest accuracy obtained was the Neural Network with 98.15%.

In this report, the data from Hungary, Cleveland clinic, and Statlog will be combined and used as one dataset. The objective of this project is to develop a ML-based decision support system for healthcare that can accurately identify coronary heart disease patients faster than CTCA. This report is structured as the following: Section I introduces the problem. Section II describes the dataset and methodology implemented in training ML algorithms. While Section III presents and discusses the results of the explored ML algorithms, Section IV concludes the contribution of this work, as well, possibilities for future work.

III. Materials and Methods

A. Data Exploration

The dataset includes 14 predicting attributes for 1190 patients. The attributes include age, sex, resting blood pressure (trestbps), degree of chest pain (cp), ranging from 1-low to 4-high, cholesterol level, maximum heart rate (thalach), exercise-induced angina (exang) along with the peak (oldpeak) and the slope of the peak's ST segment (slope), fasting blood sugar (fbs), and resting electrocardiogram (restecg). The data is labeled by a response variable which indicates whether a patient had a CHD or not. Out of the 1190 patients, 553 were diagnosed with CHD.

Initially, the data was explored by visualizing the features and searching for missing or non-logical data. Although there are no missing data, yet there are non-logical values in the cholesterol levels for 172 patients (14.5% of the patients). The cholesterol value for those patients is 0, which is not realistically possible. These values are more present for patients from the Hungary dataset. As well, chest pain type 4 is very common among heart disease patients, unlike the other three types of chest pain. The data is noticed to be biased towards men as around three-fourths of the population are men. Fig. 1 shows the block diagram that illustrates the procedure.

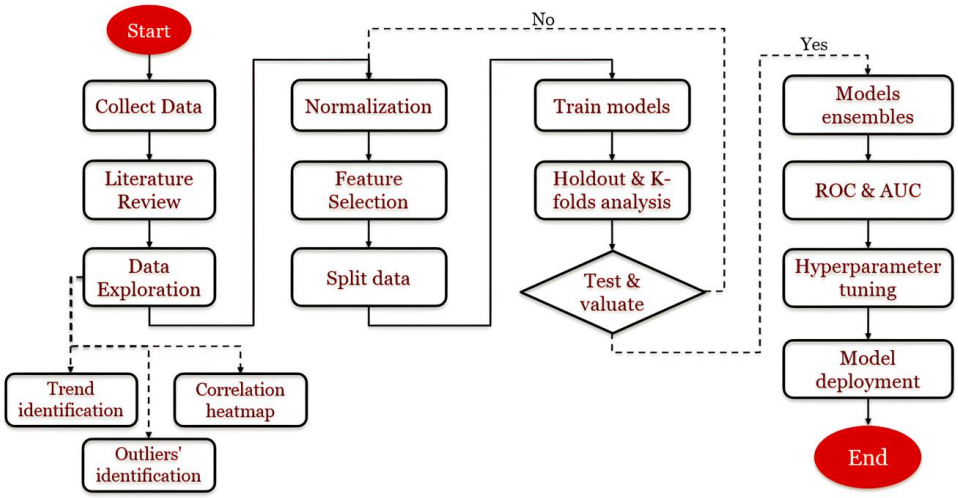

Fig. 1. A block diagram illustrating the procedure.

B. Data Imbalance

There are multiple ways to deal with data imbalance including under-sampling, over-sampling, Synthetic Minority Over-sampling TEchnique (SMOTE), and ensembles. Selecting the correct approach is critical in the field of health care to avoid inferring any misleading conclusions. While under-sampling imitates balance and expedites the training runtime, it results in a loss of classification performance due to the loss of the majority instances. Meanwhile, over-sampling increases the likelihood of overfitting and abrupt degradation in the predicted response as the model memorized patterns for the minority instances. Applying SMOTE and ensembles is better yet could be unnecessary if the imbalanced attributes were excluded during the feature selection process. Accordingly, the accuracy of each model will be assessed using the F1-scores to evaluate the need for balancing the data.

