Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2020, 8, 41-68
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease

**Imran Chowdhury Dipto[1], Tanzila Islam[2],
H M Mostafizur Rahman[3], Md Ashiqur Rahman[1]**

[1]Department of Electronics, Computing and Mathematics, University of Derby, Derby, UK
[2]Department of Computer Science and Intelligent Systems, Iwate University, Morioka, Japan
[3]BAC International Study Centre, Dhaka, Bangladesh
Email: dipto.imranchowdhury@gmail.com, tanzilamohita@gmail.com, mostafizur@bacbd.org, ashiq.rahman6868@gmail.com

How to cite this paper: Dipto, I.C., Islam, T., Rahman, H.M.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

Received: February 2, 2020
Accepted: April 6, 2020
Published: April 9, 2020

Abstract

Coronary Artery Disease (CAD) is the leading cause of mortality worldwide. It is a complex heart disease that is associated with numerous risk factors and a variety of Symptoms. During the past decade, Coronary Artery Disease (CAD) has undergone a remarkable evolution. The purpose of this research is to build a prototype system using different Machine Learning Algorithms (models) and compare their performance to identify a suitable model. This paper explores three most commonly used Machine Learning Algorithms named as Logistic Regression, Support Vector Machine and Artificial Neural Network. To conduct this research, a clinical dataset has been used. To evaluate the performance, different evaluation methods have been used such as Confusion Matrix, Stratified K-fold Cross Validation, Accuracy, AUC and ROC. To validate the results, the accuracy and AUC scores have been validated using the K-Fold Cross-validation technique. The dataset contains class imbalance, so the SMOTE Algorithm has been used to balance the dataset and the performance analysis has been carried out on both sets of data. The results show that accuracy scores of all the models have been increased while training the balanced dataset. Overall, Artificial Neural Network has the highest accuracy whereas Logistic Regression has the least accurate among the trained Algorithms.

Keywords

Coronary Artery Disease, Machine Learning, Logistic Regression, Support Vector Machine, Artificial Neural Network

1. Introduction

Coronary Artery Disease is the number one cause of deaths World-Wide and of the 56.9 million deaths reported around the world in 2016, more than 54% were because of top 10 causes of death among which Ischaemic Heart Disease (Coronary Artery Disease) and Stroke were the biggest killers and they remained the top causes of death for the last 15 years globally [1].

To function properly the Heart requires the supply of blood and the Heart muscles receive blood from Coronary Arteries. Coronary Artery Disease is the blockage or narrowing of the Coronary Arteries caused by hardening or clogging of these arteries due to the build-up of cholesterol or fatty deposits called plaque in the arteries inner walls. The plaque could restrict the flow of blood by clogging the artery or by causing abnormal artery tone or function. Therefore without a proper supply of blood, the heart becomes starved of oxygen and vital nutrients resulting in Chest Pain. If blood supply is entirely cut-off to a portion of a Heart muscle or if the energy requirements of the heart become more than the supply of blood, the result is a heart attack clinic [2].

Machine Learning is known as the Technology that is used for the development of Computer Algorithms with capabilities of mimicking the intelligence level of a human being. It is produced from ideas different fields such as Artificial Intelligence, Statistics and Probability, Computer Science, Information Theory, Psychology, Control Theory and Philosophy [3] [4] [5].

In this study, three different Supervised Machine Learning Algorithms are implemented to predict the presence of patients with CAD and finally, the performances of the Algorithms are compared to select the ideal Algorithm. The dataset used is the Z-Alizadeh Sani dataset that provides clinical records of 303 patients with a total of 54 features related to the disease.

The rest of this paper is organised as follows: in Section 2, a brief survey of existing literature related to the research topic is provided. The research methodology is described in Section 3 and the analysis of the results obtained after implementing the three Algorithms on the dataset is presented in Section 1. Then the findings from the research have been discussed in Section 5 and finally, the study is concluded with a number of future recommendations that would improve the existing research in the future in Section 6.

2. Literature Review

In 2008, Kurt [6] compared five different Algorithms for the prediction of Coronary Artery Disease which were Logistic Regression, Classification and Regression Tree, Artificial Neural Network, Radial Bias Function and Self Organising Feature Maps. The Algorithms were tested on a data set containing 1245 patients records and with the use of various predictor variables such as age, sex, body mass index and so on. The test results showed that the Neural Network was the best performer compared to the other Algorithms.

In 2009, Lavesson used Algorithms such as Bagging, AdaBoost and Naive

Bayes on the CHAPS data set. The test was done for the prediction of the Acute Coronary Syndrome from the research it was found that Naive Bayes had the highest accuracy [7]. In another study in 2010, Babaoğlu used a data set containing 480 patient data with 23 features and applied Support Vector Machine Algorithm to detect the presence of Coronary Artery Disease. In the study, a subset of features was selected using an Algorithm called Principal Component Analysis which reduced the dimensionality of the data set. The test results show that the researchers had finally achieved an accuracy score of 81.46%. It is seen from the research that the application of the Principal Component Analysis reduces the training error and time taken to testing and training of the Support Vector Machine Algorithm [8].

In 2013, Alizadehsani used different Classification Algorithms on the Z-Alzadeh Sani data which consists of random 303 patients who visited the Shaheed Rajaei Cardiovascular, Medical and Research Center in Tehran, Iran. The data set contains 216 patients who had Coronary Artery Disease (CAD) and the rest of the patients are free from the disease and a total of 54 features. With the Sequential Minimal Optimisation (SMO), Naive Bayes, Bagging with SMO and Neural Network, the researcher also introduced a feature selection Algorithm for the creation of three different features. From the test results, it is found that the accuracy of SMO produced the highest value of 94.08% when tested with the three created features [9].

A hybrid model was proposed for identification and prediction of Coronary Heart Disease (CHD) by Akila in 2015 and the model was tested on patient data who were occupational drivers and the data were collected from a medical college and hospital. The model consisted of two stages in the first stage, risk identification was carried out by classification of physical and biomedical factors using Decision Tree (DT) Algorithm. In the second stage, CHD risk identified instances using Decision Tree were analysed using Multi-Layer Perceptron (MLP) using habitation and medical history attributes. The Classification accuracies of DT and MLP were 98.66% and 96.66% [10].

In 2016, Lo collected four Heart Disease Data sets from the University of California Irvine (UCI) Machine Learning Repository, combined the data sets, removed the missing values and prepared a combined data set containing data of 822 patients of which 453 had Coronary Artery Disease (CAD) and the rest did not. The presence of disease was identified using various risk factors of CAD in the Asian population. The authors used seven Machine Learning Algorithms such as Naive Bayes, Artificial Neural Network, Sequential Minimal Optimization, K-Nearest Neighbour, Adaboost, J48 and Random Forest. The seven methods were compared against an Ensemble Method called Voting Algorithm was also used. From the study, it was found that the Voting method predicted the presence of CAD in patients with the highest accuracy [11].

Interesting research had been carried out in 2016 by Alizadehsani where the stenosis (narrowing) of the major Coronary Arteries of patients from the Z-Alizadeh Sani data set. Two feature selection methods were used to extract the best features and variables for consideration of stenosis of the major arteries were studied from a medical book. To predict the patients with stenosis, the Support Vector Machine Algorithm was used with various kernel methods and promising results were obtained. In addition, the Apriori Algorithm was used to decide on whether the arteries were stenotic [12].

In 2017, Forssen systematically implemented and evaluated two Supervised Learning Algorithms used were Logistic Regression, Penalized Logistic Regression and Random Forest and compared them to traditional regression approaches for Coronary Artery Disease prediction. The data was collected from the Clinical Cohorts in Coronary disease Collaboration (4C) study containing 3409 number of recruited patients with acute or stable chest pain from four UK National Health Service (NHS) hospitals. In order to reduce the dimensionality of the data Principal Component Analysis (PCA) was used. After running PCA six principal components were selected which had more than 95% variance of the data. After running the Supervised Algorithms in the adjusted and unadjusted forms it was found that applying PCA and adjusting the Algorithms it is seen that Penalized Logistic Regression had the highest accuracy when it was run in both adjusted and unadjusted ways [13].

A hybrid approach was proposed by Arabasadi in 2017 where the researchers tried to enhance the performance of Neural Networks by using the Genetic Algorithm. The tests were carried out on the Z-Alizadeh Sani Dataset. The results showed that the Neural Network with the use of a Genetic Algorithm produced an accuracy of 93.85% [14].

In 2018, Meng built a hybrid Algorithm called two-layer Gradient Boosting Decision Tree which was compared with two other commonly used Machine Learning Algorithms such as Support Vector Machine and Logistic Regression. The Algorithms were run on a created data set which consisted of 15,000 patient data of routine blood test results. With the use of the created data set, the researchers trained the Algorithms to classify healthy status, coronary heart disease and other diseases. The test results show that the prediction accuracy of the created Algorithm for prediction of the presence of Coronary Heart Disease and other diseases was higher than the other Algorithms trained on the data set [15].

Another study conducted by Nassif in 2018, the researchers tested Support Vector Machine, Naive Bayes and K-Nearest Neighbour Algorithms using 10-fold cross-validation on the Cleveland Heart Disease data set to compare the performance of three Machine Learning Algorithms for the prediction of Coronary Artery Disease. Three different feature selection methods were applied on the data set and feature versus risk score graphs were plotted to identify the features that are closely related to the risk of Coronary Artery Disease and seven best features were selected for input variables for the Algorithms. From the research, it is found that the Naive Bayes Algorithm was the best performer which showed an accuracy score of 84% [16].

In 2019, Shamsollahi used a data set containing clinical data of 282 patients with 58 attributes to compare the performance of various Machine Learning methods for the prediction of Coronary Artery Disease. The researchers had used both descriptive (Clustering) and predictive (Classification) methods on the data set. The K-Means Clustering Algorithm was used to cluster the data into three clusters of patients of their amount of smoking. Then to predictive (Classification) Algorithms such as C & RT, CHAID and so on were used on the Clusters. From the research, it is found that the C & RT Algorithm was the best performer as it predicted with the highest figures of accuracy for the three Clusters [17].

