RESEARCH ARTICLE

Heart Alert: A heart disease prediction system using machine learning approach and optimization techniques

Justin Allen P. Denopol, Ma. Sheila A. Magboo, Vincent Peter C. Magboo

*Corresponding author's email address: vcmagboo@up.edu.ph

Department of Physical Sciences and Mathematics, College of Arts and Sciences, University of the Philippines Manila, Manila, Philippines

ABSTRACT

Background: Cardiovascular diseases belong to the top three leading causes of mortality in the Philippines with 17.8 % of the total deaths. Lifestyle-related habits such as alcohol consumption, smoking, poor diet and nutrition, high sedentary behavior, overweight, and obesity have been increasingly implicated in the high rates of heart disease among Filipinos leading to a significant burden to the country's healthcare system. The objective of this study was to predict the presence of heart disease using various machine learning algorithms (support vector machine, naïve Bayes, random forest, logistic regression, decision tree, and adaptive boosting) evaluated on an anonymized publicly available cardiovascular disease dataset.

Methodology: Various machine learning algorithms were applied on an anonymized publicly available cardiovascular dataset from a machine learning data repository (IEEE Dataport). A web-based application system named Heart Alert was developed based on the best machine learning model that would predict the risk of developing heart disease. An assessment of the effects of different optimization techniques as to the imputation methods (mean, median, mode, and multiple imputation by chained equations) and as to the feature selection method (recursive feature elimination) on the classification performance of the machine learning algorithms was made. All simulation experiments were implemented via Python 3.8 and its machine learning libraries (Scikit-learn, Keras, Tensorflow, Pandas, Matplotlib, Seaborn, NumPy).

Results: The support vector machine without imputation and feature selection obtained the highest performance metrics (90.2% accuracy, 87.7% sensitivity, 93.6% specificity, 94.9% precision, 91.2% F1-score and an area under the receiver operating characteristic curve of 0.902) and was used to implement the heart disease prediction system (Heart Alert). Following very closely were random forest with mean or median imputation and logistic regression with mode imputation, all having no feature selection which also performed well.

Conclusion: The performance of the best four machine learning models suggests that for this dataset, imputation technique for missing values may or may not be done. Likewise, recursive feature elimination for feature selection may not apply as all variables seem to be important in heart disease prediction. An early accurate diagnosis leading to prompt intervention efforts is very crucial as it improves the patient's quality of life and diminishes the risk of developing cardiac events.

Keywords: *heart disease prediction, machine learning, imputation techniques, feature selection, support vector machine.*

Introduction

Cardiovascular diseases, cerebrovascular diseases, and neoplasms are the top leading causes of mortality in the Philippines in 2021. Around 17.8 % of the total deaths in 2020 have been attributed to heart diseases [1]. The increasing morbidity and mortality posed serious significant burden to the country's healthcare system. Lifestyle-related habits such as alcohol consumption, smoking, poor diet and nutrition, high sedentary behaviour, overweight, and obesity have been increasingly implicated in the high rates of heart disease among Filipinos [2]. To reduce the increasing mortality, it is

necessary to identify people at risk for heart disease so that early and prompt interventions on treatment and counseling can be initiated [3]. Although there has been a significant improvement in diagnostic procedures in the recent years, cardiologists and primary care physicians often encounter difficulties in the early detection and diagnosis of heart disease [4]. In most cases, patients are usually asymptomatic in the early course of the disease. It is only when there is an abnormal laboratory test results during routine check-ups such as annual physical examination of employees that clinical suspicion of heart disease is entertained. By the time a patient is overtly symptomatic like having anginal chest pain, shortness of breath, and body weakness, heart disease may already be in its advanced course. It is in this area where machine learning would be of paramount importance in developing an efficient and accurate prediction model, particularly in the early stages of the disease. This machine learning model can be implemented via a web application that could be incorporated in clinical practice by physicians. Machine learning is being rapidly developed in cardiovascular research to diagnose diseases and predict risks and outcomes. Many health researchers have applied machine learning models because of the excellent performance in pattern recognition and disease classification. However, most of the anonymized publicly available clinical datasets are noisy, inconsistent, and redundant which require proper pre-processing steps before applying any machine learning model [5]. Accurate heart disease prediction using machine learning technique is seen as an important potential support tool for health professionals in their decision-making process, enabling a more personalized treatment for patients leading to increased chances of controlling modifiable risk factors for cardiovascular disease [6-8].

