Journal of Academic Research
Vol 28, Issu 2, 2024

Optimization of Heart Disease Prediction using Supervised learning with Hyperparameter Optimization Methods

Farij O. Ehtiba[*1], Suhaila F. Elfaitouri[1], Haitham S. Ben Abdelmula[2], and Ali Elghirani[3]

[1] Libyan Academy for Postgraduate Studies, School of Basic Sciences, Department of Computer Science - Misurata, Libya.
[2] Computer Networks Department, College of Computer Technology – Zawia, Zawia, Libya.
[3] Faculty of Information Technology, Libyan International Medical University, Benghazi – Libya.

Article information

Key words
Hyper parameter tuning, Grid search, cardiovascular diseases, features selection, machine Learning.

Received 20/07/2024, Accepted 14/08/2024, Available online 24/08/2024,

Abstract

Coronary artery disease prediction is a challenging task in healthcare due to the increasing mortality rate associated with heart disease globally. Various machine learning techniques, such as logistic regression, support vector machine, K-nearest neighbors, and random forests, have been employed to create predictive models for early detection. By fine-tuning the hyperparameters, the accuracy of these models can be significantly enhanced. The study outcomes demonstrated different levels of accuracy for each algorithm, with logistic regression achieving up to 86.13% accuracy, support vector machine up to 85.71%, Random Forest exhibiting superior performance with 91.60% accuracy, and K-Nearest Neighbors emerging as the top performer with 92.44% accuracy. This research underscores the potential of utilizing relatively simple supervised machine learning algorithms to predict heart disease with exceptional accuracy, highlighting their significant utility in healthcare.

I. Introduction

Heart disease is a significant global health concern, contributing to a large number of fatalities annually [1]. The heart is the most critical organ of the human body. Cardiovascular diseases have become a significant public health concern across the nations due to their high treatment costs and time-consuming. The World Health Organization (WHO) Heart Association statistics demonstrate that 17.3 million heart patients are expected to die a year and that they should reach around 23.6 million by 2030 [2, 3]. The seriousness of the issue lies in the lack of capabilities and the unavailability of doctors and radiologists. In developing countries or low-income countries, misdiagnosis is also due to doctors' lack of experience and insufficient knowledge, which is one of the factors that makes the survival rate reach 50% of cardiovascular patients. It can be said that cardiovascular disease is a condition in which the heart does not function properly to pump the required amount of blood to other parts of the body causing heart failure [4]. The main cause of heart disease is a blockage in arteries. When these vessels get blocked or narrowed, it can lead to any heart stroke [5]. However, people who are at risk for cardiovascular disease may have high blood pressure or be overweight or high cholesterol, stress, tension, consumption of alcohol, sedentary lifestyle, obesity, and

F.ehtiba@lam.edu.ly

diabetes are the major factors that affect the heart. If the proper diagnosis is made at the appropriate time, the heart failure death rate can always be reduced. Coronary artery disease prediction is considered to be one of the most challenging tasks in the healthcare industry [6]. Timely and accurate identification of heart disease is crucial for effective treatment leading to improved patient outcomes and reduced mortality rates. With advancements in technology, particularly in the field of machine learning [6]. There is a growing interest in leveraging this Heard disease has received much attention in medical research. Healthcare organizations collect vast amounts of data related to heart disease which, unfortunately, is not "mined" to discover hidden information to make effective decisions [7]. Techniques to enhance healthcare services, including the diagnosis and prediction of various medical conditions, Machine Learning, a subset of maintaining the integrity of the specification's Aartificial Intelligence, involves the development of algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed. By leveraging machine learning techniques, it is possible to analyze vast amount of medical data, including patient demographics, clinical symptoms, laboratory tests [8], and imaging results. To identify patterns and trends associated with heart disease, this research presents a model that aims to improve the accuracy of the automated model for predicting heart disease in its early stages. A machine learning models with tuning the hyperparameters the best hyperparameters are selected and applied to the ML models, and the performance is improved. Robust models can be built by adjusting the hyperparameters. Overfitting or underfitting can be prevented by tuning the hyperparameters. is presented using a Grid search technique to achieve this goal an experiment was conducted to assess the effectiveness of machine learning approaches, where cardiology datasets were collected from the IEEE data repository. This dataset was selected from explicit data and compiled by merging five well-known cardiac treatment datasets (Long Shoreline VA, Hungarian, Cleveland, Starlog, and Switzerland). Four algorithms were used to build the predicted models for this experiment, using the socialist dataset (Support Vector Machine, Nearest Neighbor, Logistic Analysis, and Random Forest). This study was compared to the highest published research. The paper is organized as follows, an overview of related work is presented in Section 2, and Section 3 discusses the methodology. Research results and discussions are presented in Section 4. Finally, the conclusion is presented in Section 5.