3. Research Methodology

3.1. Data Collection Method

The data set is collected from the UCI Machine Learning Repository which contains a collection of data sets that are widely used by the Machine Learning community. In the repository, the information of the donors and creators of the data set, data information, attributes of the data set and other relevant information are also provided [18].

3.2. Dataset Description

The Z-Alizadeh Sani Dataset will be used for the research which contains records of 303 random patients who visited Shaheed Rajaei Cardiovascular, Medical and Research Center of Tehran, Iran. The data set contains features that are related to the diagnosis of Coronary Artery Disease. 216 patients from the data set have the disease and the rest of the patients are normal. The features are grouped into four different categories. If a patient has stenosis (narrowing) of more than 50% in one of their coronary arteries then that patient is diagnosed with the disease [9] [12] [14].

Although the data set contains one additional feature "BBB" that stands for Blood-brain barrier, however, from the research papers it is found that the feature has been removed from the data set before running the Algorithms, therefore, it can be inferred that "BBB" is a feature that is not related to Coronary Artery Disease likewise the feature will also be dropped from the data set in the current research. The label or target variable is "Cath" with values "CAD" for the presence of the disease and "Normal" for a normal patient. The dataset was donated on UCI Machine Learning Repository at 2017-11-17 [18].

3.3. Design of the Experiment

There are a total of four steps which will be followed to carry out the research. The first step involves the Exploratory Data Analysis followed by data pre-procession. After processing the data set is divided into training and test sets on which the Algorithms will be implemented and finally the Algorithms will be evaluated in the final step. **Figure 1** provides an overview of the experiment design and the

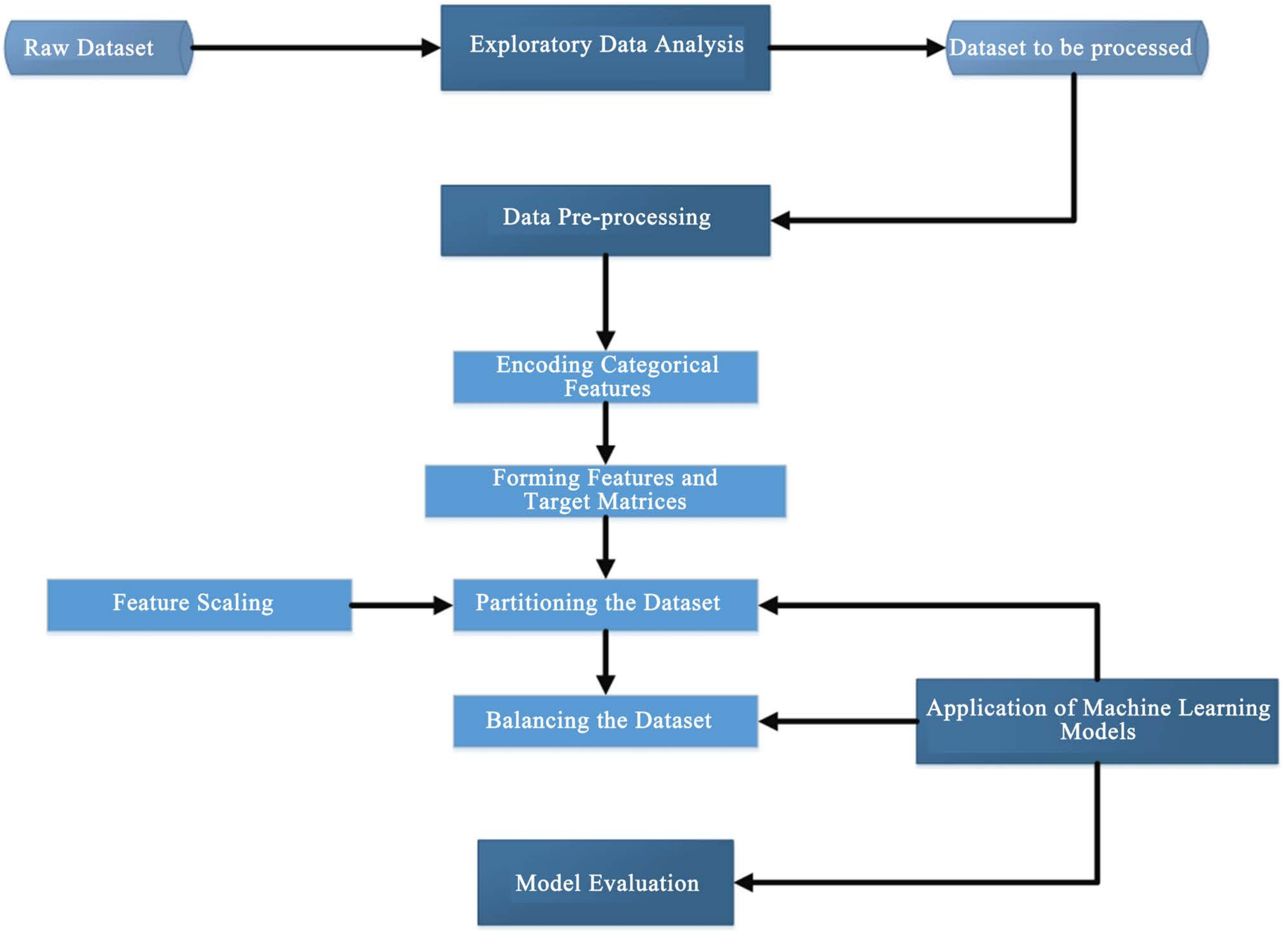


Figure 1. Design of the experiment.

following Sections describe the steps in further detail.

3.4. Exploratory Data Analysis

This step will be performed to gain useful insights into the collected data set through data visualisations and results from performing analysis. This step will help to find if the data set has any missing values, identify the Categorical features, numerical features and more.

3.5. Data Pre-Processing

Raw data are often not found in the form and shape that is required for the optimal performance of learning Algorithms. Therefore the preprocessing step is the most important in Machine Learning Applications [19]. First, the Categorical Features of the dataset are encoded to transform these features into numerical values. Then the matrices of features and the predictor variable are created followed by dividing the dataset into training and test sets. In the following step, Feature Scaling is applied to the input features. Finally, the dataset is balanced using the SMOTE Algorithm. The following Subsections describe these sub-steps involved in the Pre-processing stage for the dataset used in this research.

3.6. Encoding Categorical Features

One important aspect of Machine Learning is feature engineering. The Algorithms that will be implemented are only able to read numerical values so it is important to transform the categorical features into numerical values cat [20].

3.7. Forming Features and Target Matrices

The matrix of features to be used as input variables and the target variable will be taken. The input variable features and the target matrices will be taken into X and y variables.

3.8. Partitioning the Dataset

In this part of the Pre-processing stage, the matrix of features X and the target variable y is split using the "test_train_split" method of "model_selection" class of scikit-learn. 80% data will be used for training and the remainder will be used to Test the Machine Learning Models that will be implemented in this research. The training and test set of X will be stored into variables called "X_train" and "X_test". Likewise, the y training and test set will be stored in variables named "y_train" and "y_test".

3.9. Feature Scaling

Most data sets contain features that are of varying ranges and this is a problem since most of the Machine Learning Algorithms use Euclidean distance between two data points. If features are not scaled, such algorithms will only take in the magnitude of features and the produce various results as features with higher ranges will weigh in more in the distance calculations than features in lower ranges. Hence feature scaling is applied to suppress the explained effect to bring all the features into the same magnitude [21].

To scale the features of the data set Standardization will be used. The results of the Standardization also known as Z-score Normalization are that the features will have properties of a standard normal distribution with $\mu = 0$ and $\sigma = 1$ where μ is mean and σ is the standard deviation. The formula used to calculate the Standardization (Z-Scores) is as follows:

$$z_i = \frac{x_i - \overline{x}}{\sigma} \tag{1}$$

where x_i is the value of each feature, $\overline{x}$ is the mean of the features in a column and σ is the standard deviation of values in that column. The implementation will be performed via the use of the "StandardScaler" method from the "preprocessing" class of scikit-learn.

3.10. Balancing the Dataset

A data set balancing Algorithm called Synthetic Minority Oversampling Technique (SMOTE) will be used. SMOTE developed by [22] is an over-sampling technique used to generate synthetic minority samples. The technique combines informed oversampling of the minority class with random under-sampling of the majority class, as a result, the minority class is over-sampled with the creation of artificial sample classes of k-nearest neighbours as shown in **Figure 2**.

SMOTE balances a data set by over-sampling the minority class (by creating artificial instances of the minority class) so that it equals to the number of the majority class. The Algorithm is given as: for each minority sample:

- Find its k-nearest minority neighbours.
- Randomly select q of those neighbours.
- Randomly generate synthetic samples along the lines joining the minority sample and its q selected neighbours (q depends on the amount of oversampling desired) [23].

3.11. Machine Learning Algorithms

There are various Supervised Machine Learning Algorithms such as K-Nearest Neighbours, Decision Tree, Naive Bayes, Support Vector Machine and many more, but throughout the medical literature, it is seen that Support Vector Machine and Neural Network Algorithms are most commonly used [24]. In this research, the along with the two Algorithms mentioned the Logistic Regression will also be implemented as from Figure? It can be seen that Logistic Regression is the third most commonly used Machine Learning Algorithm in Healthcare. Hence implementing three Algorithms will make the research more relevant.