In the study by Patel *et al.* using the Framingham dataset to predict the likelihood of Chronic Heart Disease, the authors applied several machine learning models (support vector machine, decision tree, and naïve Bayes) [9]. They also reported that their models can be used by healthcare institutions to predict chronic heart disease beforehand and hence, the necessary preventive precautions can be instituted. Alsafi and Ocan applied an optimization machine learning technique for coronary heart disease diagnosis involving the Framingham heart disease dataset. They also used a feature-selector optimization model to select the best subset of chronic heart disease features and applied SMOTE (synthetic minority oversampling technique) to solve the class imbalance problem. Their results showed their proposed random forest optimization had the best performance accuracy at 90% [10]. In the study by Yazdani *et al.* [11], the authors used Weighted Associative Rule Mining (WARM) to predict heart disease applied to the Cleveland Heart dataset. Their results showed that the assignment of appropriate weight scores has improved the confidence level of prediction of heart disease. Bharti *et al.* [12], studied various machine learning algorithms (logistic regression, k-Nearest Neighbors, decision tree, support vector machine, random forest, XGBoost) and deep learning methods applied to the University of California Irvine Machine Learning Heart Disease dataset. Three approaches were made: (1) without Feature selection and with no outlier detection, (2) with feature selection but without outlier detection, and (3) with feature selection and outlier detection. The best models were obtained using the 3rd approach (with feature selection and outlier detection) with an accuracy rate of 94.2%. Lin applied six machine learning algorithms (logistic regression, k-Nearest Neighbors, adaptive boosting, random forest, and extreme gradient boosting) on Cleveland Heart Disease [13]. Results showed random forest with the best classification performance metric accuracy (84.8%), F1 score (82.9%), precision-recall area under the curve (PRC-AUC of 0.909), and receiver operating characteristic curve (ROC-AUC 0.917). Khdair made a comparative analysis of different machine learning techniques (logistic regression, support vector machine, k-Nearest Neighbors, and multilayer perceptron neural networks) to predict chronic heart disease using South African Heart Disease from KEEL machine learning repository [14]. The results showed support vector machine with the highest overall classification performance. Patro *et al.* applied k-Nearest Neighbors, naïve Bayes, support vector machine, lasso and ridge regression algorithms in predicting heart disease [15]. Support vector machine obtained the best classification performance at 92% accuracy and F1-score of 85%. Akella and Akella [3] applied different machine learning algorithms (logistic regression, random forest, support vector machine, k-Nearest Neighbors and neural networks) to predict the presence of coronary artery disease using the Cleveland Heart Dataset. All machine learning algorithms obtained accuracies greater than 80%, with the neural networks generating the best classification performance with more than 93% accuracy rate and 93% recall.

The objective of the study is to predict the presence of a heart disease using a variety of machine learning (ML) algorithms namely: Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Adaptive Boosting (AB) evaluated on an anonymized publicly available cardiovascular disease dataset. It is also aimed to create an application system based on the top-performing machine learning model that would

predict the risk of developing heart disease. Furthermore, the study also aimed to assess the effect of different imputation methods: Mean, Median, Mode, and Multiple Imputation by Chained Equations (MICE) and Recursive Feature Elimination (RFE) as a feature selection technique on the classification performance of different machine learning algorithms. Classification metrics computed were the F1-score, accuracy, and area under the receiver operating characteristic curve (AUC).

Methodology

The research framework for heart disease classification was performed in several stages. After loading the dataset, pre-processing methods were applied to the dataset. These methods were as follows: (1) data cleaning, (2) imputation techniques for missing data, (3) normalization of dataset, and (4) feature selection to remove redundant features. Data cleaning included measures to assess the presence of missing or null entries and duplicate records. The duplicate records were then dropped from the dataset. The column titles were also renamed and the spaces (" ") were removed and replaced with underscores ("_") to easily call and use the data. Data transformation was also performed, with the int values converted to float. A variety of machine learning algorithms were then tested on the dataset followed by an assessment of its performance. The model with the best performance was used to create the Heart Alert web application. Fig. 1 shows the development framework for Heart Alert.

Dataset Description

An anonymized publicly available dataset, Heart Disease Dataset (Comprehensive), taken from a machine learning data repository (IEEE Dataport) was used in this research study [16]. It is a curated dataset which combined five (5) independently available heart disease datasets, namely, Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog (Heart) with 12 common features including the target variable (diagnosis of heart disease). The seven categorical attributes are sex, chest pain type, fasting blood sugar, resting ECG, exercise angina, ST slope, and the target variable. On the other hand, the five numeric attributes include age, resting blood pressure, cholesterol level, max heart rate, and old peak. (See appendix for the description of these variables.)

Optimization Processes

The optimization processes used were imputation techniques and feature selection method. Four imputation techniques were tested on the dataset: Mean, Median, Mode, and Multivariate Imputation by Chained Equations. For the feature selection method, recursive feature elimination was used in the dataset to select the features that will have the best predictive power on heart disease classification.

Machine Learning Models

The data was divided into 80% training and 20% testing with a 10-fold cross-validation technique. Six ML algorithms were tested on the dataset: SVM, NB, RF, LR, DT, and AB. Python 3.8 and its various machine learning libraries (Scikit-learn, Keras, TensorFlow, Pandas, Matplotlib, Seaborn, and NumPy) were utilized in our simulation experiments. To assess prediction performance, the metrics computed were F1-score, accuracy, and AUC. For the top-performing models, recall or sensitivity and specificity were also assessed.

Web Application

The web application named "Heart Alert" was developed using the best-performing model. The web application used an open-source app framework in python called Streamlit to build the front end of the application. The web application was hosted on the local server.