II. RELATED WORKS

A novel approach was introduced, which aimed at identifying significant features using ML techniques to enhance the accuracy of cardiovascular disease prediction [8]. The proposed prediction model employs various combinations of features and multiple well-established classification techniques. Through the hybrid random forest with a linear model (HRFLM), an improved performance level was achieved, attaining an accuracy of 88.7% in predicting heart disease.

A clinical support system was proposed as an aid to medical specialists to predict and diagnose heart diseases, and make the best decisions [9]. Some ML algorithms were applied in this study such as Naïve Bayes and KNN, SVM, RF, and DT to predict heart failure disease using risk factor data retrieved from medical files. Several experiments have been performed to predict the use of the HD UCI dataset, and the best result with NB when using both Cross-

validations with an accuracy of 82.17 percent and split-test training and with an accuracy of 84.28 percent.

Early detection of heart disease has been proposed for the potential to save numerous lives [10]. ML offers an effective approach for decision-making in the context of heart disease. In study, the Cleveland heart disease dataset was utilized, and various ML algorithms, including Random Forest, Decision Tree, and a Hybrid model (combining Random Forest and Decision Tree), were employed. The experimental findings demonstrate an accuracy level of 88.7% in the heart disease prediction model using the hybrid approach.

The prediction of coronary heart disease (CHD) was investigated using a range of predictive methodologies including K- Nearest Neighbors, Binary Logistic Classification, and Naive Bayes [11]. Furthermore, it explores ensemble modeling techniques such as bagging, boosting, and stacking in comparison to traditional classifiers to improve prediction accuracy. Analysis of the 'Cardiovascular Disease Dataset' containing 70,000 patient records reveals a notable 1.96% increase in accuracy with bagged models compared to conventional methods. Boosted models achieve an average accuracy of 73.4%, with the highest AUC score reaching 0.73 .However, the stacked model, integrating KNN, random forest classifier, and SVM emerges as the most effective, achieving a final accuracy of 75.1%.

Heart diseases were detected using various machine learning algorithm and were implemented for analysis of the heart attack in minimum time using dataset from Kaggle [12]. KNN and Logistic regression algorithm were implementated for heart disease detection. The results of disease detection using KNN and Logistic regression algorithm show effective which KNN attained 71.4% accuracy, Logistic regression attained 87.9% accuracy and Hybrid algorithm attained 89%.

The study intended to provide effective heart disease prediction system (EHDPS) which developed a neural network system for predicting the risk level of heart disease [13]. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The back propagation multi-layer perceptron neural network has been employed as the training algorithm. The experimental results show that the system using neural networks predicts heart disease with 100% accuracy.

Supervised machine learning algorithms namely SVM, KNN, and Naive Bayes are used to predict heart diseases [14]. The machine learning algorithms were implemented and the results show that Naive Bayes algorithm predicts the heart disease with the accuracy of 86.6%.