C. Feature Selection

Feature selection methods, either qualitative or quantitative, aim to filter the most predicative features.

One of the aspects involved in quantitative filtering methods is the correlation between each attribute and the response. Multiple techniques were explored to normalize the data and compare their corresponding variances and weights, such as chi-squared, MinMax Scaler, log-transformation, and *z*-scores. While the *z*-scores work best with normally distributed data (with little to no skewness), it must be supported with a continuity correction in the case of binary data. log-transformation offer a great solution for skewed distributions, yet do not outperform the chi-squared analysis in the case of categorical attributes. Since half of the attributes are categorical, the chi-squared distribution and MinMax Scaler are good techniques for normalizing the data. MinMax Scaler surpasses the chi-squared distribution as it better represents continuous distributions for non-categorical attributes given the absence of outliers.

After exploring the correlation of each attribute with the response variable, its *p*-value, variance, and weight in the principal components with explained variance larger than 5%, the features with the highest correlation and lowest *p*-value were selected. The interaction between the features was also considered using their correlation with one another. Wrapper and embedded methods, model performance-based feature selection methods, were also applied. Further pre-processing and data analysis was carried out to engineer the best features. It is important to note that all pre-processing was done on the training data and applied later to the test data.

D. Models Training and Evaluation

The data was split using an 80/20 holdout. While 80% of the data was used for training and validation purposes, 20% was used as a testing subset (new observations). A set of classification models were trained and tuned using a 10-folds cross validation. Cross-validation was implemented on the training data to ensure that all observations have the chance of appearing in both training and validation datasets, as well, to avoid overfitting. The investigated classification models are Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), and Gradient Boosting Classifier (GBC). The evaluation metrics for each model include the accuracy, specificity, sensitivity, and F1-scores. To assess the reliability of the selected features, each model was trained and tuned using the original dataset (excluding the cholesterol column) and the selected 7 features.

E. Hyperparameter Tuning

In order to improve the accuracy of the tested models, hyperparameter tuning was done for each model using three hyperparameter tuning methods. The methods used were GridSearch, RandomSearch, and BayesSearch. First, a set of parameter ranges was set before running any hyperparameter tuning methods. Then hyperparameter tuning models were run to check for the best parameters foreach model and apply using five folds cross validation on 80% of the data which was the training data. Next, the accuracy score was obtained by using the 20% data which was the testing data. Furthermore, a Stacked Generalization model was utilized. The Stacked Generalization model is a type of meta-learning algorithms to learn the best way of combining two or more machine learning algorithms. Since Stacked Generalization uses a base of different machine learning algorithms, the stacking ensembles are often heterogeneous. For this particular application the top three models based on accuracy were used for the base level and the meta-learner used was the fourth best accuracy. An accuracy score was obtained by using the testing data and a confusion matrix was plotted.

IV. Result and Discussion

Fig. 2 shows the histogram of patients with and without CHD as a function of blood cholesterol. It is observed that zero cholesterol values are imbalanced, as most of them are for patients with CHD. To deal with the outliers in the cholesterol values, extensive literature review and visuals were used. The collected cholesterol values are not well correlated with the response, CHD diagnosis, which implies one of the following: 1) the cholesterol value for some patients was collected without fasting (cholesterol level is highly dependent on the fat and carbohydrates level consumed in the last meal [14]), or 2) cholesterol is not directly related to the investigated heart disease, unlike stress levels (controversial studies showed no relation between cholesterol and CHD [15]). To investigate the reasons behind the missing cholesterol values, patients with zero cholesterol were compared against other patients for each attribute. the distribution for any attribute, e.g., age group or sex, had the same trends in both subsets. As the missing cholesterol values is not related to the values of any of the available attributes, the values are Missing Completely at Random (MCAR).