Results and Discussion

The performance metrics of the various models highlighting the effects of the imputation techniques are seen in Table 1. When there is no imputation technique applied to the dataset, the best-performing model was obtained by SVM with 90.2% accuracy, 91.2% F1-score, and 0.902 AUC. On closer inspection, there was a prominent increase in the accuracy rates by 4.3 percentage points for RF with mean and median imputation and for LR with mode imputation. On the other hand, a prominent decrease in the accuracy rates was observed in SVM with mode imputation and MICE imputation (6.5 - 8.7%), NB with median imputation (4.3%), DT for all types of imputation (4.4% - 5.5%). The rest of the models did not exhibit any prominent changes in the accuracy rates. Likewise, almost the same pattern was observed for the changes in the AUC values. A prominent decrease in the AUC was observed in SVM with mode and MICE imputation, DT for all types of imputation. A prominent increase in the AUC was noted for LR with mode imputation. For F1-scores, a prominent decrease was evident in the SVM with mode and MICE imputation and DT with MICE imputation. On the other hand, a prominent increase in F1-score was seen in RF with mean and median imputations. Nevertheless, SVM without imputation

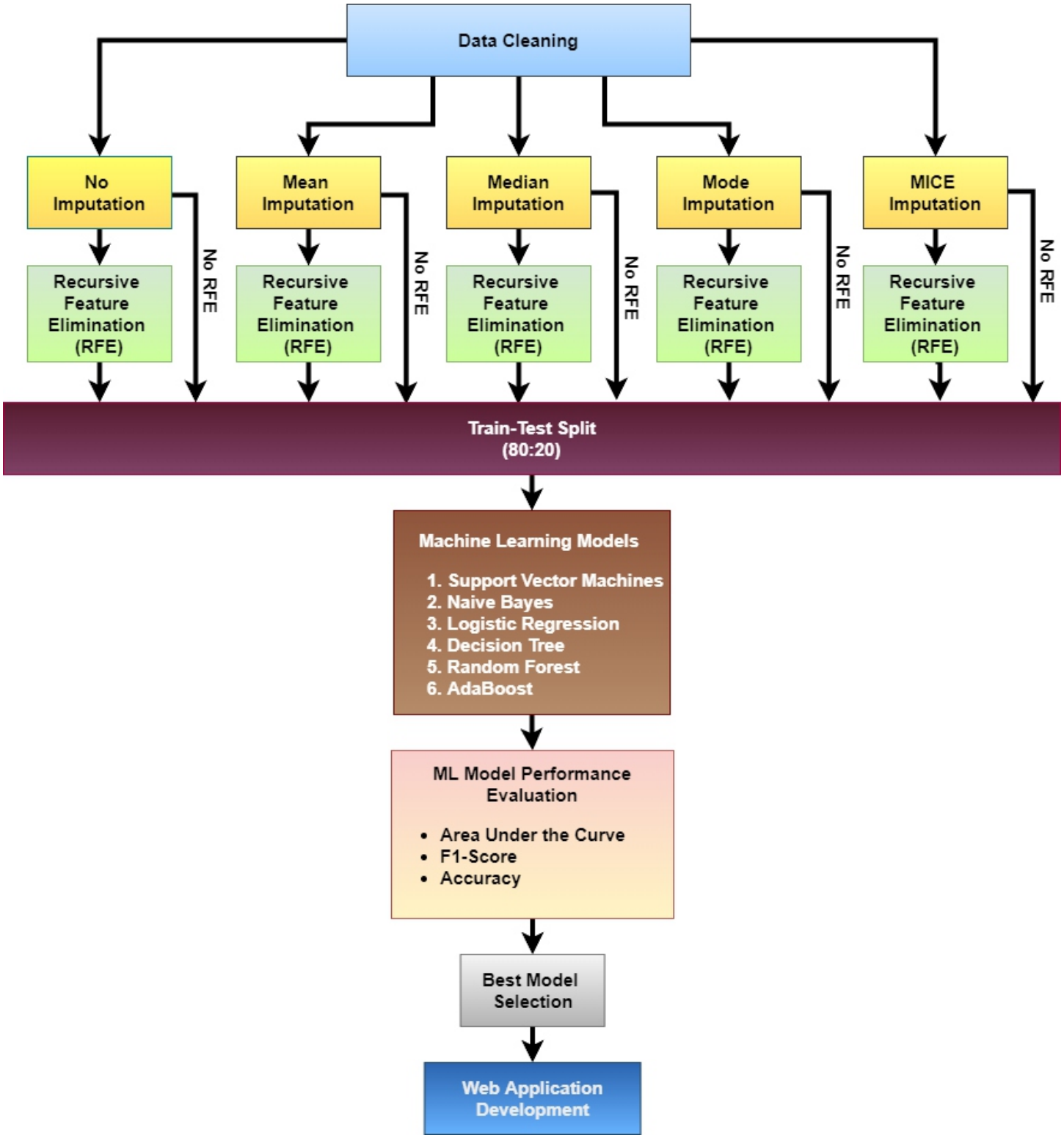


Figure 1. Heart Alert Development Framework

remains to be the top-performing model at 90.2% accuracy followed by RF with mean and median imputation and LR with mode imputation, all with 89.1% accuracy rate. Despite the seeming improvement in the performance metrics for some ML models (RF with mean and median imputation and LR with mode imputation), this is still not sufficient to surpass the performance of SVM without imputation. This pattern suggests that the effects of imputation techniques