III. MATERIALS AND METHOD

A. Proposed methodology

Original datasets were collected and data preprocessing is done on the collected data. Relevant features were selected using Chi squared tests and prediction methods such as KNN, LR, SVM, and RF were tuned using the hyper parameter optimization techniques Grid search. The models were validated and analyzed to predict the heart disease. In the proposed model, 5- cross validation is used to validate the data. Figure 1 shows the flowchart of the proposed model.

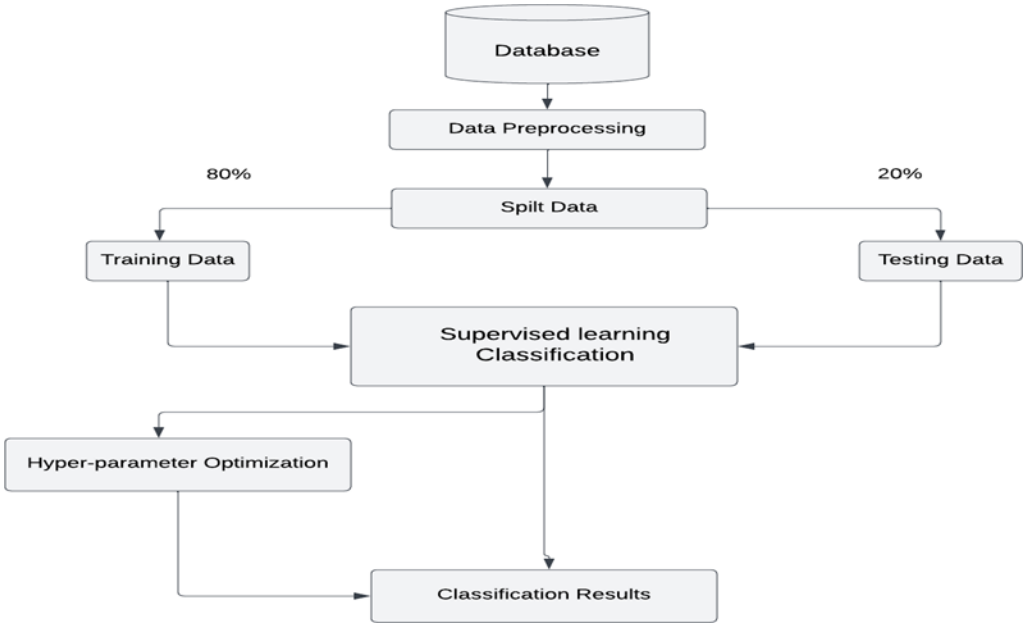

Figure 1: The architecture of the proposed model

B. Dataset Description

This cardiovascular disease dataset is a consolidation of five individual datasets, these datasets, previously known as Long Shoreline VA. Hungarian, Cleveland, Starlog, and Switzerland, were initially separate entities. However, they have been merged to create a comprehensive resource for scientific inquiry. With a combined total of 1190 samples, each comprising 12 attributes. This integrated dataset offers a rich and extensive resource for research purpose. Combining these datasets offers researchers a comprehensive perspective on various factors associated with heart disease. This amalgamation provides access to a more diverse and complete of cardiovascular health each centered around heart-related conditions as described in Table 1.

C. Preprocessing of Dataset

Preprocessing the dataset is crucial for ensuring a robust representation. Various techniques, including the removal of attribute missing values, as well as StandardScaler (SS) and Min-Max Scaler normalization have been employed to enhance the dataset's quality and usability.

