Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 694-705
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease

Keshab R. Dahal[1*], Yadu Gautam[2]

[1]Department of Statistics, Truman State University, Kirksville, MO, USA
[2]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
Email: *kdahal@truman.edu

How to cite this paper: Dahal, K.R. and Gautam, Y. (2020) Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease. *Open Journal of Statistics*, **10**, 694-705.
https://doi.org/10.4236/ojs.2020.104043

Received: July 25, 2020
Accepted: August 16, 2020
Published: August 19, 2020

Abstract

Cardiovascular disease (CVD) is a leading cause of death across the globe. Approximately 17.9 million of people die globally each year due to CVD, which comprises 31% of all death. Coronary Artery Disease (CAD) is a common type of CVD and is considered fatal. Predictive models that use machine learning algorithms may assist health workers in timely detection of CAD which ultimately reduces the mortality. The main purpose of this study is to build a predictive model that provides doctors and health care providers with personalized information to implement better and more personalized treatments for their patients. In this study, we use the publicly available Z-Alizadeh Sani dataset which contains random samples of 216 cases with CAD and 87 normal controls with 56 different features. The binary variable "Cath" which represents case-control status, is used the target variable. We study its relationship with other predictors and develop classification models using the five different supervised classification machine learning algorithms: Logistic Regression (LR), Classification Tree with Bagging (Bagging CART), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These five classification models are used to investigate the detection of CAD. Finally, the performance of the machine learning algorithms is compared, and the best model is selected. Our results indicate that the SVM model is able to predict the presence of CAD more effectively and accurately than other models with an accuracy of 0.8947, sensitivity of 0.9434, specificity of 0.7826, and AUC of 0.8868.

Keywords

Machine Learning, Classification Model Comparison, Coronary Artery Disease, Data Mining

1. Introduction

Coronary Artery Disease (CAD) is one of the most common types of Cardiovascular disease (CVD). According to the current statistics from World Health Organization (WHO), 17.9 million of people are dying globally in yearly due to CVD, which is 31% of all deaths. It is the number one cause of death around the world in 2020 [1]. CAD is considered a fatal illness that causes the death of millions of people every year globally. In the United States of America, 365,914 people died because of CAD in 2017. About 18.2 million (6.7%) Americans who are 20 and older have CAD, and CAD is the cause of death for 20% of Americans that are 65 and younger [2]. India is predicted to be the country hardest hit by CAD. By 2020, it is estimated that at least 1.4 million citizens will die of heart disease, and one out of four cardiac patients globally will be Indian [3].

These facts illustrate the importance of dealing with CAD. There have been numerous efforts applied during the previous years to include clinical decision support systems and artificial intelligence to predict the CAD. Such predictive models provide doctors and health care providers with personalized information to implement better and more personalized treatments for their patients. Often health care data are very large with several information collected over a large number of patients, which is impractical to analyze using standard statistical techniques. Machine learning approaches are very powerful and efficient tools to study and analyze such large-scale multi-dimensional dataset. Because of that, for decades, machine learning and the other artificial intelligence have been successfully used and have proven to be helpful in medicine [4].

Several studies have been conducted using various machine learning algorithms and different datasets in order to detect CAD. In 2019, M. Abdar *et al.* [5] compared 10 machine learning algorithms to investigate the CAD detection using accuracy and F1-score as the performance matrices. However, the authors did not use sensitivity, specificity, and area under the receiver operating character (ROC) curve, which are also critical information for the model comparison. In 2017, H. Forssen *et al.* [6] compared 6 machine learning algorithms to investigate the CAD detection using accuracy, area under the ROC curve (AUC), sensitivity, and specificity as the performance matrices. The model they preferred has a very low specificity of 0.339. In 2020, A.B. Akella & V. Kaushik [7] showed that neural network is the best machine learning algorithm to detect CAD. Their claim is based on the following matrices: Accuracy = 0.9303, Recall = 0.9380, F1 score = 0.8984, AUC = 0.796, and Mean = 0.88. Based on these matrices, it can be estimated that the specificity of their best model could be less than 0.60. In other words, the false positive rate (FPR) (also known as type I error), is about 40%, which is significantly high. In 2020, I.C. Dipto *et al.* [8] claimed that neural network is the best machine learning algorithm to detect CAD, for the algorithm achieving an average accuracy of 0.9325 and an AUC of 0.98. However, they applied SMOTE algorithm to balance the dataset. In 2012, R. Alizadehsani *et al.* [9] used four machine learning algorithms; Naïve Bayes, C4.5, AdaBoost, and SMO

to detect CAD. However, none of the algorithms could achieve the satisfactory performance.