Table 1. *Performance Metrics for Heart Disease Prediction - Assessment of Imputation Techniques*

Imputation Used	ML Model	Accuracy	F1-score	AUC
No Imputation	**SVM**	**90.2**	**91.2**	**0.902**
	NB	85.3	85.7	0.853
	RF	84.8	86.0	0.846
	LR	84.8	87.4	0.836
	DT	81.0	81.1	0.810
	AB	85.9	88.1	0.857
Mean Imputation	SVM	86.4	88.2	0.858
	NB	88.6	88.9	0.886
	RF	**89.1**	**90.9**	**0.887**
	LR	82.1	84.2	0.815
	DT	75.5	78.3	0.750
	AB	84.8	87.7	0.854
Median Imputation	SVM	86.4	87.4	0.863
	NB	81.0	82.2	0.811
	RF	**89.1**	**91.1**	**0.883**
	LR	84.2	85.6	0.846
	DT	76.6	78.3	0.772
	AB	87.5	87.2	0.875
Mode Imputation	SVM	81.5	84.1	0.816
	NB	83.7	84.7	0.837
	RF	84.2	85.7	0.839
	LR	**89.1**	**89.4**	**0.891**
	DT	75.5	78.3	0.764
	AB	86.4	86.9	0.866
MICE Imputation	SVM	83.7	86.0	0.830
	NB	84.8	86.0	0.848
	RF	85.3	87.7	0.850
	LR	82.1	83.4	0.820
	DT	76.1	76.8	0.761
	AB	88.0	89.5	0.881

on the performance are largely dependent on the ML model and possibly the nature of the dataset.

The comparative performance metrics to assess the effects of RFE are shown in Table 2. Generally, there were no prominent changes in the accuracy, AUC, and F1-scores across all ML models when RFE was applied to the dataset. This suggests that for this specific dataset, feature selection may not be useful and perhaps all features are important for predicting the presence of heart disease. It should be noted that the dataset for this study was an aggregation of five cardiovascular datasets of which only the common attributes or predictors were extracted. These common attributes may already represent the important and non-

Table 2. *Comparative Performance Metrics of ML Models with Recursive Feature Elimination*

Feature Selection	ML Model	Accuracy	F1-score	AUC
Without Imputation, Without RFE	SVM	90.2	91.2	0.902
	NB	85.3	85.7	0.853
	RF	84.8	86.0	0.846
	LR	84.8	87.4	0.836
	DT	81.0	81.1	0.810
	AB	85.9	88.1	0.857
Without Imputation, With RFE	SVM	89.1	89.8	0.890
	NB	85.3	85.1	0.853
	RF	85.3	86.6	0.852
	LR	84.2	84.7	0.842
	DT	79.3	80.0	0.794
	AB	85.3	87.6	0.847

Table 3. *Confusion Matrix and Performance Metrics of the Best-performing ML Models*

	SVM (No Imputation No RFE)	RF (Mean Imputation, No RFE)	RF (Median Imputation, No RFE)	LR (Mode Imputation, No RFE)
Actual	**Predicted** + -	**Predicted** + -	**Predicted** + -	**Predicted** + -
With Heart Disease Without Heart Disease	[[93 13] [5 73]]	[[100 14] [6 64]]	[[102 13] [7 62]]	[[84 13] [7 80]]
Accuracy	90.2	89.1	89.1	89.1
Recall/Sensitivity	87.7	87.7	88.7	86.6
Specificity	93.6	91.4	89.9	91.9
Precision	94.9	94.3	93.6	92.3
F1-score	91.2	90.9	91.1	89.4

redundant predictors for heart disease. As such, feature selection may no longer be suitable. Nonetheless, the best model for both scenarios with or without RFE, appears to be SVM with accuracy rates of 90.2% and 89.1%, respectively.

The confusion matrix of the best performing ML Models in heart disease prediction based on simulation experiments (with and without imputation technique, with and without feature selection or RFE) is seen in Table 3. The performance of the top four models seems to be similar or comparable to each other across all metrics suggesting that for this dataset, imputation technique may or may not be applied for missing values. Likewise, RFE for feature selection was not applied as all variables seem to be important in heart disease prediction. Nonetheless, the best model is SVM without any imputation technique and no feature selection as it obtained the highest metric values.