Table1: Attributes of the dataset

Attribute	Code given	Unit	Data Type
Sex	Sex	1, 0	Binary
Age	Age	In years	Numeric
Chest pain type	Chest pain	1, 2, 3, 4	Nominal

	type		
Resting blood pressure	Resting bp s	In mm Hg	Numeric
Exercise-induced angina	Exercise angina	0, 1	Binary
Serum Cholesterol	Cholesterol	In mg/dl	Numeric
old peak = ST	Old peak	depression	Numeric
The slope of the peak exercise ST segment	ST slop	0, 1, 2	Nominal
Resting electrocardiogram result	Resting ECG	0, 1, 2	Nominal
Fasting blood sugar	Fasting blood sugar	1,0>120 mg/dl	Binary
Maximum heart rate achieved	Max heart rate	71-202	Numeric
Class	target	0,1	Binary

D. Feature selection

After data pre-processing, the selection of features becomes necessary for the subsequent process. Generally, feature selection (FS) plays a crucial role in constructing a classification model. This step involves reducing the number of input features in a classifier, aiming to improve its performance. One common method used for feature selection is employing Chi-squared tests [15], as shown in Figure 2.

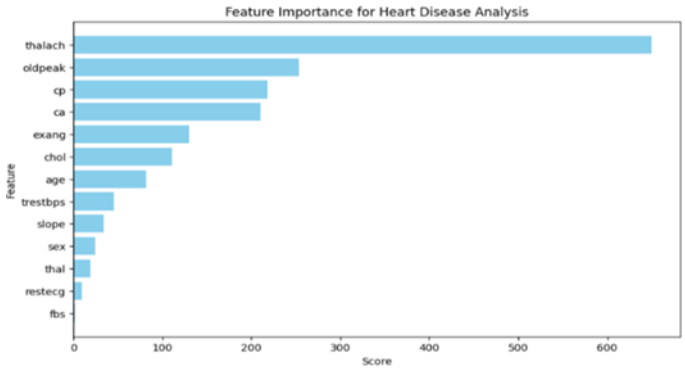

Figure 2: Feature Importance

E. Supervised Learning Classification

Based on Figure 1, four classifier techniques have been selected to predicate the input data. This section gives brief discussion about them.

- K-nearest Neighbor (KNN).

  KNN is a simple and intuitive classification algorithm that works by finding the most similar instances (nearest neighbors) to a given data point based on distance metrics (e.g., Euclidean distance) [18]. It assigns the majority class among its k nearest neighbors to the data point as its predicted class.

- Logistic Regression (LR)

Logistic Regression is a statistical method used for binary classification tasks. Despite its name, it is primarily used for classification rather than regression [19]. It models the probability that a given data point belongs to a particular class using a logistic function, which outputs values between 0 and 1 [20].

- Random Forest (RF)

Random Forest is an ensemble learning method based on decision trees[16, 17]. It builds multiple decision trees during training and combines their predictions to obtain a more accurate and stable result. Each decision tree in the forest is trained on a random subset of the training data and features, which helps to reduce overfitting and improve generalization.

- Support Vector Machine (SVM).

SVM is a powerful classification algorithm that works by finding the hyperplane that best separates the data points of different classes. It aims to maximize the margin between the hyperplane and the nearest data points (support vectors) [17, 21]. SVM can handle both linear and non-linear classification tasks using different kernel functions.

F. Hyper parameter Optimization

Hyper parameter optimization is crucial in machine learning to enhance model performance. Hyperparameters are parameters that are set before the learning process begins, unlike model parameters [22], which are learned from the data There are different types of hyperparameters in ML algorithms that must be tuned to improve the result as shown in Table 2.

Table 2. - ML Hyperparameters used in Grid Search

Classifiers	Hyperparameter	Definition	Default
SVM	C	Regularization parameter.	1
	Gamma	Kernel coefficient for ‘rbf’, ‘poly’, and ‘sigmoid’.	scale
KNN	n-neighbors	n-neighbors Number of neighbors	5
	Weights	Weight function used in prediction	uniform
	Metric	Metric to use for distance computation	manhattan
RF	n_estimators	The number of trees in the forest	100
	max_features	The number of features to consider when looking for the best split	sqrt
	max_depth	The maximum depth of the tree.	None
	min_samples_leaf	The minimum number of samples required to be at a leaf node	2

	min_samples_leaf	The minimum number of samples required to be at a leaf node	1
LR	C	Regularization parameter	0.1
	Solver	Optimization algorithm for logistic regression	'lbfgs'