It has been very challenging to determine which model type to apply to a machine learning task in order to make a precise prediction. Every model has some merits and demerits [10]. It can be difficult to compare the relative merits of the models. In this study, we implement five different supervised classification machine learning approaches to predict the CAD using the publicly available Z-Alizadeh Sani dataset. The five implemented machine learning approaches are: Logistic Regression (LR), Classification Tree with Bagging (Bagging CART), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Lastly, the performance of the algorithms is compared in order to select the best model.

Rest of the article is organized as follow: In Section 2, we discuss data description and preprocessing. In Section 3, a different classification machine learning will be discussed, followed by model comparison and selection of the best model in Section 4. In Section 5, we discuss the results. In Section 6, we summarize the main findings and conclude the manuscript.

2. Data Description and Preprocessing

2.1. Data Source

In this study, we use the publicly available Z-Alizadeh Sani dataset obtained from the UCL Machine Learning Repository, which contains a large collection of datasets that have been widely used by the Machine Learning Community. Detailed information about the dataset such as: name, type, level, and other relevant information are provided [11].

2.2. Data Description

The Z-Alizadeh Sani dataset contains the records of 303 random patients who visited Shaheed Rajaei Cardiovascular, Medicine, and Research Centre of Tehran, Iran. Among the patients who visited, 216 have been diagnosed with CAD and the rest 87 were normal. Every entry of the dataset contains information about the patient such as: age, sex etc. The dataset contains 56 features that are arranged in four groups: demographic, symptoms and examinations, ECG, and laboratory and echo features. The target variable "Cath" is binary with labels "Cad" and "Normal". The "Cad" stands for the presence of CAD, and the "Normal" stands for normal patients. In 2017, the dataset was donated to the UCL Machine Learning Repository [9] [12] [13].

2.3. Feature Selection

The feature with a negligible effect on the response variable is called an irrelevant feature. A common example of an irrelevant feature is a serial number. In predictive modeling, we are often confronted with many inputs (explanatory variables). Some of these inputs may not have any relation to the target variable.

An initial screening can eliminate irrelevant variables and keep the number of inputs to a manageable size [14]. The irrelevant features increase the noise in the dataset. There are different ways to denoise [15]. Dropping irrelevant features is one of the most common ways. There are many feature selection methods that automatically drop the irrelevant features. We have used variable selection node available in SAS Enterprises Miner to drop the irrelevant features because it handles both categorical and numerical variables. We have chosen the Chi-square criteria because our target variable Cath is binary. The brief summary of the variables selected using variable selection node based on Chi-square criteria, including response variable with role, type, level, and range is summarized in Table 1. The relative importance plot of the input variables (those selected using variable selection node) with respect to the target is given in Figure 1.

2.4. Data Partition

The data is split into two parts—training and testing in the ratio 3:1. First, we train the data that contains 227 observations, and then move on to test the data that contains 76 observations. Train data is used to find the relationship between target and predictor variables while the test data assesses the performance of the model. The main purpose of the splitting data is to avoid overfitting. If overfitting

Table 1. Summary of the variable name, role, type, level, and range of those selected using variable selection node.

Variable name	Variable Role	Variable Type	Variable Level	Variable Range
Typical Chest Pain	Input	Characteristic	Nominal	Yes, No
Age	Input	Numerical	Interval	30 - 86
Regional Wall Motion Abnormality (Region RWMA)	Input	Numerical	Discrete	0, 1, 2, 3, 4
Hypertension (HTN)	Input	Characteristic	Nominal	Yes, No
Low Density Lipoprotein (LDL)	Input	Numerical	Interval	18 - 232
Tinversion	Input	Characteristics	Nominal	Yes, No
Nonanginal	Input	Characteristics	Nominal	Yes, No
High Density Lipoprotein (HDL)	Input	Numeric	Interval	35 - 111
Diabetes Mellitus (DM)	Input	Characteristics	Nominal	Yes, No
Current Smoker	Input	Characteristics	Nominal	Yes, No
Potassium (K)	Input	Numeric	Interval	3.0 - 6.6
Body Mass Index (BMI)	Input	Numeric	Interval	18.12 - 40.90
Weight	Input	Numeric	Interval	48.0 - 120.0
Length	Input	Numeric	Interval	140 - 188
Erythrocyte Sedimentation Rate (ESR)	Input	Numeric	Discrete	1 - 90
Cath	Target	Characteristics	Nominal	CAD, Normal