In this study, a web application named "Heart Alert" was created based on the best model for predicting heart disease. The Heart Alert System as seen in Figure 2 is a form containing the following fields: age, sex, chest pain type, resting blood pressure, cholesterol level in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak, and heart rate slope. Sex has two radio buttons, both for male and female options. Likewise, fasting blood sugar also has two radio buttons, both for lower than 120 mg/dl and higher than 120 mg/dl options. The following accepts numerical values: age ranging from 1 to 121, resting blood

sugar ranging from 1 to 500, cholesterol level ranging from 1 to 1000, max heart rate achieved ranging from 1 to 300, while old peak accepts a numerical input up to two decimal places. A dropdown box was used for chest pain type with four options (typical angina, atypical angina, non-anginal pain, and asymptomatic); resting rlectrocardiographic result with three options (normal, ST-T wave abnormality, and possible or definite left ventricular hypertrophy); induced angina with two options (yes and no); heart rate slope with three options: (upsloping or better heart rate with exercise (uncommon), flatsloping, or minimal change (typical healthy heart), and downsloping, or signs of unhealthy heart). Figure 2A shows a low risk for heart disease prediction output shown at the bottom of the prediction result page while the high risk for heart disease prediction output is shown in Figure 2B. Most of the studies in the literature pertain to the development of machine learning models with acceptably good performance. However, majority of these works were not deployed in clinical work due to the absence of a functioning application that a physician can utilize. With this application, Filipino doctors can avail of its use as a decision support tool particularly in the early stages of the disease, suggesting its potential utility in clinical practice.

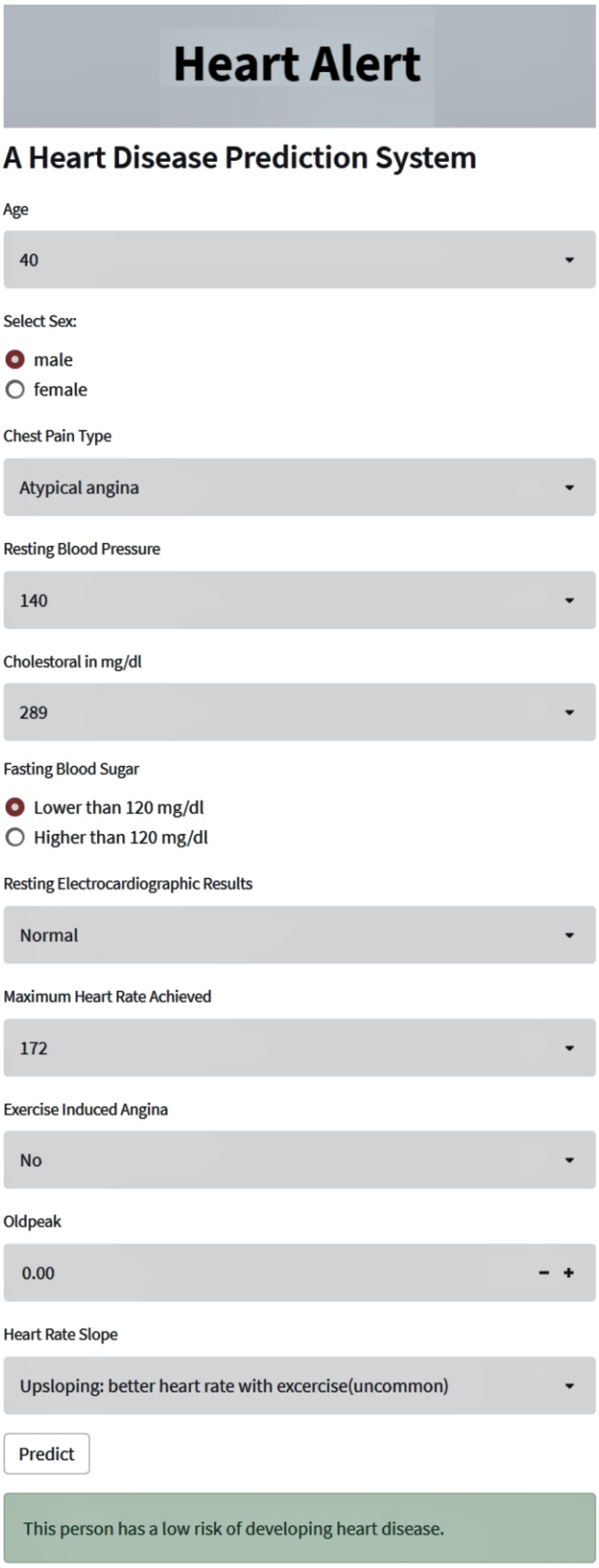
Four different types of imputation methods were analyzed in this research. Imputation is a method for replacing missing data with some substitute value as opposed to simply removing the data from the dataset. Usually in medical datasets, any missing values directly affect the accuracy of clinical decision-making [17]. A study indicated that the seriousness of missing values depends on how much data is missing, the pattern of missing data, and the mechanism underlying the missingness of the data [18]. They also reported deletion from the dataset to be the simplest method to use, however, it also introduced bias in analysis particularly when the missing data is not randomly distributed. The study by Khan and Hoque [19]reported that a simple solution of ignoring observations with missing values is not a problem when there are very few observations. If there is a large number of observations with missing values, this causes a significant loss of information and a decrease in the statistical power of the study. A simple way to do the imputation is to substitute the mean, median, or mode for the missing data. This is an easy step and very suitable for small numerical datasets. However, this simple imputation method may also produce bias or unrealistic results on high-dimensional data sets and seems to be performing poorly on big data sets [18]. But it should be noted that this method does not consider correlations between predictors as well as account for uncertainty in the imputations [19]. On the other hand, multiple imputation methods produce many values for replacing a missing value using different simulation modes. They are complex in nature but they do not suffer from biased values like single imputation. MICE is one the most common and flexible method of multiple imputation [20]. It iteratively fits a predictive model for each variable with missing values, conditional on other variables in the data [21]. Javadi *et al.*, indicated the use of MICE imputation using a mix of parametric and nonparametric methods. Though the default setting in MICE implementation is for imputation models to include variables as linear terms only with no interactions, they have indicated that the omission of important nonlinear terms may lead to biased results [22].