IV. Evaluation metrics

The proposed model was assessed using four performance metrics: accuracy, precision, recall, and the F-measure. These metrics were calculated as shown in questions (1), (2), and (3) respectively [11]. The evaluation technique employed involved the use of a confusion matrix, also known as a contingency table.

$$Accuracy = {}^{TP}/_{(TP + FP)} \quad (1)$$

$$Precision = {}^{TP}/_{(TP + FP)} \quad (1)$$

$$Recall = {}^{TP}/_{(TP + FN)} \quad (2)$$

$$F - Measure = {}^{(TN + TP)}/_{(TN + TP + FN + FP)} \quad (4)$$

V. Result and Discussion

After preprocessing the data, it's essential to split the dataset into a training set and a testing set with a ratio of 80% for training and 20% for testing. This split allows for modeling on the training data and evaluating the model's performance data. Table 3 presents the classification results obtained using SVM, KNN, LR, and RF classifiers, highlighting the best performance achieved with the RF the best results, as shown in Figure 3.

Table 3: Classification results

Classifiers	Acc	Precision	Recall	F1-score
SVM	85.71	88.80	84.73	86.72
KNN	92.44	93.13	93.13	93.13
LR	86.13	87.12	87.79	87.45
RF	91.60	93.70	90.84	92.25

After finishing the classification stage and before moving on to the optimization stage, with eight features identified, the hyperparameter ranges for all classifiers must be chosen as indicated in Table 4, the optimization phase starts with the application of Grid Search (RS) to

the dataset using all classifiers. The method is utilized to determine the optimal values for the parameters specified for each classifier, which significantly impact the classification outcome.

Table 4 : Classifiers Hyperparameter values in experiments

Classifiers	Hyperparameter	Values used in experiments
SVM	C	[0.1,1,10,100]
	Gamma	[0.1,0.01,0.001,0.0001]
KNN	n-neighbors	(1,30)
	Metric	['euclidean', 'manhattan', 'minkowski']
LR	C	[15,10,100]
	Solver	solver': ['liblinear]'
RF	n_estimators	[100,200,50]
	max_features	[None, 10, 20]
	min_samples_leaf	[2, 6, 10,14,16,20]
	min_samples_leaf	[1, 2, 4]

Table 5 presents the best test accuracy achieved along with the optimal values of hyperparameters for each classifier. Additionally, it highlights the best result obtained when using the Random Forest (RF) classifier show in Table (6) Result using hyperparameter tuning.

Table 5: Optimization result

Classifier	Hyperparameter	Optimal values	Acc
svm	C	10	85.71%
	Gamma	0.01	
	'kernel'	'rbf'	
KNN	n-neighbors	22	92.4%
	Metric	Manhattan 92.44%	
LR	C	10	86.13%
	Solver	'liblinear'	
RF	n_estimators	100	91.60%
	max_features	10	
	min_samples_leaf	1	
	min_samples_leaf	2	

Before hyperparameter tuning, the performance of the classifiers varied significantly, as shown in Figure 4. Logistic Regression and Support Vector Machine had lower accuracies compared to Random Forest and K-Nearest Neighbors. However, after tuning the hyperparameters, all classes showed improvements in their performance metrics Random Forest achieved an accuracy of 91.60% before and after tuning, with consistent precision, recall, and F1.