Logistic Regression

Logistic Regression (LR) Model is used for predicting binary outcomes. In predicting the LR equation the maximum-likelihood ratio to determine the statistical significance of the variables [25]. LR is ideal for problems where the task is to predict the presence or absence of a characteristic or outcomes that are based on values of predictor variables. LR model is similar to a Linear Regression model however, it is suitable for models where the outcome is binary [6]. LR is based on the Logistic Function $P(y)$ defined as:

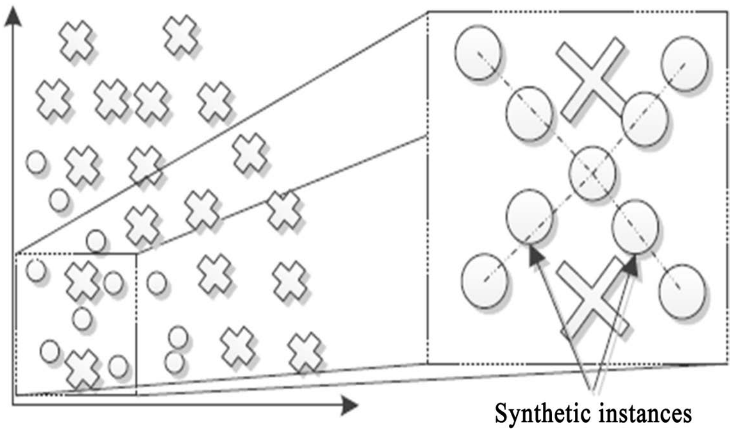


Figure 2. Synthetic Instances created by SMOTE [23].

$$P(y)=\frac{e^{y}}{1+e^{y}}=\frac{1}{1+e^{-y}} \quad (2)$$

LR model for P independent variables can be written as:

$$P(y=0)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_PX_P)}} \quad (3)$$

Here $P(y=0)$ is the presence of Coronary Artery Disease and $\beta_0,\beta_1,\cdots,\beta_P$ are regression coefficients. There is a linear model hidden within the LR model and the mathematical Logarithm of the ratio of $P(y=0)$ to $1-p(y=0)$ gives a linear model in *X*:

$$g(X)=\ln\left(\frac{P(y=0)}{1-P(y=0)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_PX_P \quad (4)$$

where $g(X)$ has some properties of a Linear Regression model and the independent variables: "*X*" could be a combination of both continuous and categorical variables [25].

3.12. Support Vector Machine

It was invented by Vapnik in 1979 and proposed for solving Classification and Regression problems by Vapnik in 1995. Support Vector Machine or SVM for short is a Supervised Learning Algorithm that uses a non-linear mapping to transform the original training data into higher dimensional space, within this new dimension it searches for the linear optimal separating hyperplane (decision boundary) that separates the tuples of one class from another. Data can always be separated by a hyperplane with an appropriate nonlinear mapping to a sufficiently high dimension. The algorithm finds the hyperplane using support vectors ("essential" training tuples) and margins (defined by the Support Vectors) [26].

Mathematics

From the research paper of [26] the mathematics is explained as; let the data set D be given as $(X_1,y_1),(X_2,y_2),\cdots,(X_D,y_D)$, where X_i is the set of training tuples with associated class labels, y_i. Each y_i can take one of the two values, either +1 or −1 (*i.e.*, $y_i \in +1,-1$). In an SVM, the idea is to find the hyperplane that maximises the minimum distance from any training data point as shown in **Figure 3**. It is expected that the hyperplane with a larger margin to be more accurate at classifying future data tuples than hyperplane with the smaller margin. Hence, SVM searches for the hyperplane with the largest margin. A separating hyperplane can be written as, $W\cdot X+b=0$ where W is a weight vector and b is a bias. Thus any point that lies above the separating hyperplane satisfies $W\cdot X+b>0$ Similarly, any point that lies above the separating hyperplane satisfies $W\cdot X+b<0$. The weights can be adjusted so that the hyperplanes defining the sides of margin can be written as $H_1: W\cdot X+b\geq 1$ for $y_i=+1$, $H_2: W\cdot X+b\geq 1$ for $y_i=-1$. So, any tuple that falls on or above H_1 belongs

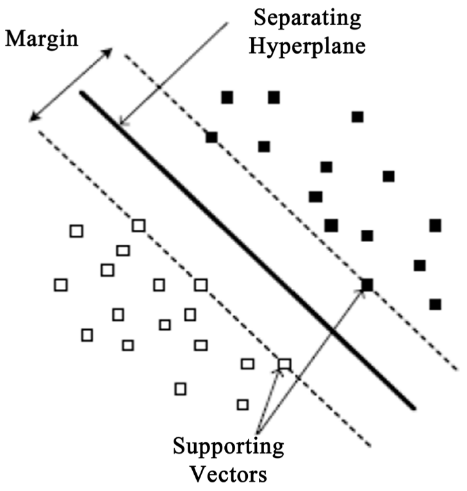


Figure 3. Maximum margin and optimal hyperplane [26].

to class +1, and any tuple that falls on or below H_2 belongs to class −1. Combining the two inequalities of equations gives $y_i(W \cdot X + b) \geq 1$, for all *i*. The above problem can be solved by introducing the Lagrange multipliers ($\alpha_i \geq 0$ ($i = 1, 2, \cdots, m$)) The patterns x_i which correspond to non-zero Lagrange coefficients are called support vectors. The resultant decision function has the following form:

$$y(x) = \operatorname{sgn}\left[\sum_{i=1}^{m} \alpha_i y_i (x_i, x) + b\right] \tag{5}$$

However, Equation (5) is applicable to data samples that are linearly separable. In such cases where data is linearly inseparable a kernel function is used to transform the data into a higher-dimensional space without actually transforming them into that space. This notion is known as the kernel trick which allows the transformation of data to large dimensions for Classification problems [26]. In situations where the data samples are not linearly separable the resultant function is given as follows:

$$y(x) = \operatorname{sgn}\left[\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right] \tag{6}$$

where $K(x_i, x)$ is the kernel function equals to $(\phi(x_i), \phi(x))$ and $\phi(x)$ is the non-linear space from the original space to high dimensional space [26]. The four basic kernels are given as follows where γ, r and d are kernel parameters [27]:

- *Linear*: $K(x_i, x_j) = x_i^{\mathrm{T}} x_j$.
- *Polynomial*: $K(x_i, x_j) = (\gamma x_i^{\mathrm{T}} x_j + r)^d \; \gamma > 0$.
- *Radial Bias Function* (*RBF*): $K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0$.
- *Sigmoid*: $K(x_i, x_j) = \tanh(\gamma x_i^{\mathrm{T}} x_j + r)$.

3.13. Artificial Neural Network

Inspired from capabilities of the human brain for its incredible processing capabilities due to interconnected neurons. Artificial Neural Networks (ANN) are designed by processing units called Perceptrons. They contain one layer and are able to solve linearly separable problems and to solve non-linear problems Multilayer Perceptrons are used which contains an input layer, one or more hidden layers and an output layer [28].

Multi-layer Perceptron is a supervised learning algorithm that learns a function $f(.): R^m \rightarrow R^o$ where m is the number of dimensions for input and o is the number of dimensions for output. MLP can learn a non-linear approximator for either classification or regression given a set of features $X = x_1, x_2, \cdots, x_m$ and a target y. **Figure 4** represents the diagram of an MLP where the layer on the left is the input layer with a set of neurons representing the input features $(x_1, x_2, \cdots, x_n)$, layer in the middle is the hidden layer and the neurons of this layer transforms the values from the input layer with a weighted linear summation $w_1, w_1 + w_2, w_2 + \cdots + w_m, x_m$ followed by a non-linear activation function. The final layer is the output layer that receives from the hidden layer and transforms them into output values [29].

3.13.1. Steps in Backpropagation Algorithm

The Backpropagation Algorithm (BA) is the most commonly used learning techniques in Artificial Neural Network, following are the steps as described by [28]:

- All network weights are initialised to small random numbers.
- Training data is received as input and output is computed for each unit with the equation below known as Sigmoid Function where $\overline{w}$ is the vector of weight values and $\overline{X}$ is the vector of input values in the network:

$$o = \alpha(\overline{w}, \overline{X}) = \alpha(y) = \frac{1}{1 + e^{-y}} \quad (7)$$

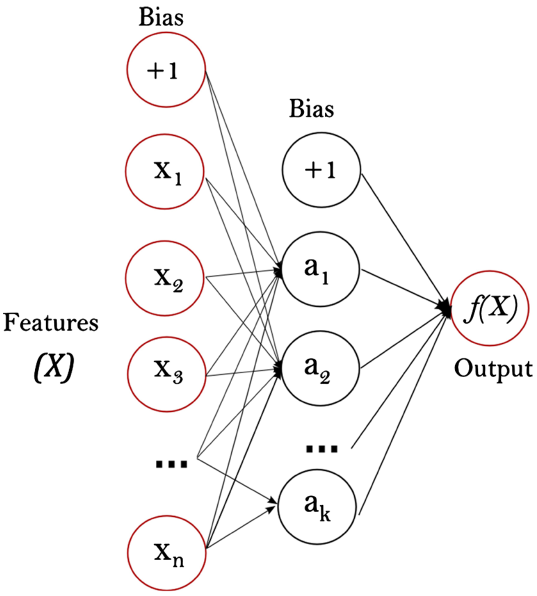


Figure 4. MLP with one hidden layer [29].