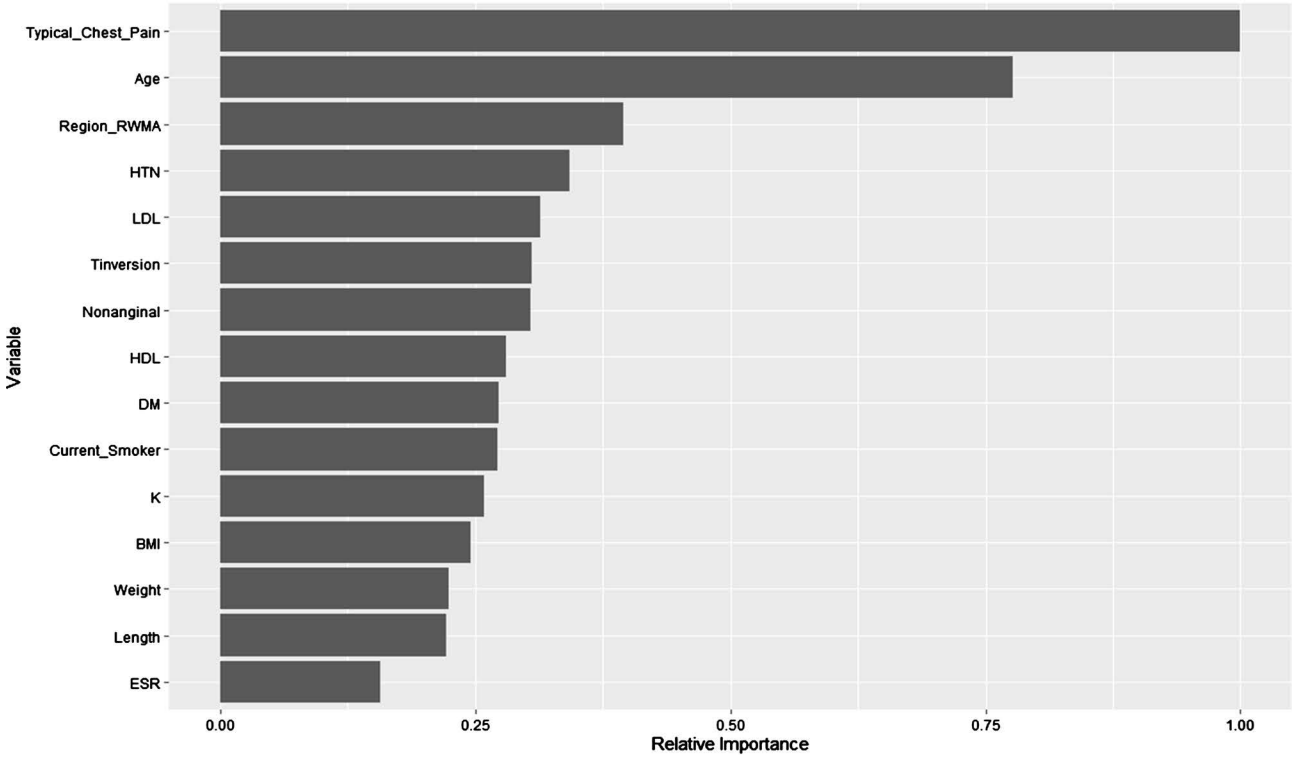

Figure 1. Relative importance plot of the features selected using variable selection node.

occurs, the machine learning algorithm could perform exceptionally in the training dataset, but perform poorly in the testing dataset.