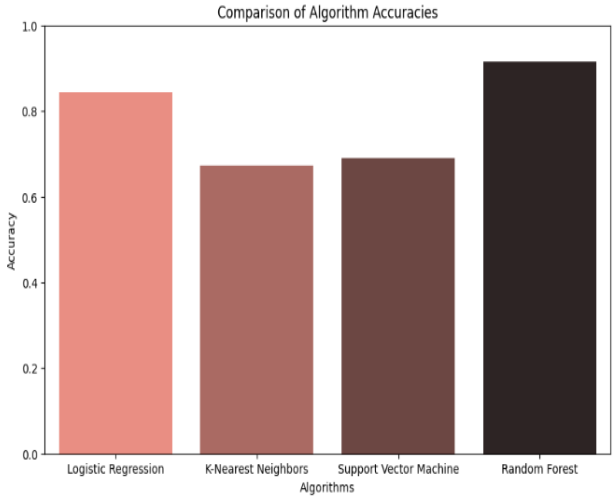

Figure 3: Results of Various Class

LR showed an increase in accuracy, from 84.45% to 86.13% after tuning, along with improvements in precision, recall, and F1 score. K- Nearest Neighbors had the most significant improvement, with accuracy increasing from 67.23% to 92.44% after tuning, resulting in higher precision, recall, and F1 score. Likewise, Support Vector Machine's accuracy increased from 68.91% to 85.71% after tuning, with improvements in precision, recall, and F1 score. Finally, Random Forest is the best algorithm for this dataset, followed by Logistic Regression and Support Vector Machine. K-Nearest Neighbors has the lowest performance across all metrics.

Table 6: Results using hyperparameter tuning

Classifiers	Acc	Precision	Recall	F1- score
SVM	85.71	88.80	84.73	86.72
KNN	92.44	93.13	93.13	93.13
LR	86.13	87.12	87.79	87.45
RF	91.60	93.70	90.84	92.25

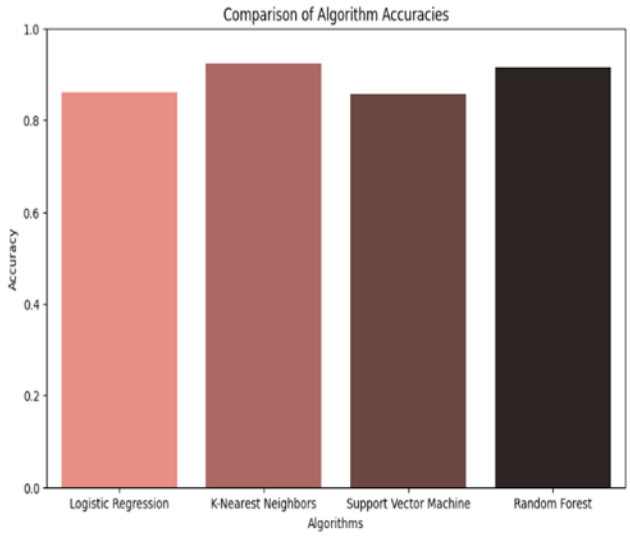

Figure 4: Results using hyperparameter tuning

VI. CONCLUSION

In conclusion, heart disease remains a significant global health concern, with high mortality rates and substantial economic burdens. The timely and accurate prediction of heart disease is crucial for effective treatment and prevention strategies. In this study, machine learning techniques were employed to develop predictive models for early detection of heart disease. By leveraging algorithms such as LR, SVM, K-NN, and RF. A comprehensive dataset comprising patient demographics, clinical symptoms, and diagnostic tests were analyzed, before hyperparameter tuning, the classifiers exhibited varying levels of performance, with LR and SVM lagging behind RaFand K-NN. However, after tuning the hyperparameters using Grid Search, significant improvements were observed across all classifiers. RF maintained high accuracy of 91.60% both before and after tuning, demonstrating consistent precision, recall, and F1 score. LR and SVM also showed notable enhancements in accuracy, with LR increasing from 84.45% to 86.13% and SVM increasing from 68.91% to 85.71% after tuning. These improvements were accompanied by enhancements in precision, recall, and F1 score for both classifiers. K-NN experienced the most substantial improvement, with its accuracy skyrocketing from 67.23% to an impressive 92.44% after tuning. This dramatic increase in accuracy translated into higher precision, recall, and F1 score, highlighting the effectiveness of hyperparameter tuning in optimizing its performance. Based on the post-tuning results, RF emerged as the best-performing algorithm for the dataset, followed closely by LR and SVM. Meanwhile, K-NN despite significant improvement, still lagged behind the other classifiers in terms of performance metrics. Overall, the optimization phase, through hyperparameter tuning using Grid Search, proved to be instrumental in enhancing the classification outcomes of all classifiers, ultimately leading to more accurate and reliable predictions on the dataset. The findings underscore the importance of hyperparameter tuning heart disease prediction.