- Then error computation step is started. BP algorithm works as follows: Error signal (δ) which is calculated for each network output is propagated to all neurons in the network as input.
- Error term δ_k calculated for each network output unit using the following equation where o_k network output for output unit *k* and indicates desired output for output unit *k*:

$$\delta_k \leftarrow o_k\left(1-o_k\right)\left(t_k-o_k\right) \tag{8}$$

- Error term δ_k calculated for each hidden unit *h* as below where w_{kh} denotes network weight from hidden unit *h* to output unit *k*:

$$\delta_h \leftarrow o_h\left(1-o_h\right)\sum_{k\in outputs} w_{kh}\delta_k \tag{9}$$

- Each network weight is updated as follows where $\Delta w_{ji}=\eta\delta_j x_{ji}$ and η is the learning rate and x_{ji} denotes the input from unit *i* into unit *j* [3]:

$$w_{ji} \leftarrow w_{ji}+\Delta w_{ji}\,. \tag{10}$$

3.13.2. Designed MLP for the Research

The designed MLP consists of an Input Layer, Hidden Layer and an Output Layer as shown in Figure 5 where *X* denotes the input features, *n* denotes the final feature and *A* is the activation function. The number of neurons will be the same as the input features such as “Age”, “Weight” and so on in the input layer

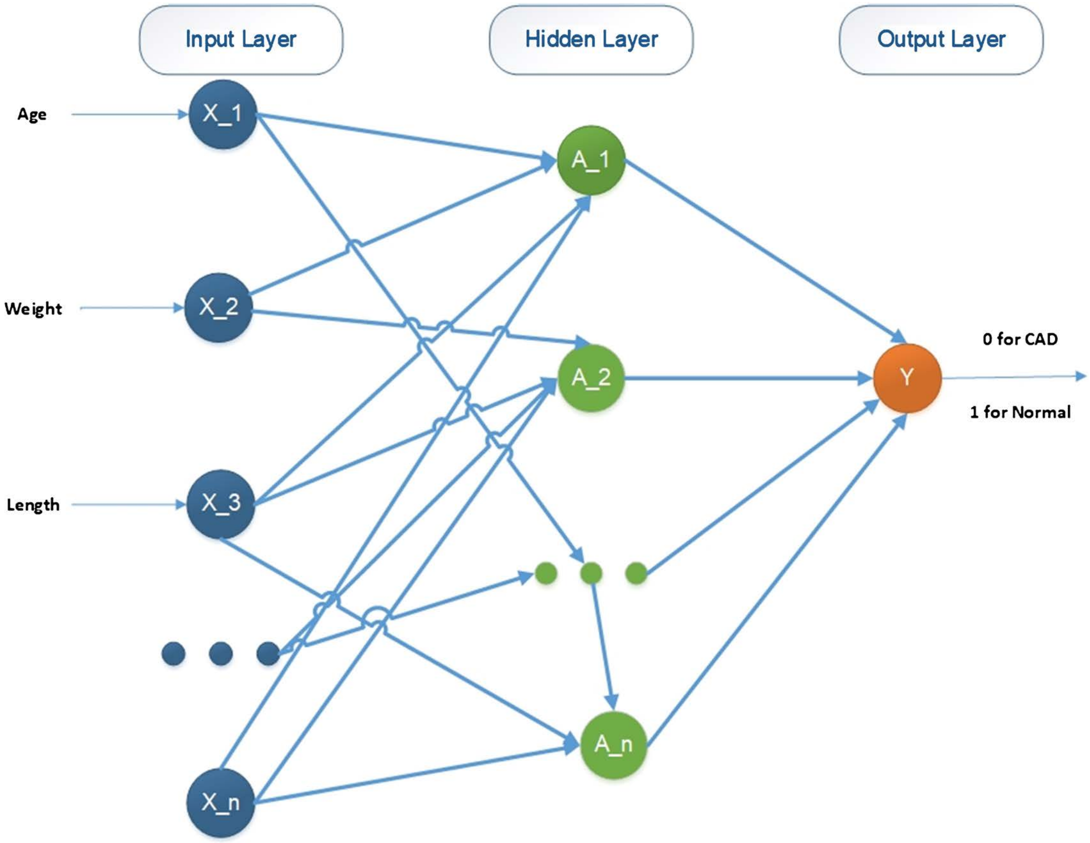


Figure 5. Design of multi-layer perceptron for the research.

where $X\_1 \dots X\_n$ represents the number of input features. The number of neurons in the hidden layer is represented as $A\_1 \dots A\_n$, as the Network will solve a binary classification problem hence, the output layer will consist of one neuron shown as Y. The hidden layer neurons will be activated by the Rectified Linear Unit function and the neuron on the output layer will be activated by the Logistic Function as shown by Equation (2).

3.14. Performance Evaluation Methods

3.14.1. Confusion Matrix

It is an evaluation metric which is used to describe the performance of a classifier by calculating evaluation parameters and is shown in Table 1 Where TP = True Positive *i.e.* positive instances that are classified as actual and correct prediction of CAD, FP = False Positive *i.e.* negative instances that are classified as positive, FN = False Negative *i.e.* positive instances that are classified as negative, TN = True Negative *i.e.* negative instances that are classified as negative. The values of the true positive rate and false positive rate are generated using a confusion matrix.

3.14.2. Stratified K-Fold Cross Validation

The designed experiment uses two steps to evaluate the implemented Algorithms. Firstly, a stratified K-fold Cross-validation technique will be used for validating the implemented models. In this validation technique, the folds are selected so that each class labels are distributed equally in each fold. The target variable is binary hence the experiment comes under dichotomous classification, this means that each fold contains roughly the same proportions of the two types of class labels. The data set will be divided into k subsets where $k = 10$, each time one of the k subsets will be used as the test set and the $k - 1$ subsets will be used as a training set. Therefore every data point will be part of the test set exactly once and gets to be in training set $k - 1$ times. The average results from the k folds will be taken and a single estimation will be produced. $k = 10$ is taken because 10 is the standard value which is ideally used in research [30]. Figure 6 shows the illustration of the technique when $K = 5$.

3.14.3. Accuracy

The performance of the Algorithms will be compare based on the accuracy obtained on the prediction of CAD given by Equation (11). Accuracy of the models will be obtained in each fold at the end of training with this technique there would be 10 accuracies per model. The average of the accuracies will be obtained along with the standard deviation of the accuracies will also be calculated to

Table 1. Confusion matrix.

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

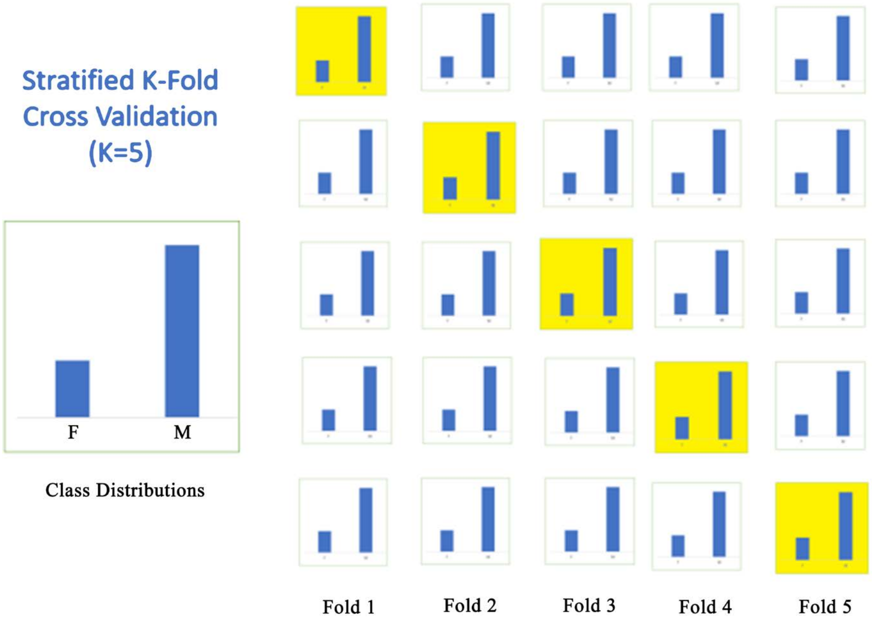


Figure 6. K-fold cross validation with $k = 5$ [31].

understand the variance.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \tag{11}$$

3.14.4. ROC and AUC

Receiver operating Characteristic curve (ROC) and area under the curve (AUC) will be obtained. ROC-AUC plot of each model will be generated to visualise the mean accuracy of each model. ROC curve is based on two metrics, True Positive Rate (TPR) and False Positive Rate (FPR). True positive rate (TPR), also known as sensitivity, hit rate or recall, is given as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

Intuitively this metric corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. The higher the TPR, the fewer positive data points will be missed. False-positive rate (FPR) or fall-out is given as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{13}$$

FPR can also be generated from specificity as:

$$\text{FPR} = 1 - \text{Specificity} \tag{14}$$

where specificity is defined as:

$$\text{FPR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{15}$$

This metric corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points. In other words, the higher the FPR, the more negative data points will be miss classified. To combine FPR and TPR into one single metric *i.e.* to generate AUC, two former metrics with a different threshold are calculated and then plotted on a single graph with FPR values on *x*-axis and TPR values on the *y*-axis. The resulting curve is called AUROC as shown in Figure 7.

4. Results Analysis

4.1. Results of Exploratory Data Analysis

Statistical Analysis

It was found from the literature that in previous experiments the feature "BBB" was not used so this feature was removed from the dataset. Figure 8 shows the results of running the code for obtaining the descriptive statistics of the modified dataset.

It is seen that the dataset contains a total of 303 rows and 55 columns with column names of "Age" to "Cath". The data-types include 5 floats, 29 integers and 21 objects. Hence the total number of Numeric and Categorical Features are 34 and 21 also the dataset does not contain any missing values.

4.2. Results of Data Pre-Processing

At the first stage, the raw dataset is processed by One-Hot encoding all the categorical variables into dummy variables and to avoid the dummy variable trap the

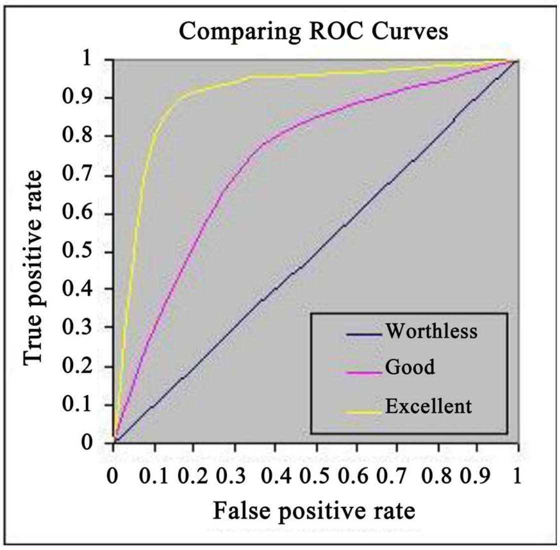
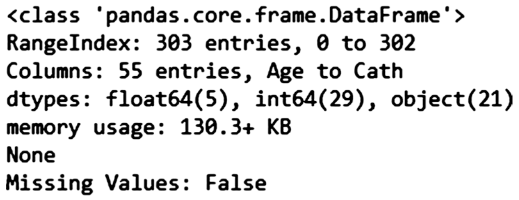


Figure 7. ROC-AUC curve explanation [30].