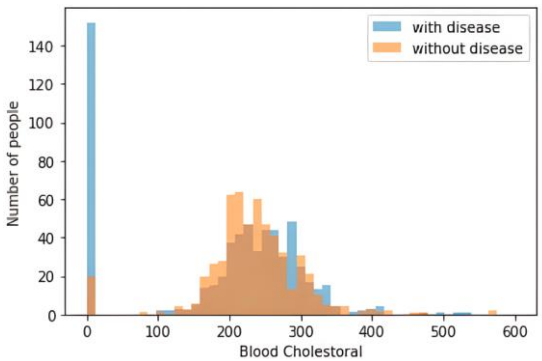

Fig. 2. Distribution of people with and without heart disease as a function of blood cholesterol.

Fig. 3 shows the correlation heatmap between all attributes from the original dataset. The two attributes with the highest correlation to the response variable are chest pain (cp) and exercise-induced angina (exang). It was noticed that filtering out patients with zero cholesterol values (around 14% of the patients) affects its correlation to the response variable (changes from −0.2 to +0.11). On the other hand, the correlation of other attributes with the response variable experienced slight to no effects. Then, feature engineering process was conducted according to

the correlation values, *p*-values, variances, principal components, wrapper, and embedded methods.

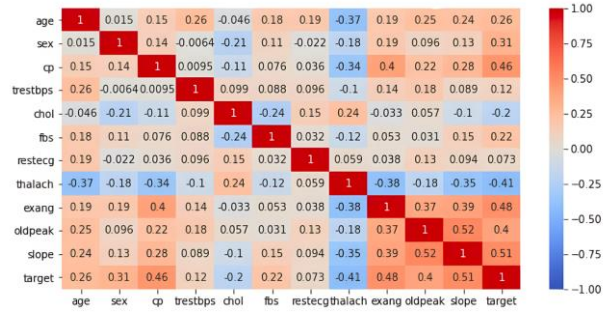

Fig. 3. Heatmap for correlations between attributes-original dataset.

Figs. 4 and 5 show the results obtained from the wrapper and embedded methods, respectively. The selected features were picked according to the ones proposed by the embedded method along with the correlation and variance statistics. The selected 7 features are (age, chest pain, fasting blood sugar, maximum heart rate, exercise induced angina, and its peak's ST segment slope) which will be used to predict the response variable (identify CHD patients). Considering that cholesterol had small variances, the cholesterol attribute was dropped without the need to filter or replace the zero-valued cells.

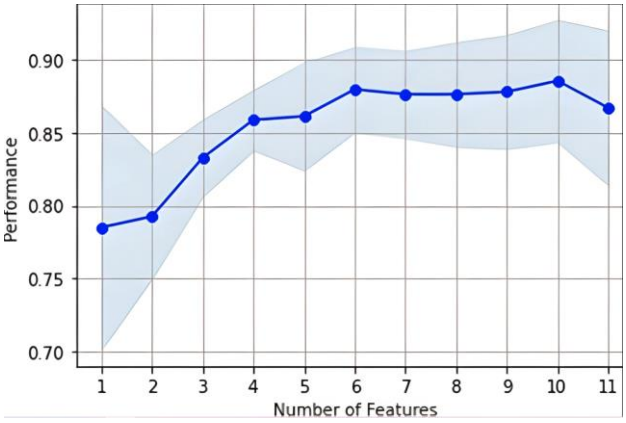

Fig. 4. Model performance as a function of the number of selected features—Wrapper method (Sequential Forward Selection).

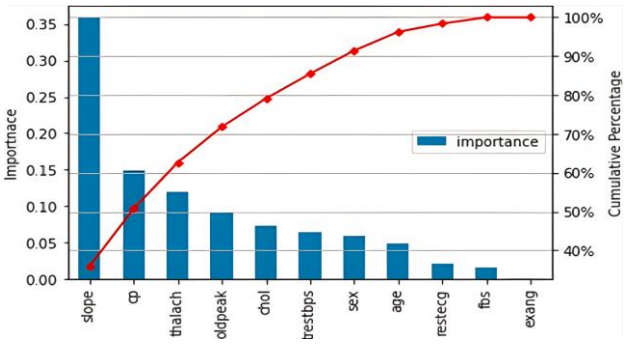

Fig. 5. Feature importance and cumulative percentage—Embedded method.