References

[1] Singh, A. and R. Kumar. Heart disease prediction using machine learning algorithms. in 2020 international conference on electrical and electronics engineering (ICE3). 2020. IEEE.

[2] Terrada, O., et al., Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease. Advances in Science, Technology and Engineering Systems Journal, 2020. 5(5): p. 269-277.

[3] Yadav, S.S., et al. Automated cardiac disease diagnosis using support vector machine. in 2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA). 2020. IEEE.

[4] Gupta, A., et al., MIFH: A machine intelligence framework for heart disease diagnosis. IEEE access, 2019. 8: p. 14659-14674.

[5] Alotaibi, F.S., Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 2019. 10(6).

[6] Kavitha, M., et al. Heart disease prediction using hybrid machine learning model. in 2021 6th international conference on inventive computation technologies (ICICT). 2021. IEEE.

[7] Chetankumar, Y.S., D.K. Singh, and S. Menaria, A review on: heart disease prediction using data mining. J Emerg Technol Innov Res (JETIR), 2019.

[8] Mohan, S., C. Thirumalai, and G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 2019. 7: p. 81542-81554.

[9] El Hamdaoui, H., et al. A clinical support system for prediction of heart disease using machine learning techniques. in 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). 2020. IEEE.

[10] 10. Valarmathi, R. and T. Sheela, Heart disease prediction using hyper parameter optimization (HPO) tuning. Biomedical Signal Processing and Control, 2021. 70: p. 103033.

[11] Shorewala, V., Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 2021. 26: p. 100655.

[12] Morya, R. and S. Singh. Prediction model of heart diseases based on hybrid model. in Journal of Physics: Conference Series. 2022. IOP Publishing.

[13] Singh, P., S. Singh, and G.S. Pandi-Jain, Effective heart disease prediction system using data mining techniques. International journal of nanomedicine, 2018. 13(sup1): p. 121-124.

[14] Anitha, S. and N. Sridevi, Heart disease prediction using data mining techniques. Journal of analysis and Computation, 2019.

[15] Ray, S., et al. Chi-squared based feature selection for stroke prediction using AzureML. in 2020 Intermountain Engineering, Technology and Computing (IETC). 2020. IEEE.

[16] Sekulić, A., et al., Random forest spatial interpolation. Remote Sensing, 2020. 12(10): p. 1687.

[17] Boateng, E.Y., J. Otoo, and D.A. Abaye, Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. Journal of Data Analysis and Information Processing, 2020. 8(4): p. 341-357.

[18] Xiong, L. and Y. Yao, Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm. Building and Environment, 2021. 202: p. 108026.

[19] Das, A., Logistic regression, in Encyclopedia of Quality of Life and Well-Being Research. 2024, Springer. p. 3985-3986.

[20] Schober, P. and T.R. Vetter, Logistic regression in medical research. Anesthesia & Analgesia, 2021. 132(2): p. 365-366.
[21] Cervantes, J., et al., A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 2020. 408: p. 189-215.
[22] Yang, L. and A. Shami, On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 2020. 415: p. 295-316.
[23] Manu Siddhartha. (2020). Heart Disease Dataset (Comprehensive). IEEE Dataport. https://dx.doi.org/10.21227/dz4t-cm36