3. Machine Learning Algorithms

There are various machine learning algorithms that are available to solve the classification problems such as: Logistic Regression, Random Forest, and Support Vector Machine. We have implemented the following approaches in this study:

3.1. Logistic Regression

Logistic Regression (LR) model is used for predicting binary outcomes. It is a statistical model that in its basic form uses as a sigmoid function to model a binary response variable, taking on values 1 and 0 with probability π and 1 − π respectively. A logistic regression model is given below as:

$$\operatorname{logit}\left(\Pr\left(Y=1\right)\right)=\beta_0+\sum_{j=1}^{p}X_j\beta_j \tag{1}$$

where,

$$\operatorname{logit}\left(\Pr\left(Y=1\right)\right)=\ln\left(\frac{\Pr\left(Y=1\right)}{1-\Pr\left(Y=1\right)}\right) \tag{2}$$

LR is one of the most popular and commonly used method to solve classification problem, especially when the response variable is binary [16]. The method is simple, and convenience always comes first in the mind of a statistician [17].

We fitted the LR model using the variables selected in the previous step with the help of glm command of R package [18].

3.2. Classification Tree with Bagging

Classification tree (CART) is a powerful alternative to more traditional approaches of land cover classifications. Trees provide a hierarchical and nonlinear classification method and are suited to handling non-parametric training data, as well as categorical or missing data. By revealing the predictive hierarchical structure of the independent variables, the tree allows for great flexibility in data analysis and interpretation [19]. CART is simple and useful for interpretation. It is a statistical model which is used to predict a qualitative response. In this model, we predict that each observation belongs to the most commonly occurring class of training observations in the region which it belongs to. To build the CART model, we used the Gini index in order to evaluate the quality of the split.

CART is a non-robust, meaning that a small departure from the validity of the model effects the performance badly. However, Bagging is a machine learning algorithm obtained by aggregating CART, and causes the predictive performance of the CART to improve substantially. In Bagging, we obtain n bootstrap samples from the existing training data. For each sample, a CART is fitted using all predictors. Finally, the average of the resulting predictions is obtained. Bagging always prevents the model from overfitting [20]. We fitted the classification tree with Bagging (Bagging CART) model using 1000 bootstrap samples using random Forest command of the R package [21].

3.3. Random Forest

Random Forest (RF) is one of the most popular machine learning algorithms. It is obtained by improving the Bagging algorithm since it selects trees that are not correlated. In Bagging, we build several CART on bootstrapped training samples. However, when building these CART using the RF algorithm, a random sample of $m < p$ predictors is chosen to build the model. Each time the collection of m predictors is different so that the CARTs are decorrelated. Typically, we choose the $m = \sqrt{p}$ predictors while building RF [20]. We fitted the RF model with $m = 4$ using 1000 bootstrap samples using random Forest command of the R package [21].

3.4. Support Vector Machine

Support vector machine (SVM) is one of the most popular and powerful machine learning algorithms introduced in the early 90's. It is a supervised machine learning model that can be used to solve both regression and classification problems. SVM is equipped with various kernels, such as: linear, polynomial, radial, and sigmoid. The performance of SVM is based on the actual class boundary and the kernel used. For the linear kernel, the tuning parameters are cost and gamma, which controls the bias-variance trade-off the statistical learning technique.

A small value of these tuning parameters overfits the data whereas a large value underfits [20]. The 10-fold cross validation is used to choose the best tuning parameters. We used the grid technique to find the optimal parameters cost and gamma by varying cost 0.01 to 10 and gamma 0.01 to 1, for which it yields cost and gamma to be 0.02 and 0.01 respectively. A SVM model equipped with the linear kernel using the tuning parameters cost = 0.02 and gamma = 0.01 is fitted by the help of svm command of the R package [22].

3.5. K-Nearest Neighbors

K-Nearest Neighbors (KNN) model takes a completely different approach than the other classification models. To fit KNN model, no assumption is needed. In fact, it is completely nonparametric. KNN can outperform other classification models if the assumptions are not met [16]. In KNN, the parameter k characterizes the tradeoffs between variance and bias. The small and large value of k overfit and underfit the data, respectively. There is not a strong basis for the selection of the value of k [23]. It has been a common practice to choose k equals 10 so we fitted the KNN model using k equals 10 with KNN command of the R package [24].