A wrapper method using RFE as a feature selection technique was illustrated in this research work. Feature selection removes redundant features which can be expressed by other attributes as well as irrelevant features which do not contribute to the performance of the model [23,24]. The reduction in the number of attributes leads to a reduction of the computational complexity of prediction algorithms which in turn increases the accuracy rate of the models as well as avoids potential overfitting [25]. RFE reduces model complexity by removing attributes one at a time until it automatically finds an optimal number of features based on the cross-validation score of the model [26,27].

The results of this study are comparable to the finding of other studies that uutilized machine learning in heart disease prediction [3-5,7,9-15]. Our findings suggest the utility of machine learning methods to make a reliable and accurate prediction of heart disease. An early accurate diagnosis leading to prompt intervention efforts is highly important as it warrants prompt intervention methods to improve the patient's quality of life, diminish the risk for developing heart failure and other cardiac events [3,6-9,28]. This research, thus, provide useful insights in the development of automated models that can assist healthcare professionals in the assessment of heart diseases. Using data analytics and machine learning, physicians can have a better understanding of cardiovascular diseases contributing to prompt and improved diagnosis thereby leading to early institution of treatments for patients.

Conclusion

Cardiovascular diseases belong to the top three leading causes of mortality in the Philippines with 17.8% of the total



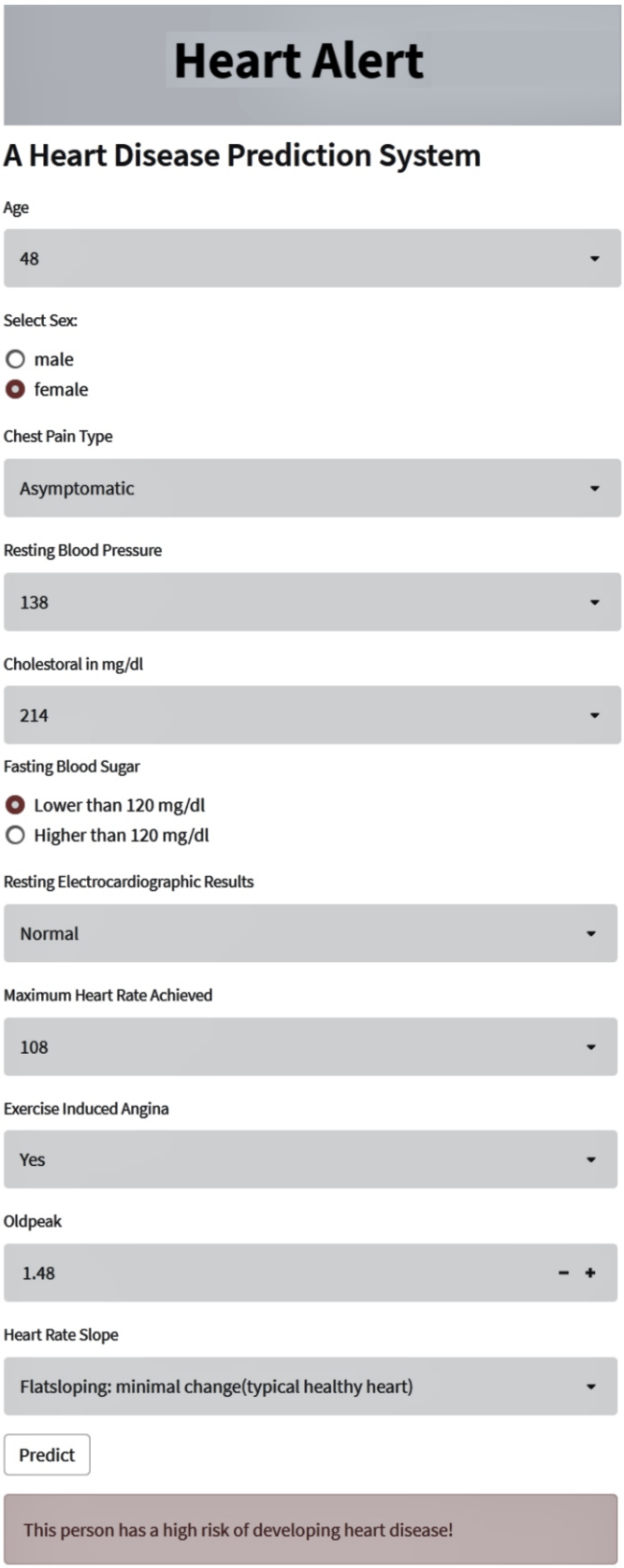


Figure 2. *Heart Alert System - for Heart Disease Prediction Output. (A) Low Risk for Heart Disease, (B) High Risk for Heart Disease.*