Figs. 6 and 7 show the results for the trained models using all attributes, except cholesterol, and the engineered 7 attributes. The best performing model was RF, with an accuracy and sensitivity of 89.5% and 91.7%, respectively, when trained using the selected 7 features.

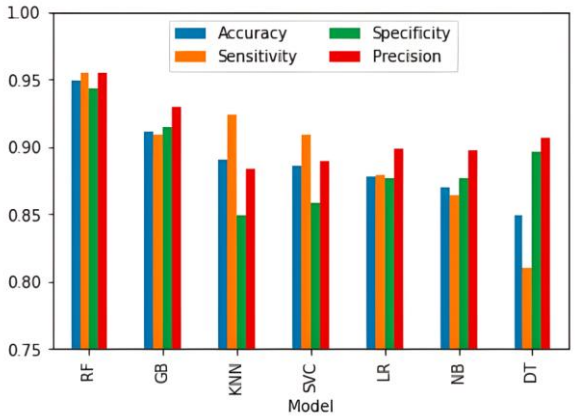

Fig. 6. Models' performance metrics using all attributes, except cholesterol.

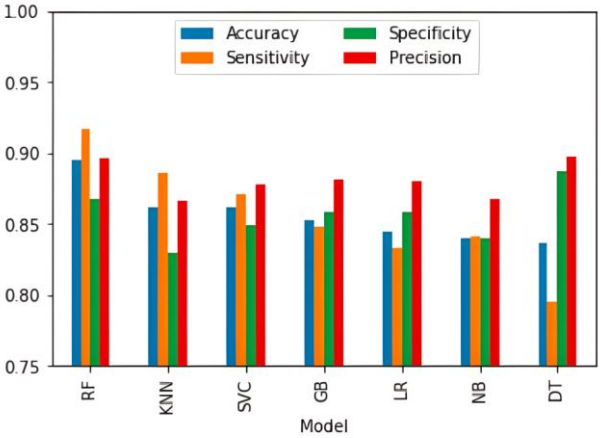

Fig. 7. Models' performance metrics using the engineered 7 features.

Table I lists the average test accuracy for each of the models using all attributes and only the selected features. It is noticed that the maximum difference in the scores is around 6% which reflects well-engineered features. Using the selected features is more computationally efficient (average training period reduced by around 35%) without sacrificing the models' performance. After calculating the F1-scores for each model, it was found that the maximum difference between any of the F1-score and the model accuracy is 4%.

TABLE I. ACCURACY OF EACH MODEL USING ALL ATTRIBUTES, EXCEPT CHOLESTEROL, AND THE SELECTED 7 ATTRIBUTES

Model	Accuracy (11 Features)	Accuracy (7 Features)
RF	95.80%	89.50%
KNN	88.23%	86.13%
SVC	89.50%	86.13%
GB	92.44%	85.29%
LR	86.97%	84.45%
NB	86.97%	84.03%
DT	86.55%	83.61%

Figs. 8 and 9 show the confusion matrices for the best performing model, RF, using all attributes and the selected 7 features, respectively. In the case of identifying heart

disease, the most critical value is the number of False Negatives (FN). Thus, tuning the models aims to minimize the number of FN instances without increasing the number of False Positive (FP) incidents. In other words, to improve the F1-score for the 'diseased' patients without forfeiting that for 'healthy' patients.

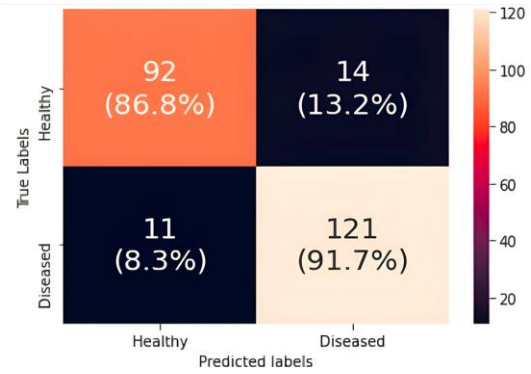

Fig. 8. Confusion matrix for RF model using the engineered 7 features—pre-tuning.