4. Model Comparisons

To determine which model had the better performance, they were trained on the training dataset and fit to the test dataset where they retrieved the following matrices: Sensitivity, Specificity, Accuracy, and area under the receiver operating characteristic curve (AUC). We compute the confusion matrix for each model as shown in **Table 2**.

The proportion of the actual positive cases that is correctly predicted as positive is called sensitivity. It is also called true positive rate (TPR) and is given in Equation (3).

$$\begin{aligned}\text{Sensitivity} &= \text{True positive rate}\left(\text{TPR}\right)\\ &= \frac{\text{True positive}\left(\text{TP}\right)}{\text{True positive}\left(\text{TP}\right)+\text{False negative}\left(\text{FN}\right)}\end{aligned} \tag{3}$$

The proportion of the actual negative cases that is correctly predicted as negative is called specificity. It is also called true negative rate (TNR) and is given in Equation (4).

$$\begin{aligned}\text{Specificity} &= \text{True Negative rate}\left(\text{TNR}\right)\\ &= \frac{\text{True negative}\left(\text{TN}\right)}{\text{True negative}\left(\text{TN}\right)+\text{False positive}\left(\text{FP}\right)}\end{aligned} \tag{4}$$

Table 2. Confusion matrix.

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

The proportion of the actual negative cases that is incorrectly predicted as positive is called type I error. It is also called false positive rate (FPR) and is given in Equation (5).

$$\begin{aligned}\text{Type I Error} &= \text{False Positive rate}(\text{FPR}) \\ &= \frac{\text{False positive}(\text{FP})}{\text{True negative}(\text{TN}) + \text{False positive}(\text{FP})}\end{aligned} \tag{5}$$

The proportion of the actual positive cases that is incorrectly predicted as negative is called type II error. It is also called false negative rate (FNR) and is given in Equation (6).

$$\begin{aligned}\text{Type II Error} &= \text{False Negative rate}(\text{FNR}) \\ &= \frac{\text{False negative}(\text{FN})}{\text{True positive}(\text{TP}) + \text{False negative}(\text{FN})}\end{aligned} \tag{6}$$

The proportion of the cases that is predicted accurately is called the accuracy and is defined by Equation (7).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{7}$$

There is a direct relation between the sensitivity, specificity, Type I error, and Type II error. Sensitivity is 1-Type II error, whereas specificity is 1-Type I error. Our goal is to minimize both types of errors. In other words, we want sensitivity, and specificity as large as possible.

Sensitivity and specificity are inversely proportional to each other, meaning that as the sensitivity increases, the specificity decreases, and vice-versa.

Receiver operating characteristic (ROC) curve is commonly used to characterize the sensitivity/specificity tradeoffs for a binary classifier. The ROC curve is obtained by plotting the false positive rate (1-specificity) on x-axis against the sensitivity on y-axis at various threshold settings.

Area under the ROC curve (AUC) is one of the most important matrices to measure the performance of the model. Its value lies between 0 and 1. A model is said to be an excellent if its AUC is close to 1. The higher the AUC, the better the model, and vice-versa. We used the roc command of R package to compute the AUC of ROC curve of each model [25].

The model with the highest statistics, which are: sensitivity, specificity, accuracy, and AUC is considered the best model.

5. Results

The summary of the performance statistics from the five models are presented in Table 3. The sensitivity of all models is reasonable. There is not a significant difference in sensitivity among the models. The RF model has highest sensitivity of 0.9623 whereas the KNN model has lowest sensitivity of 0.9245. In Table 3, the highest value of the performance matrices is highlighted. Based on these performance metrices, LR, RF, and SVM models outperformed the Bagging CART

and KNN models. The performance of the SVM model is outstanding because it has an accuracy of 0.8947, sensitivity of 0.9434, specificity of 0.7826, and AUC of 0.8868. The sensitivity of the RF and the AUC of the LR model is slightly higher than SVM, however; their other performance matrices are less than SVM. In fact, SVM has significantly greater values of specificity and accuracy than RF and LR; therefore, neither RF nor LR model can be considered as a better model than SVM.