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Columns: 55 entries, Age to Cath
dtypes: float64(5), int64(29), object(21)
memory usage: 130.3+ KB
None
Missing Values: False
```

Figure 8. Descriptive statistics.

last columns were dropped. After encoding features, it was found that the "Exertional CP" had unique value in the feature columns, so it was dropped from the dataset. Dropping this feature will be beneficial for the implemented Machine Learning models as it only contains only one value so the Algorithms will not learn anything from this feature as shown in Figure 9.

Then further steps of taking matrices of features, splitting the matrices and feature scaling are applied as mentioned in the Research Methodology chapter. The implemented Algorithms will be implemented on the scaled features and also the X features will be re-sampled and the Algorithms will also be tested on the resampled features.

4.3. Results of SMOTE

From the Exploratory Data Analysis section, it was found that 28.7% of the patients are Normal and 71.3% of the patients were diagnosed with Coronary Artery Disease. This shows that there is an unequal distribution of labelled classes in the dataset. In order to fix the issue, the SMOTE Algorithm is implemented. The results of implementing SMOTE are shown in the following two figures. Figure 10 shows the distribution of the target classes before applying SMOTE and the result of applying SMOTE is shown in Figure 11 from which it is seen that the target class count of both the classes is 173.

4.4. Results of Model Evaluation

4.4.1. Logistic Regression: Imbalanced Target Data

Logistic Regression is implemented first on the processed data with imbalanced classes. From Figure 12 the Confusion Matrix of the model shows that it has returned with TP = 40 and TN = 12 and an accuracy score of 85.25%. The average accuracy score of K-Fold cross-validation obtained is 81.83% with a Standard

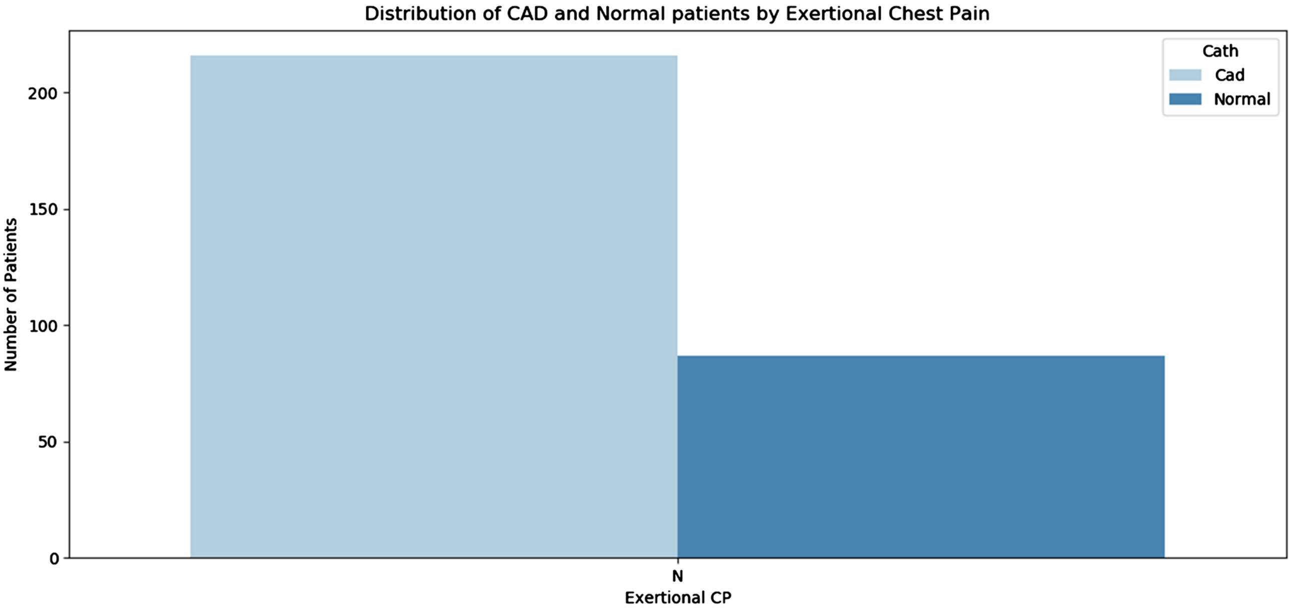


Figure 9. Distribution of patients with exertional CP.

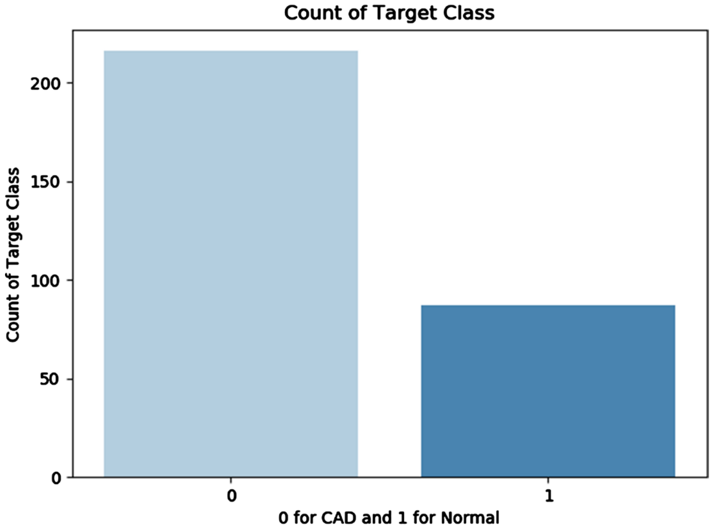


Figure 10. Count of target classes before SMOTE.

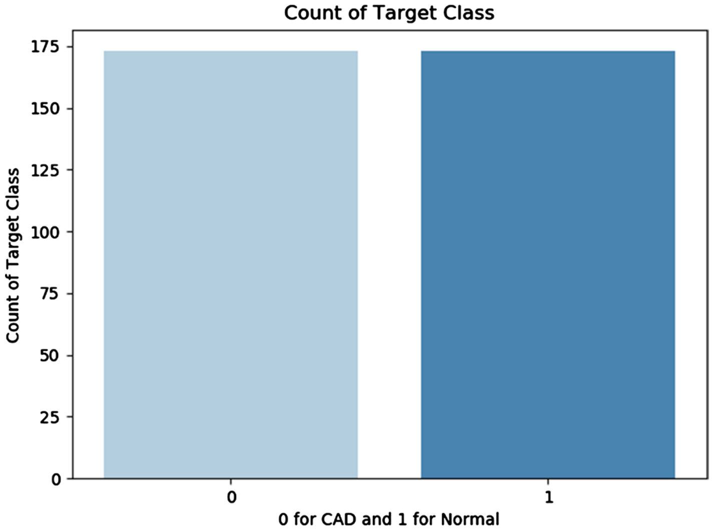


Figure 11. Count of target classes after SMOTE.

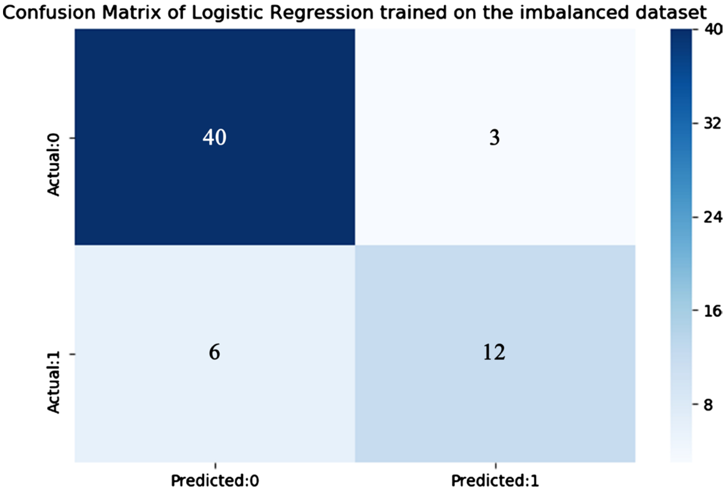


Figure 12. Confusion matrix of logistic regression trained on imbalanced dataset.

Deviation of ±5.28%. The ROC-AUC curve shows that the mean AUC obtained is 0.88 with minimum and maximum figures of 0.78 and 1 as shown in Figure 13.

4.4.2. Logistic Regression: Balanced Target Data

However, the value of TP, FN decreases while the values of TF and FP increase by 1 and accuracy stayed the same when Logistic Regression is run on the balanced dataset as shown in Figure 14.

From Figure 15 the average accuracy score increased to 89.61% and the standard deviation also decreases when the Algorithm is run on the balanced dataset. The mean AUC score achieved on the balanced dataset is 0.94.

4.4.3. Support Vector Machine: Imbalanced Target Data

The Confusion Matrix of Support Vector Machine in Figure 16 shows that the accuracy is about 87%. TP and TN were 41 and 12. Values of FN and FP are 2

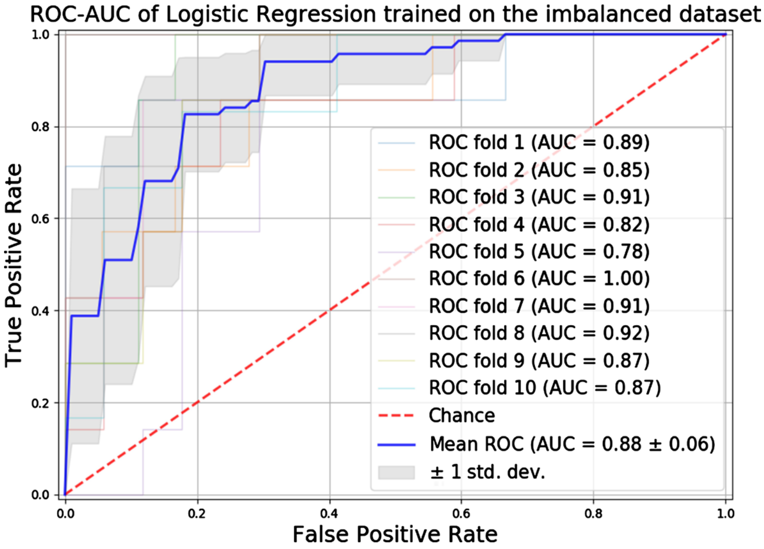


Figure 13. ROC-AUC of logistic regression trained on imbalanced dataset.

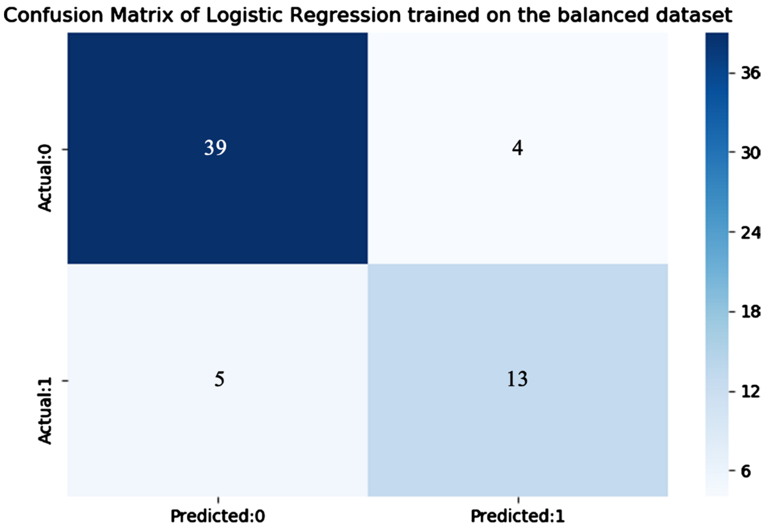


Figure 14. Confusion matrix of logistic regression trained on balanced dataset.

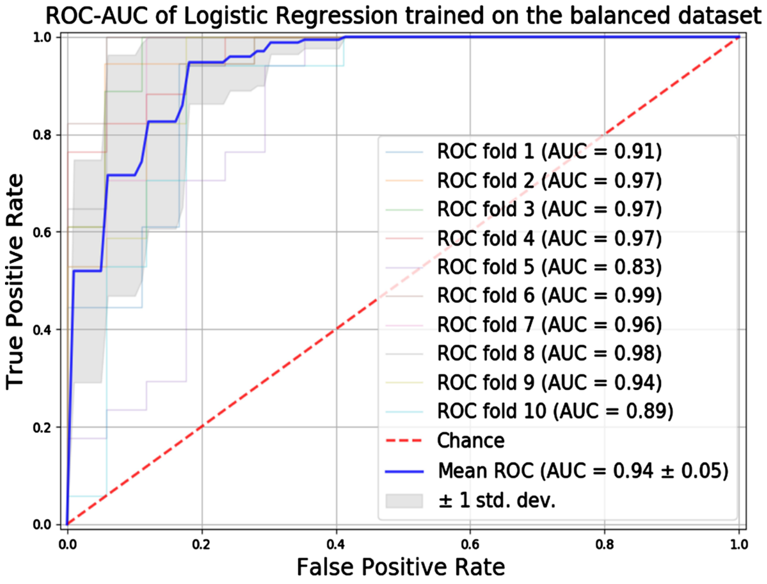


Figure 15. ROC-AUC of logistic regression trained on balanced dataset.

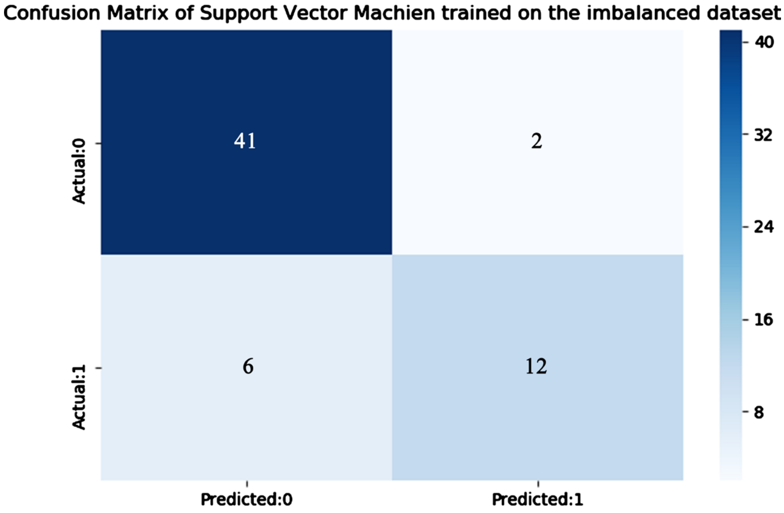


Figure 16. Confusion matrix of support vector machine trained on imbalanced dataset.

and 6. The average cross-validation accuracy obtained is the same as the Logistic Regression. Compared to the Logistic Regression the average AUC obtained is 0.91 with a standard deviation of ±0.05. The maximum figure is 1 whereas the lowest figure is 0.82 as shown in Figure 17.

4.4.4. Support Vector Machine: Balanced Target Data

Significant improvements in performance are observed when the Support Vector Machine is trained on the balanced dataset. Although the value of TN and FP stayed the same although TN improved from 12 to 14 and FN reduced from 6 to 4 as shown in Figure 18.

The improvement in performance is further displayed by the ROC-AUC curve showing a mean AUC of 0.98 with a standard deviation of 0.02 as illustrated in Figure 19.

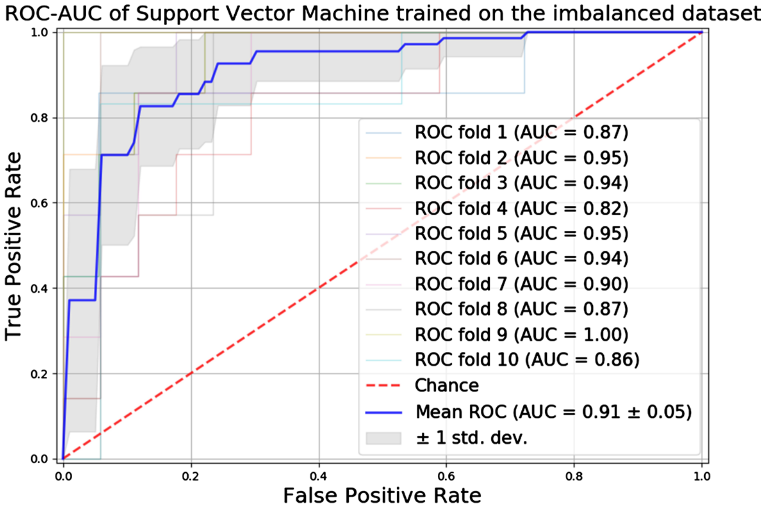


Figure 17. ROC-AUC of support vector machine trained on imbalanced dataset.

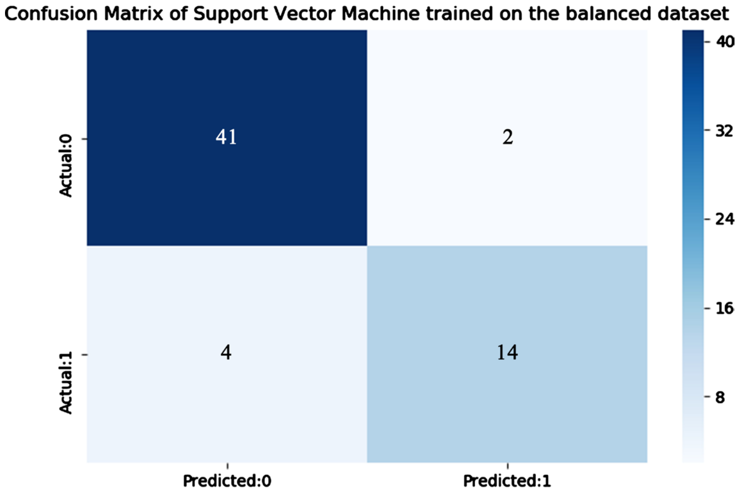


Figure 18. Confusion matrix of support vector machine trained on imbalanced dataset.

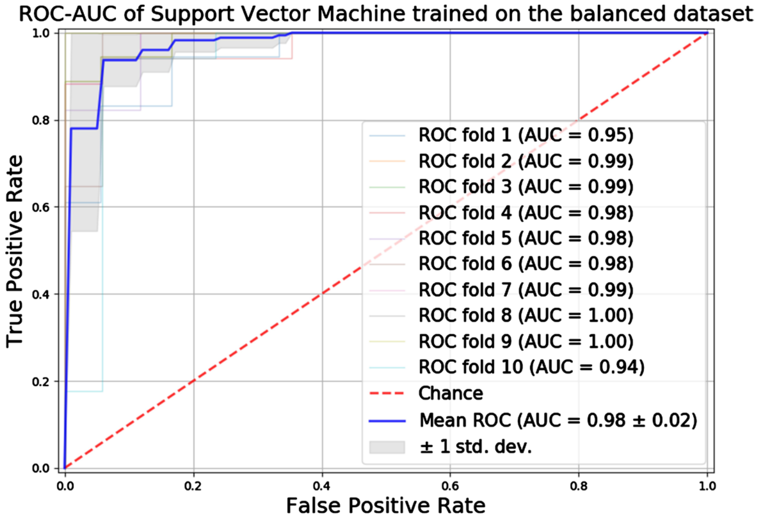


Figure 19. ROC-AUC of support vector machine trained on the balanced dataset.

4.4.5. Artificial Neural Network: Imbalanced Target Data

Like Support Vector Machine the Artificial Neural Network trained on the imbalanced dataset produced a TP of 41, however, the FP value is the same FN is 7 and TN is 11 (Figure 20). The cross-validation score obtained is 83.08% which is higher than both Logistic Regression and Support Vector Machine. Although The mean AUC obtained from the ROC-AUC curve is 0.90 with a standard deviation of +0.05. The value of mean AUC is 0.01 less than Support Vector Machine trained on the imbalanced target classes as shown in Figure 21.