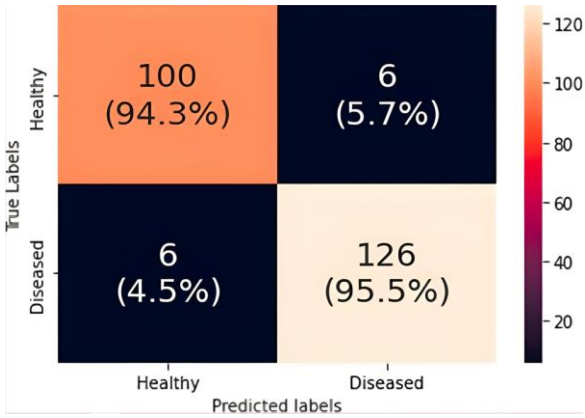

Fig. 9. Confusion matrix for RF model using the engineered 7 features after Grid Search tuning.

The tuning parameters used varied between different models, however, the main parameters investigated dealt with the number of iterations, max depth, kernels, and 'C' values. However, one common parameter that was fixed for all models was the number of folds used to run the hyperparameter tuning models which was five folds. Table II shows the accuracy of each model before and after tuning using the selected 7 attributes.

TABLE II. ACCURACY OF EACH MODEL BEFORE AND AFTER TUNING USING THE SELECTED 7 ATTRIBUTES

Model	Test Accuracy [pre-tuning]	Test Accuracy [post-tuning]		
		Grid Search	Random Search	Bayes Optimization
RF	89.50%	94.96%	90.34%	90.76%
KNN	86.13%	87.82%	86.55%	86.55%
SVC	86.13%	89.08%	84.87%	89.08%
GB	85.29%	84.03%	84.87%	85.29%
LR	84.45%	87.39%	79.83%	79.83%
NB	84.03%	86.97%	86.97%	86.97%
DT	83.61%	80.25%	81.51%	80.25%

Previous work in Ref. [16] was implemented on the same dataset and achieved a classification accuracy of 86.9% with diagnosis rate of 93.3% using RF model. In this study, the tuned RF model achieved an accuracy of 94.96% with a diagnosis rate of 94.3%. As both models were based on a 10-fold cross-validation and 80/20 split, the accuracies are comparable. The accuracy improvement is mainly related to the difference in handling the cholesterol attribute. While the cholesterol attribute was completely dropped in this study, patients with zero cholesterol were filtered in [16]. Another aspect that may drop the accuracy is data imbalance. Balancing the patients in terms of sex attribute was not conducted in [16], however, it is not needed in this study as sex attribute was dropped during the feature selection process.

The results could be compared with previous work conducted on smaller datasets including [5, 6]. While the size of the dataset comprises of 1190 patients in this study, only 573 patients were used in [5, 6]. In Ref. [5], the most accurate model was reported as the decision tree with 99% accuracy and no cross- validation. The second most accurate was the random forest with 93% accuracy. In Ref. [6], the highest accuracy was around 98%. As both studies lack the cross-validation step, the results could be comparable to the highest accuracies among the 10-folds in this study. Among the trained 10-folds, RF had the highest accuracy of 98.9% after tuning which outperforms the performance reported in [5]. Given the dataset size of 573 patients, the unconsidered imbalance, the absence of cross-validation, and the unreported number of branches for the decision tree model with 99%, the model results are highly questionable.