Figure 2 compares the performances of the five machine learning algorithms by using ROC curves. With the ROC plot, we can visualize the tradeoff between the sensitivity and the specificity. As seen in Figure 2, all models except KNN, perform well. There is a small difference between the ROC curves of the LR, Bagging CART, RF, and SVM. After combining this with the result obtained from the model performance matrices in Table 3, it can be concluded that the SVM is able to predict the presence of CAD more effectively and accurately than other models.

A possible cause of KNN suffering from a poor performance is whenever the class distribution of the Cath is skewed [26]. Most of the voting will raise conflict when there is a huge class that dominates prediction. There will also be a tendency for new data to be voted into additional popular classes. Figure 3 verifies the fact that the number of positive cases (Cad) is almost three times more than the number of negative cases (Normal). As a result, it is unsuitable to use KNN

Table 3. Model performance metrices obtained using test dataset.

Model method	Sensitivity	Specificity	Accuracy	AUC	95% CI for Accuracy
LR	0.9434	0.6957	0.8684	**0.9032**	(0.7713, 0.9351)
Bagging CART	0.9434	0.6522	0.8553	0.8687	(0.7558, 0.9255)
RF	**0.9623**	0.6087	0.8553	0.8782	(0.7558, 0.9225)
SVM	0.9434	**0.7826**	**0.8947**	0.8868	**(0.8031, 0.9534)**
KNN	0.9245	0.2174	0.7105	0.5894	(0.5951, 0.8089)

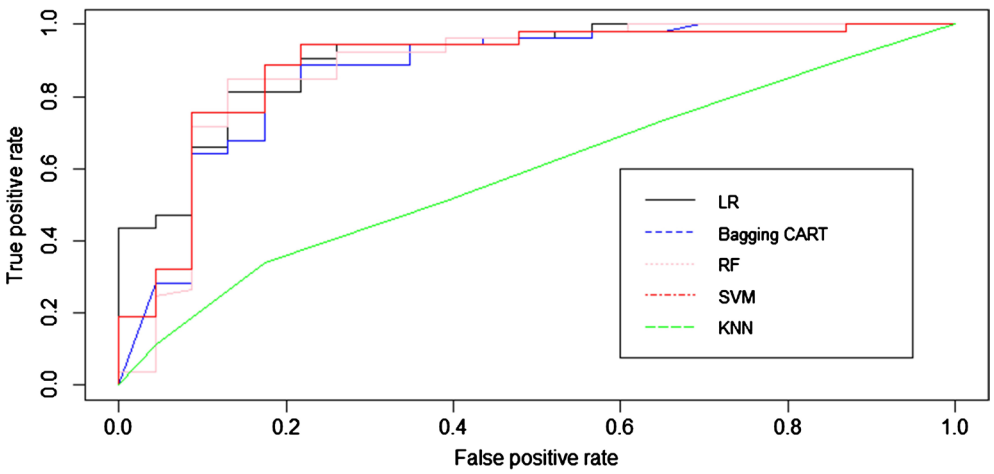

Figure 2. ROC curve for LR, Bagging CART, RF, SVM, and KNN.