4.4.6. Artificial Neural Network: Balanced Target Data

The highest accuracy score from the Confusion Matrix is found when the Artificial Neural Network is trained on the balanced target classes with the model being 91.80%. From Figure 22 the Confusion Matrix shows the highest TP of 42, lowest FP of 1 among the trained Algorithms on the balanced dataset despite FN

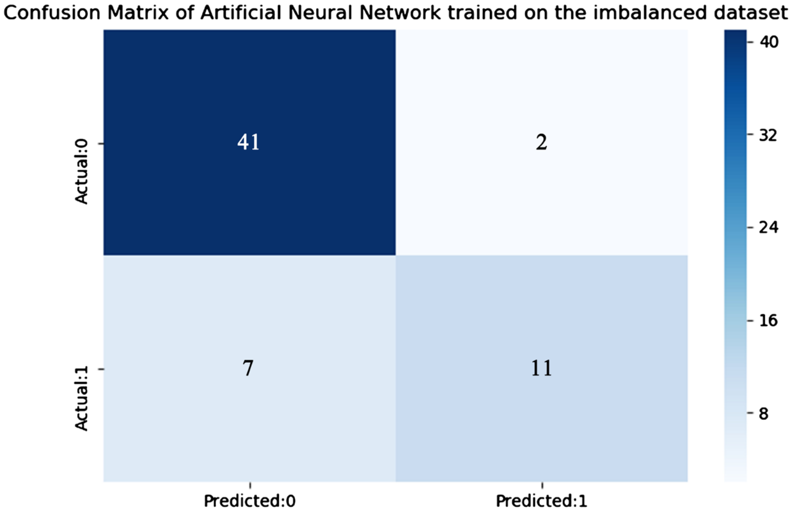


Figure 20. Confusion matrix of artificial neural network trained on the imbalanced dataset.

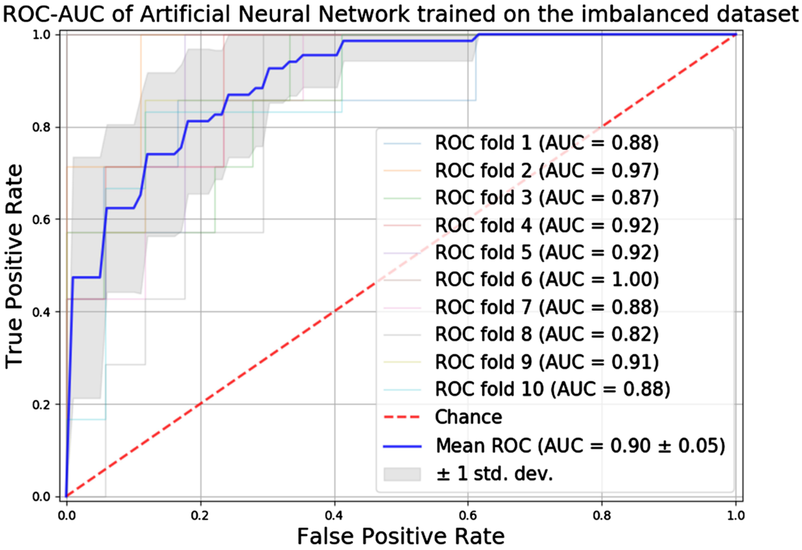


Figure 21. ROC-AUC of artificial neural network trained on the imbalanced dataset.

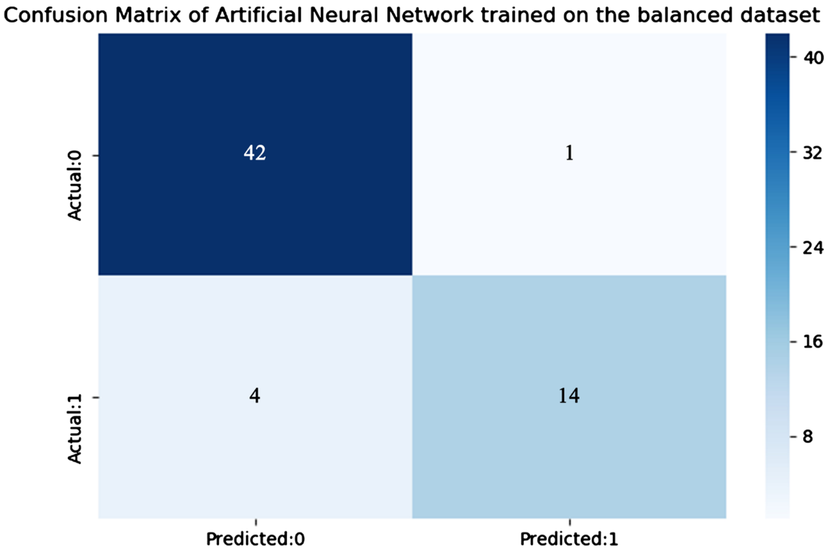


Figure 22. Confusion Matrix of Artificial Neural Network trained on the balanced dataset.

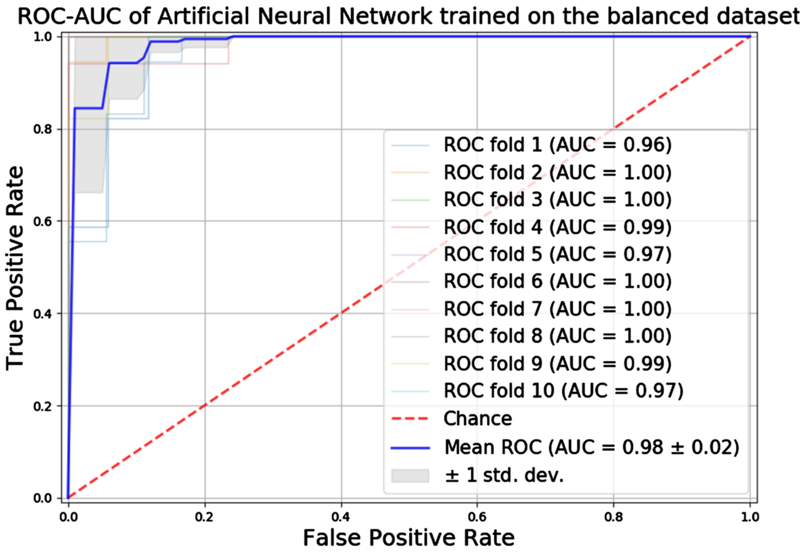


Figure 23. ROC-AUC of Artificial Neural Network trained on the balanced dataset.

and TN are the same as Support Vector Machine trained on the balanced dataset.

The highest cross-validation score is achieved with a low standard deviation and the mean AUC obtained is the same as Support Vector Machine's AUC with the same mean value of the standard deviation of the mean AUC value. The average cross-validation score obtained is 93.35 and the standard deviation is +2.56%. The results show that the model was quite stable with AUC of 1 appearing five times, 0.99 and 0.97 twice and the lowest score being 0.96 as shown in Figure 23.

5. Discussion of Findings

5.1. Impacts of Training on Imbalanced and Balanced Datasets

The results obtained from training the models on the dataset with imbalanced class shows that Support Vector Machine performed better than the other two Algorithms as it had the highest accuracy. This is also evident from the results obtained from the ROC curves of all the Algorithms trained on the imbalanced dataset. However, from the ROC curves, it is seen that the Support Vector Machine performed better as shown in Table 2 and Table 3.

Drastic improvements in results are seen after applying SMOTE. The results from Table 4 and Table 5 shows that the performance of all the Algorithms improved significantly as Logistic Regression and Support Vector Machine saw increases in performance by 7.78% and 6.19% and the model's performance were much stable as depicted by their reduction in Standard Deviations. The most significant improvement in performance is observed in the performance of ANN with an increase of accuracy score from 83.08% to 93.35% which is an increase of 10.27%. Moreover, ANN was much stable during training on both forms of the dataset and it was much stable when trained with a balanced dataset as the value of Standard Deviation reduced. As expected with the improved results the classifiers AUCs improved with SVM and ANN having the same values.

Table 2. Average cross-validation accuracies of algorithms trained on the imbalanced dataset.

Algorithm	Average Accuracy and Standard Deviation
Logistic Regression	81.83% ± 5.28%
Support Vector Machine	85.18% ± 7.99%
Artificial Neural Network	83.08% ± 5.34%

Table 3. Mean AUCs of Algorithms trained on the imbalanced dataset.

Algorithm	Mean AUC and Standard Deviation
Logistic Regression	0.88 ± 0.06
Support Vector Machine	0.91 ± 0.05
Artificial Neural Network	0.90 ± 0.05

Table 4. Average cross-validation accuracies of algorithms trained on the balanced dataset.

Algorithm	Average Accuracy and Standard Deviation
Logistic Regression	89.61% ± 4.96
Support Vector Machine	91.37% ± 3.50
Artificial Neural Network	93.35% ± 2.56

Table 5. Mean AUCs of algorithms trained on the balanced dataset.

Algorithm	Mean AUC and Standard Deviation
Logistic Regression	0.94 ± 0.05
Support Vector Machine	0.98 ± 0.02
Artificial Neural Network	0.98 ± 0.02

Results show that all the Algorithms trained on the balanced dataset produces better performance. According to [32] in certain areas such as fraud detection, medical diagnosis and risk management, severe imbalance class distribution is relatively common and is a concerning problem. ML Algorithms are built to reduce errors. As the probability of instances that belong to the majority class is greatly high in imbalanced datasets, the Algorithms are most likely to classify new observations to the majority class. Also, in real life, the cost of False Negative is usually much larger than False Positive, yet ML algorithms penalise both at a similar weight.

5.2. Ideal Machine Learning Model

To evaluate the performance of the implemented Algorithms various matrices are used and one of them is Confusion Matrix which gives results of the various aspects of a model from which it is possible to calculate performance measures such as Accuracy, False Positive Rate and so on. However, the accuracies obtained from the Matrix are not enough to find an accurate measure of a models accuracy as the dataset is split at a particular point. Hence, the K-Fold Cross Validation is used to split the dataset K times where (K = 10). The value 10 is chosen because this is the commonly used value found in the existing literature. In the K-Fold Cross Validation method, the entire dataset is split K number of times where one set is kept as a test set and the remainder as training set as discussed in the Methodology chapter and finally, the average of accuracies obtained from every training is calculated. The results have been already discussed in the previous Subsection. In this Subsection, the results obtained from the Confusion Matrix will be discussed. As found from the experiment that imbalanced dataset strongly reduces the predictive capabilities of Machine Learning models, therefore, only the results obtained from the Confusion Matrices of Algorithms trained on the balanced dataset are considered for comparison and they are provided in the following **Table 6**.