deaths. Lifestyle-related habits such as alcohol consumption, smoking, poor diet and nutrition, high sedentary behavior, and obesity have been increasingly implicated in the high rates of heart disease among Filipinos leading to a significant burden on the country's healthcare system. Six machine learning algorithms aiming to predict the risk of a person having heart disease were tested. An assessment of the effects of different optimization techniques as to the imputation and feature selection methods was also done. The best-performing model was obtained by support vector machine with no imputation nor feature selection technique obtaining a 90.2% accuracy, 87.7% sensitivity, 93.6% specificity, 94.9% precision, 91.2% F1-score, and an AUC of 0.902. Following very closely were random forest with mean or median imputation and logistic regression with mode imputation, all having no feature selection which also performed well. The performance of the best four machine learning models suggests that for this dataset, an imputation technique for missing values may or may not be applied. Additionally, a recursive feature elimination for feature selection need not be applied as all variables seem to be important in heart disease prediction. An early accurate diagnosis leading to prompt intervention efforts is very crucial as it improves the patient's quality of life and diminishes the risk of developing cardiac events. The potential deployment of these machine learning models in clinical practice can further enhance the diagnostic acumen of health professionals. The primary limitation of this research is the use of small datasets due to the unavailability of large and open-source cardiovascular datasets.

Future enhancement of this work should focus on the inclusion of other techniques for imputation and other feature selection methods. The determination of feature importance of attributes coupled with the interpretability of ML results using methods of explainable AI for better understanding of health professionals is also highly recommended Likewise, datasets combining symptoms, clinical laboratory results, and/or cardiac imaging features such as echocardiography or myocardial perfusion imaging can also be explored to generate superior diagnostic accuracy in predicting heart disease. It is also recommended to test the web application among Filipino patients to determine its suitability in local clinical practice. These findings are promising and have generated useful insights in the development of automated models with high accuracy and reliability which can be of use to health professionals in heart disease assessment.

References

1. Philippine Statistics Authority. (2021) Causes of death in the Philippines. https://psa.gov.ph/content/causes-deaths-philippines-preliminary-january-december-2021
2. Cacciata MC, Alvarado I, Jose MM, Evangelista LS. (2021) Health determinants and risk factors for coronary artery disease among older Filipinos in rural communities. European journal of cardiovascular nursing, 20(6):565–571. https://doi.org/10.1093/eurjcn/zvaa039
3. Akella A, Akella S. (2021) Machine learning algorithms for predicting coronary artery disease: efforts toward an open-source solution. Future science OA, 7(6), FSO698. https://doi.org/10.2144/fsoa-2020-0206
4. Baashar Y, Alkawsi G, Alhussian H, *et al.* (2022) Effectiveness of Artificial Intelligence Models for Cardiovascular Disease Prediction: Network Meta-Analysis, Vol 2022, Article ID 5849995. https://doi.org/10.1155/2022/5849995
5. Ghosh P, Azam S, Jonkman M, *et al.* (2021) Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. IEEE Access, Access, IEEE, 9 : 1 9 3 0 4 – 1 9 3 2 6 . https://doi.org/10.1109/ACCESS.2021.3053759
6. Magboo VPC, Magboo MSA. (2021) Machine Learning Classifiers on Breast Cancer Recurrences. Procedia Computer Science 192 2742–2752. ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.09.044
7. Kavitha M, Gnaneswar G, Dinesh R, Sai YR, Suraj RS. (2021) Heart Disease Prediction using Hybrid machine Learning Model, 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597
8. Magboo MSA, Coronel AD. (2019) Data Mining Electronic Health Records to Support Evidence-Based Clinical Decisions. In: Chen YW., Zimmermann A., Howlett R., Jain L. (eds) Innovation in Medicine and Healthcare Systems, and Multimedia. Smart Innovation, Systems and Technologies, vol 145. Springer, Singapore. https://doi.org/10.1007/978-981-13-8566-7_22
9. Patel J, Khaked AA, Patel J, Patel J. (2021) Heart Disease Prediction Using Machine Learning. In: Singh, P.K., Wierzchoń, S.T., Tanwar, S., Ganzha, M., Rodrigues, J.J.P.C. (eds) Proceedings of Second International Conference on Computing, Communications, and Cyber-Security. Lecture Notes in Networks and Systems, vol 203. Springer, Singapore.