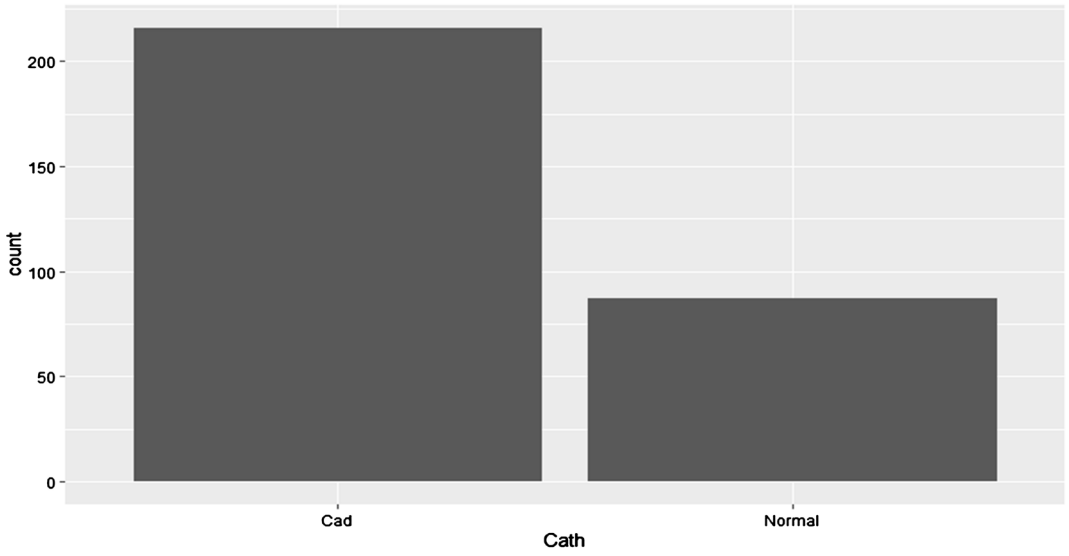

Figure 3. Distribution of target variable Cath.

in this dataset.

A possible cause for the poor performance of Bagging CART and RF when compared to SVM is the fact that the actual class boundary is not complex non-linear. In addition, SVM algorithm outperforms Bagging CART and RF if the dataset has small number of observations with no missing values [27].

6. Conclusion

In conclusion, we used logistic regression (LR), classification tree with Bagging (Bagging CART), random forest (RF), support vector machine (SVM), and k nearest neighbors (KNN) to learn the detection of coronary artery disease (CAD), utilizing the publicly available Z-Alizadeh Sani dataset. The performance of the models is gauged by comparing the following performance matrices: sensitivity, specificity, accuracy, area under the ROC curve (AUC) of the testing data. Our results indicate that the SVM model is able to predict the presence of CAD more effectively and accurately than other models with an accuracy of 0.8947, sensitivity of 0.9434, specificity of 0.7826, and AUC of 0.8868. Further research might be necessary to improve in the performance of the machine learning algorithm before this method translated into clinical solution. Such improvements might include, but are not limited to, using other machine learning algorithms such as artificial neural network, using more data, or exploring other ways of extracting important features before feeding to the machine learning algorithm.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization. Cardiovascular Disease. https://www.who.int/health-topics/cardiovascular-diseases/#tab%20=%20tab_1

[2] CDC Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm#:~:text=Coronary%20Artery%20Disease,killing%20365%2C914%20people%20in%202017.&text=About%2018.2%20million%20adults%20age,have%20CAD%20(about%206.7%25)&text=About%202%20in%2010%20deaths,less%20than%2065%20years%20old

[3] Enas, E.A. and Kannan, S. (2008) How to Beat the Heart Disease Epidemic Among South Asians. A Prevention and Management Guide for Asian Indians and Their Doctors. Downers Grove: Advanced Heart Lipid Clinic USA, 2007. *Indian Heart Journal*, **60**, 161-175.

[4] Dudchenko, A., Ganzinger, M. and Kopanitsa, G. (2020) Machine Learning Algorithms in Cardiology Domain: A Systematic Review. *The Open Bioinformatics Journal*, **13**, 25-40. https://doi.org/10.2174/1875036202013010025

[5] Abdar, M., Książek, W., Acharya, U.R., Tan, R.S., Makarenkov, V. and Pławiak, P. (2019) A New Machine Learning Technique for an Accurate Diagnosis of Coronary Artery Disease. *Computer Methods and Programs in Biomedicine*, **179**, Article ID: 104992. https://doi.org/10.1016/j.cmpb.2019.104992

[6] Forssen, H., Patel, R., Fitzpatrick, N., Hingorani, A., Timmis, A., Hemingway, H. and Denaxas, S. (2017) Evaluation of Machine Learning Methods to Predict Coronary Artery Disease Using Metabolomic Data. *Studies in Health Technology and Informatics*, **235**, 111-115.

[7] Akella, A.B. and Kaushik, V. (2020) Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts toward an Open Source Solution. *BioRxiv*. https://doi.org/10.1101/2020.02.13.948414

[8] Dipto, I.C., Islam, T., Rahman, H.M. and Rahman, M.A. (2020) Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease. *Journal of Data Analysis and Information Processing*, **8**, 41-68. https://doi.org/10.4236/jdaip.2020.82003

[9] Alizadehsani, R., Habibi, J., Sani, Z.A., Mashayekhi, H., Boghrati, R., Ghandeharioun, A. and Bahadorian, B. (2012) Diagnosis of Coronary Artery Disease Using Data Mining Based on Lab Data and Echo Features. *Journal of Medical and Bioengineering*, **1**, 26-29.