From the table, it can be seen that ANN and SVM have higher TP and TN values than LR. When selecting an ideal model, the FN and FP values must be taken into consideration. FN of ANN is 1 meaning that, out of the test set of data (61 patients) the model for one patient predicted that the patient has CAD but actually the patient is normal. The FP value is 4 which means that four patients were classified to have the disease but they are actually normal. In contrast, LR results with values of 4 and 5 for FN and FP with TN = 13 and TP = 39 whereas, SVM shows similar results compared to ANN.

Table 6. Results of confusion matrices of algorithms trained on the balanced dataset.

Algorithm	TP	FN	FP	TN
Logistic Regression	39	4	5	13
Support Vector Machine	41	2	4	14
Artificial Neural Network	42	1	4	14

6. Conclusion and Future Recommendations

In this research, a prototype system for detection of Coronary Artery Disease is built using Logistic Regression, Support Vector Machine and Artificial Neural Network for comparison of the Algorithms. The dataset used for the research contains medical records of patients who visited Shaheed Rajaei Cardiovascular, Medical and Research Center of Tehran, Iran. After performing Statistical Analysis on the data set, it was found that the dataset does not contain any missing values, however, from the Exploratory Data Analysis, it is evident that there is a class imbalance in the dataset as patients with CAD are higher than Normal patients, to solve this issue, SMOTE Algorithm is applied on the dataset to balance the dataset. Then the Algorithms have been compared on both balanced and imbalanced datasets and required pre-processing steps were carried out before the Algorithms were implemented. Results show that the performance of Support Vector Machine and Artificial Neural Network significantly improved when trained on the balanced dataset however, the overall accuracy of Logistic Regression stayed the same on both sets of data. Various performance matrices were used in the research and the accuracies were cross-validated and their ROC curves were plotted for each fold. Overall, the Artificial Neural Network had the highest average accuracy of 93.35% ± 2.56% and AUC of 0.98 ± 0.02, whereas the Support Vector Machine came quite close with an accuracy of 91.37% ± 3.50% with the same AUC value. In contrast, the Logistic Regression performed CAD prediction with an accuracy of 89.61% ± 4.96% with an AUC value of 0.94 ± 0.05.

Future Recommendations

A limitation of this research is the size of the dataset, hence working on a larger dataset with more features could be a better extension to this research. therefor a larger dataset containing patients with different geographic locations could be ideal. High Blood Cholesterol is another risk factor which is not present in the dataset. Heavy drinking of alcohol, use of drugs could lead to causes of increased blood pressure, stroke and so on could also be considered as contributing risk factors [33]. Also, risk factors for women such as Menopause and emerging non-traditional features for women mentioned in the research work of [34] which are; preterm delivery, Hypertensive disorders of pregnancy, Gestational diabetes, Autoimmune disease, Breast Cancer treatment and depression.

Acknowledgements

There are a number of people whom I would like to express my acknowledgements. First, I would like to thank my Supervisor Mr. H M Mostafizur Rahman for providing his constructive feedback on which I have reflected on to prepare better work. A special thanks to Mrs. Tanzila Islam from whom I was inspired about studying Artificial Intelligence during the time I was completing Higher National Diploma. Then I would like to thank my dear friend Md Ashiqur Rahman with whom I discussed all the relevant aspects of this research. Finally, I would like to express my gratitude towards Dr. Zahra Alizadeh Sani, Roohallah Alizadehsani and Mohamad Roshanzamir who are donors and creators of the dataset. They have published the data online and it is free to use for research purposes. Without their contribution, it would not have been possible to conduct this research.

Conflicts of Interest

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

References

[1] WHO (2018) The Top 10 Causes of Death. https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] Cleveland Clinic (2019) Coronary Artery Disease. https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease

[3] Mitchell, T. (1997) Machine Learning. McGraw-Hill Higher Education, New York.