V. CONCLUSION

In this paper, multiple machine learning algorithms were implemented to classify coronary heart disease patients using the data from 3 clinics (Hungary, Cleveland, and Statlog). The trained models included Logistic Regression, Support Vector Classifier, K-Nearest Neighbor, Random Forest, Decision Tree, Naïve Bayes, and Gradient Boosting Classifier. In addition to a 80/20 data split, a 10-fold validation was conducted for all models. Using the selected features, the Random Forest algorithm outperformed all the models with an accuracy of 89.50% (pre-tuning) and 94.96% (post-tuning). The study anticipated weak links between cholesterol levels and CHD. The Random Forest model is to provide a decision support system for healthcare sector to identify coronary heart disease patients along with CT coronary angiography. The Algorithm could improve the accuracy of detecting heart disease and reduce the need for immense health care expenses and a long diagnosis process. Future plans for using ML for the diagnosis of coronary heart disease may include the development of more sophisticated algorithms that can handle the complexity of the disease and incorporate a greater number of risk factors. There may also be efforts to improve the interpretability of ML models, to allow for a better understanding of the reasoning behind a diagnosis.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Ahmad Hammoud and Ayman Karaki conducted the research and analyzed. Ahmad Hammoud wrote part of the introduction, methodology, results, and conclusion. Ayman Karaki wrote the other part of the introduction, literature review, results, and conclusion. Reza Tafreshi, Md. Wahid, and Shameel Abdulla reviewed and all authors had approved the final version.

References

[1] Cardiovascular diseases (CVDs). (2022). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] C. W. Wong, J. Tafuro, Z. Azam, D. Satchithananda, S. Duckett, D. Barker, A. Patwala, F. Z. Ahmed, C. Mallen, and C. S. Kwok, "Misdiagnosis of heart failure: A systematic review of the literature," *Journal of Cardiac Failure*, vol. 27, no. 9, pp. 925–933, 2021.

[3] Diagnosing Coronary Artery Disease. (2022). [Online]. Available: https://nyulangone.org/conditions/coronary-artery-disease/diagnosis

[4] The SCOT-HEART investigators, "CT coronary angiography in patients with suspected angina due to coronary heart disease (SCOT-HEART): An open-label, parallel-group, multicentre trial," *Lancet*, vol. 385, no. 9985, pp. 2383–2391, 2015. https://doi.org/10.1016/S0140-6736(15)60291-4

[5] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 14, no. 12, December 2016.

[6] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," *IETE Journal of Research*, vol. 68, no. 4, pp. 2488–2507, 2020.

[7] J. Sony, U. Ansari, D. Sharma, and S. Sony, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Science and Engineering*, vol. 3, 2011.

[8] A. S. Sadh and N. Shukla, "Association rules optimization: A survey," *International Journal of Advanced Computer Research (IJACR)*, vol. 3, no. 9, pp. 111–115, 2013.

[9] A. Dubey, R. Patel, and K. Choure, "An efficient Data Mining and ant Colony Optimization technique (DMACO) for heart disease prediction," *International Journal of Advanced Technology and Engineering Exploration*, vol. 1, no. 1, pp. 1–6, 2014.

[10] 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, Apr. 2006.

[11] P. Chandra, M. Jabbar, and B. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," in *Proc. 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2012, pp. 628–634.

[12] Statlog. Heart Disease Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

[13] Cleveland Clinic Foundation. Heart Disease Data Set. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[14] Cleveland Clinic. (2022). Why you should no longer worry about cholesterol in food. [Online]. Available: https://health.clevelandclinic.org/why-you-should-no-longer-worry-about-cholesterol-in-food/#:~:text=%E2%80%9CThe%20body%20creates%20cholesterol%20in,the%20body%20in%20the%20liver

[15] U. Ravnskov, D. M. Diamond, R. Hama *et al*., "Lack of an association or an inverse association between low-density-lipoprotein cholesterol and mortality in the elderly: A systematic review," *BMJ Open*, vol. 6, no. 6, e010401, 2016. doi: 10.1136/bmjopen-2015-010401

[16] M. Pal and S. Parija, "Prediction of heart diseases using random forest," *J. Phys.: Conf. Ser.*, vol. 1817, 012009, 2021. doi: 10.1088/1742-6596/1817/1/012009