https://doi.org/10.1007/978-981-16-0733-2_46
10. Alsafi HES, Ocan ON. (2021) A novel intelligent machine learning system for coronary heart disease diagnosis. Appl Nanosci. https://doi.org/10.1007/s13204-021-01992-4
11. Yazdani A, Varathan KD, Chiam YK, *et al.* (2021) A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med Inform Decis Mak 21, 194. https://doi.org/10.1186/s12911-021-01527-5.
12. Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. (2021) Prediction of heart disease using a combination of machine learning and deep learning. Computational Intelligence and Neuroscience, 2021. vol. 2021, Article ID 8387680, 11 pages. https://doi.org/10.1155/2021/8387680
13. Lin Y. (2021) Prediction and Analysis of Heart Disease Using Machine Learning. 2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI), Robotics, Automation and Artificial Intelligence (RAAI), 2021 IEEE International Conference On, 53–58. https://doi.org/10.1109/RAAI52226.2021.9507928
14. Khdair, Hisham (2021) Exploring Machine Learning Techniques for Coronary Heart Disease Prediction. International Journal of Advanced Computer Science and Applications, 12 (5):28-36
15. Patro SP, Padhy N, Chiranjevi D. (2021) Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning. Evolutionary Intelligence, 14(2),941. https://doi.org/10.1007/s12065-020-00484-8
16. Manu Siddhartha. (2020) Heart Disease Dataset (Comprehensive). IEEE Dataport. https://dx.doi.org/10.21227/dz4t-cm36
17. Wang H, Tang J, Wu M. *et al.* (2022) Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. BMC Med Inform Decis Mak 22, 13. https://doi.org/10.1186/s12911-022-01752-6
18. Emmanuel T, Maupong T, Mpoeleng D, *et al.* (2021) A survey on missing data in machine learning. J Big Data 8, 140. https://doi.org/10.1186/s40537-021-00516-9
19. Khan SI, Hoque ASML. (2020) SICE: an improved missing data imputation technique. J Big Data 7, 37 https://doi.org/10.1186/s40537-020-00313-w
20. Laqueur HS, Shev AB, Kagawa R. (2022) SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations. American journal of epidemiology, 191(3): 516–525. https://doi.org/10.1093/aje/kwab271
21. Beesley LJ, Bondarenko I, Elliot MR, Kurian AW, Katz, SJ, Taylor JM. (2021) Multiple imputation with missing data indicators. Statistical methods in medical research, 30(12), 2685–2700. https://doi.org/10.1177/09622802211047346
22. Javadi S, Bahrampour A, Saber MM, Garrusi B, Baneshi MR. (2021) Evaluation of Four Multiple Imputation Methods for Handling Missing Binary Outcome Data in the Presence of an Interaction between a Dummy and a Continuous Variable. Journal of Probability and Statistics, vol. 2021, Article ID 6668822, 14 pages, 2021. https://doi.org/10.1155/2021/6668822
23. Magboo VPC, Magboo MSA. (2021). Imputation Techniques and Recursive Feature Elimination in Machine Learning Applied to Type II Diabetes Classification. In 2021 4th Artificial Intelligence and Cloud Computing Conference (AICCC '21). Association for Computing Machinery, New York, NY, USA, 201-207. https://doi.org/10.1145/3508259.3508288
24. Demircioğlu A. (2021) Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights Imaging 12, 172. https://doi.org/10.1186/s13244-021-01115-1
25. Magboo VPC, Magboo, MSA. (2022). Prediction Models for COVID-19 in Children. In: Chen Y-W *et al.* (eds.) Innovation in Medicine and Healthcare. Smart Innovation, Systems and Technologies, vol 308. Springer, Singapore. https://doi.org/10.1007/978-981-19-3440-7_2
26. Chang W, Ji X, Wang L, *et al.* (2021) A Machine-Learning Method of Predicting Vital Capacity Plateau Value for Ventilatory Pump Failure Based on Data Mining. Healthcare 9, 1306. https://doi.org/10.3390/healthcare9101306
27. Li D, Wang Y, Hu W, *et al.* (2021) Application of Machine Learning Classifier to Candida auris Drug Resistance Analysis. Frontiers in Cellular and Infection Microbiology, vol 11. https://www.frontiersin.org/article/10.3389/fcimb.2021.742062
28. Magboo VPC, Abu PAR. (2022) Deep Neural Network for Diagnosis of Bone Metastasis. In 2022 The 5th International Conference on Software Engineering and Information Management (ICSIM) (ICSIM 2022). Association for Computing Machinery, New York, NY, USA, 144-151. https://doi.org/10.1145/3520084.3520107

APPENDIX

Appendix A. *Description of Variables of Comprehensive Heart Disease Dataset*

Attribute	Variable Type	Description	Feature Values
Sex	Categorical	Gender of patient	0: Female, 1: Male
Chest Pain Type	Categorical	Type of chest pain experience by patient	1 : Typical 2 : Atypical Angina 3 : Non-anginal pain 4 : Asymptomatic
Fasting Blood Sugar	Categorical	Blood sugar levels on fasting >120 mg/dl	0: False 1: True
Resting ECG	Categorical	Result of Electrocardiogram while at rest	0: normal 1: ST-T wave abnormality 2: ventricular hypertrophy
Exercise Angina	Categorical	Exercise induced angina	1: yes 0: no
ST Slope	Categorical	ST segment measured in terms of slope during peak exercise	1: Upsloping 2: Flat 3: Downsloping
Target	Categorical		0: no disease 1: disease
Age	Numeric	Age of patient in years	
Resting BP	Numeric	Resting blood pressie (in mmHg on admission to the hospital)	
Cholesterol	Numeric	Serum cholesterol in mg/dl	
Max Heart Rate	Numeric	Maximum heart rate achieved	
Old Peak	Numeric	ST depression induced by exercise relative to rest	