[4] Alpaydin, E. (2014) Introduction to Machine Learning. MIT Press, Cambridge.

[5] Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin.

[6] Kurt, I., Ture, M. and Kurum, A.T. (2008) Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease. *Expert Systems with Applications*, **34**, 366-374. https://doi.org/10.1016/j.eswa.2006.09.004

[7] Lavesson, N., Halling, A., Freitag, M., Odeberg, J., Odeberg, H. and Davidsson, P. (2009) Classifying the Severity of an Acute Coronary Syndrome by Mining Patient Data. *The Swedish AI Society Workshop*, 27-28 May 2009, Vol. 35, 55-63.

[8] Babaoğlu, I., Fındık, O. and Bayrak, M. (2010) Effects of Principle Component Analysis on Assessment of Coronary Artery Diseases Using Support Vector Machine. *Expert Systems with Applications*, **37**, 2182-2185. https://doi.org/10.1016/j.eswa.2009.07.055

[9] Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B. and Sani, Z.A. (2013) A Data Mining Approach for Diagnosis of Coronary Artery Disease. *Computer Methods and Programs in Biomedicine*, **111**, 52-61. https://doi.org/10.1016/j.cmpb.2013.03.004

[10] Akilaand, S. and Chandramathi, S. (2015) A Hybrid Method for Coronary Heart Disease Risk Prediction Using Decision Tree and Multi Layer Perceptron. *Indian Journal of Science and Technology*, **8**, 1-7. https://doi.org/10.17485/ijst/2015/v8i34/85947

[11] Lo, Y.-T., Fujita, H. and Pai, T.-W. (2016) Prediction of Coronary Artery Disease Based on Ensemble Learning Approaches and Co-Expressed Observations. *Journal of Mechanics in Medicine and Biology*, **16**, Article ID: 1640010. https://doi.org/10.1142/S0219519416400108

[12] Alizadehsani, R., Zangooei, M.H., Hosseini, M.J., Habibi, J., Khosravi, A., Roshanzamir, M., Khozeimeh, F., Sarrafzadegan, N. and Nahavandi, S. (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] 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.

[14] 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

[15] Meng, N., Zhang, P., Li, J., He, J. and Zhu, J. (2018) Prediction of Coronary Heart Disease Using Routine Blood Tests.

[16] Nassif, A.B., Mahdi, O., Nasir, Q., Talib, M.A. and Azzeh, M. (2018) Machine Learning Classifications of Coronary Artery Disease. *International Joint Symposium on Artificial Intelligence and Natural Language Processing*, Pattaya, 15-17 November 2018, 1-6. https://doi.org/10.1109/iSAI-NLP.2018.8692942

[17] Shamsollahi, M., Badiee, A. and Ghazanfari, M. (2019) Using Combined Descriptive and Predictive Methods of Data Mining for Coronary Artery Disease Prediction: A Case Study Approach. *Journal of AI and Data Mining*, **7**, 47-58.

[18] UCI Machine Learning Repository (2019) Z-Alizadeh Sani Data Set. https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani

[19] Raschka, S. (2015) Python Machine Learning. Packt Publishing Ltd., Birmingham.

[20] Liu, Y. (2018) Encoding Categorical Features. https://towardsdatascience.com/encoding-categorical-features-21a2651a065c

[21] Asaithambi, S. (2017) Why, How and When to Scale Your Features. https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e

[22] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) Smote: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[23] Poolsawad, N., Kambhampati, C. and Cleland, J. (2014) Balancing Class for Performance of Classification with a Clinical Dataset. *Proceedings of the World Congress on Engineering*, Vol. 1, 1-6.

[24] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. and Wang, Y. (2017) Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology*, **2**, 230-243. https://doi.org/10.1136/svn-2017-000101

[25] Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. John Wiley & Sons, Hoboken, 398. https://doi.org/10.1002/9781118548387

[26] Gudadhe, M., Wankhade, K. and Dongre, S. (2010) Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network. *IEEE International Conference on Computer and Communication Technology*, Allahabad, 17 September, 2010, 741-745. https://doi.org/10.1109/ICCCT.2010.5640377

[27] Hsu, C.-W., Chang, C.-C., Lin, C.-J., *et al.* (2003) A Practical Guide to Support Vector Classification.

[28] Karayılan, T. and Kılıç, Ö. (2017) Prediction of Heart Disease Using Neural Network. *IEEE International Conference on Computer Science and Engineering*, Helsinki, 21-23 August 2017, 719-723. https://doi.org/10.1109/UBMK.2017.8093512

[29] Scikit-Learn (2019) 1.17. Neural Network Models (Supervised). https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[30] Sonkar, P. (2017) Application of Supervised Machine Learning to Predict the Mortality Risk in Elderly Using Biomarkers.

[31] Ray, S. (2018) Improve Your Model Performance Using Cross Validation (in Python and R). https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r

[32] Minh, H. (2018) The Top 10 Causes of Death. https://medium.com/jamesblogs/handling-imbalanced-data-in-classification-problems-7de598c1059f

[33] Healthline Editorial Team (2018) Risk Factors for Coronary Artery Disease (CAD). https://www.healthline.com/health/coronary-artery-disease/risk-factors

[34] Garcia, M., Mulvagh, S.L., Bairey Merz, C.N., Buring, J.E. and Manson, J.E. (2016) Cardiovascular Disease in Women: Clinical Perspectives. *Circulation Research*, **118**, 1273-1293. https://doi.org/10.1161/CIRCRESAHA.116.307547