[10] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning. Springer, New York, 3-7. https://doi.org/10.1007/978-1-4614-7138-7

[11] UCL Machine Learning Repository (2020) Z-Alizadeh Sani Data Set. https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani#

[12] Alizadehsani, R., Zangooei, M.H., Hosseini, M.J., Habibi, J., Khosravi, A., Roshanzamir, M., *et al.* (2016) Coronary Artery Disease Detection Using Computational Intelligence Methods. *Knowledge-Based Systems*, **109**, 187-197. https://doi.org/10.1016/j.knosys.2016.07.004

[13] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H. and Yarifard, A.A. (2017) Computer Aided Decision Making for Heart Disease Detection Using Hybrid Neural Network-Genetic Algorithm. *Computer Methods and Programs in Biomedicine*, **141**, 19-26. https://doi.org/10.1016/j.cmpb.2017.01.004

[14] Sarma, K.S. (2017) Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications. 3rd Edition, SAS Institute, Cary.

[15] Lin, E.B., Abayomi, O., Dahal, K., Davis, P. and Mdziniso, N.C. (2016) Artifact Removal for Physiological Signals via Wavelets. 8*th International Conference on Digital Image Processing* (*ICDIP* 2016), Chengdu, 29 August 2016, Article ID: 1003355.

https://doi.org/10.1117/12.2244906

[16] Dahal, K.R., Dahal, J.N., Goward, K.R. and Abayami, O. (2020) Analysis of the Resolution of Crime Using Predictive Modeling. *Open Journal of Statistics*, **10**, 600-610. https://doi.org/10.4236/ojs.2020.103036

[17] Dahal, K.R. and Amezziane, M. (2020) Exact Distribution of Difference of Two Sample Proportions and Its Inferences. *Open Journal of Statistics*, **10**, 363-374. https://doi.org/10.4236/ojs.2020.103024

[18] Manning, C. (2007) Logistic Regression (With R) Changes.

[19] Hansen, M., Dubayah, R. and Defries, R. (1996) Classification Trees: An Alternative to Traditional Land Cover Classifiers. *International Journal of Remote Sensing*, **17**, 1075-1081. https://doi.org/10.1080/01431169608949069

[20] Gareth, J., Daniela, W., Trevor, H. and Robert, T. (2013) An Introduction to Statistical Learning: With Applications in R. Springer, Berlin.

[21] Liaw, A. and Wiener, M. (2002) Classification and Regression by Randomforest. *R News*, **2**, 18-22.

[22] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., *et al.* (2019) Misc Functions of the Department of Statistics (e1071). TU Wien, R Package Version 1.7-0.1.

[23] Gong, A. and Liu, Y. (2011) Improved KNN Classification Algorithm by Dynamic Obtaining K. In: Shen, G. and Huan, X., Eds., *International Conference on Electronic Commerce*, *Web Application*, *and Communication*, Springer, Berlin, Heidelberg, 320-324. https://doi.org/10.1007/978-3-642-20367-1_51

[24] Crookston, N.L. and Finley, A.O. (2008) Yaimpute: An R Package for kNN Imputation. *Journal of Statistical Software*, **23**, 1-16. https://doi.org/10.18637/jss.v023.i10

[25] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C. and Müller, M. (2011) pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics*, **12**, Article No. 77. https://doi.org/10.1186/1471-2105-12-77

[26] Coomans, D. and Massart, D.L. (1982) Alternative K-Nearest Neighbour Rules in Supervised Pattern Recognition: Part 1. K-Nearest Neighbour Classification by Using Alternative Voting Rules. *Analytica Chimica Acta*, **136**, 15-27. https://doi.org/10.1016/S0003-2670(01)95359-0

[27] Lee, Y.S., Oh, H.J. and Kim, M.K. (2005) An Empirical Comparison of Bagging, Boosting and Support Vector Machine Classifiers in Data Mining. *Korean Journal of Applied Statistics*, **18**, 343-354.