Thailand Statistician
July 2025; 23(3): 677-691
http://statassoc.or.th
Contributed paper

Predicting Ischemic Heart Disease and Determining Its Risk Factors: A Comparison of Various Classification Methods in Machine Learning

Muhammad Yaqoob* and Farhat Iqbal
Department of Statistics, University of Balochistan, Quetta, Pakistan
*Corresponding author; e-mail: mohammadyaqoob065@gmail.com

Received: 10 September 2021
Revised: 13 September 2022
Accepted: 4 October 2023

Abstract

This study was conducted to identify the important risk factors of ischemic heart disease (IHD) amongst the population of Balochistan and to determine the most accurate machine learning (ML) algorithms for the prediction of IHD. The data were collected from 300 individuals (100 IHD cases and 200 control cases) on common risk factors of IHD. The risk factors included marital status, physical activity, socioeconomic position, type of oil used for cooking, diet, body mass index, blood pressure, random blood sugar, history of known disease, and cholesterol level. We employed linear discriminant analysis (LDA), artificial neural networks (ANN), naïve Bayes (NB) and random forest (RF) classification methods. The data were randomly partitioned into training (70%) and testing (30%) sets. The classification methods were evaluated based on their accuracy rates, sensitivity, specificity, positive and negative predictive values, and area under the receiver operating characteristic curve. The results of the study indicated that ANN was the most accurate classification method, with an accuracy of 88.89%, followed by NB, LDA and RF, with accuracy rates of 86.67%, 85.56% and 84.44%, respectively. Moreover, in most classification methods, blood pressure, cholesterol levels, physical activity, diet, BMI, and family history were found as the important factors for developing the risk of IHD. The study's results indicated that ML methods, especially ANN, can be employed for accurately predicting the state of IHD and determining the important risk factors.

Keywords: Ischemic heart disease, risk factors, prediction, machine learning algorithms.

1. Introduction

Over the last several decades, cardiovascular disease has become one of the leading causes of death in developing countries and developed countries (WHO 2017). Ischemic heart disease (IHD) is a form of chronic cardiovascular disease that occurs when the blood supply gets low in the heart area because of the buildup of lipid deposit plaque in the walls of arteries. The risk of IHD develops due to some controllable risk factors such as physical activity, obesity, smoking, cholesterol level, and diet, etc., and some uncontrollable risk factors such as age, gender, and family history (Bhatia 2010). The impact of these risk factors on IHD is not uniform in different regions across the world. The risk factors

which are significantly common in different regions of the world for developing IHD are smoking habits, age, blood pressure, and cholesterol level. Likewise, physical activity is negatively associated with the development of IHD among the population of Southern Europe (Keys et al. 1984). Similarly, these risk factors of IHD are not uniform among the people in South Asian countries like Pakistan, India, and Bangladesh (Bhopal 1999). In a particular region, the identification of risk factors is essential for physicians to accurately predict the condition of patients with IHD.

Among various statistical methods, epidemiologists have employed classification methods to determine the significant risk factors of IHD. The classification methods are data analysis tools used to construct a function that distinguishes data in multiple (binary or categorical) classes according to their features. The widely used statistical classification method for determining risk factors and predicting the outcome of a disease is logistic regression (LR). The LR is a form of the parametric regression model in which the response variable is a dichotomous variable and the explanatory variables may be of any type (dichotomous, categorical and continuous) (Peng et al. 2002). The statistical classification methods provide valid results in determining the association of risk factors with the disease when their assumptions are fulfilled. But, in real-world problems, the data may not always support the assumptions of a particular classification method. In such situations, alternative methods have been employed by the researcher to model and predict the response variable.

Among different classification methods, machine learning (ML) algorithms have been proven as powerful tools for researchers. ML is the branch of computer science that utilizes past data without relying on rule-based programming to create a detectable pattern from a given problem and use the learned knowledge for future predictions. Popular ML methods used to tackle classification problems on basis of their similar features are Linear Discernment Analysis (LDA) by Fisher (1936), Artificial Neural Network (ANN) proposed by McCulloch and Pitts (1943), Naïve Bayes (NB) by Good (1950), Decision Trees (DT) by Breiman et al. (1984) and Random Forest (RF) suggested by Breiman (2001) among others. In the past, numerous studies have utilized ML methods for predicting and determining risk factors for different kinds of diseases such as cancer (Delen et al. 2005; Tran et al. 2021), kidney disease (Kate et al. 2016; Luo et al. 2021), diabetes (Olisah et al. 2022) and hypertension (Adavi et al. 2016) etc. Moreover, ML methods are also used for gene identification of various diseases (Wazir et al. 2020; Hamraz et al. 2021) and feature selection (Khan et al. 2019; Qasim and Algamal 2021).

The predictive performance of ML algorithms for binary classification problems can be evaluated by using different Evaluation methods. Some popular methods of evaluation of classification methods are sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV). The measure of sensitivity (true positive rate) used to determine how the given classification method is correctly identified true positive cases from all positive cases. Therefore, sensitivity can be measured by using the ratio of true positive cases from true positive and false negative cases. Conversely, specificity (true negative rate) is used to measure the performance of classification methods for identifying true negative cases. In this method, the fraction of true negative cases by the sum of true negative and false positive cases can be employed to measure the specificity (Trevethan 2017). The PPV is the fraction between the total number of true-positive cases and the sum of the total number of true-positive cases and false-positive cases. It measures the probability of the positive cases given that it was predicted positive. Whereas the NPV is the fraction between the total number of true negative cases and the sum of the total number of true negative cases and false-negative cases. It measures the probability of the negative cases, given that it was predicted as negative.

For cardiovascular disease risk prediction, Weng et al. (2017) used four ML algorithms: RF, LR, gradient boosting, and ANN. In their study, it was observed that ML algorithms provide accurate classification methods for cardiovascular risk prediction. Goldstein et al. (2016) compared statistical

methods with different ML algorithms for the risk prediction of cardiovascular disease. Their study revealed that ML algorithms could improve the prediction ability of cardiovascular disease than the statistical method in the presence of non-linear relationships, correlated predictors and interaction between predictors among the observations. Ayatollahi et al. (2019) compared two ML algorithms for predicting coronary artery disease among the Iranian population. The various risk factors of cardiovascular disease in the Pakistani environment are discussed in different literature. Abbas et al. (2009) inferred that smoking is the most common risk factor among men than women, whereas cholesterol level, body mass index and diabetes mellitus are more common among women compared to men. For both men and women, blood pressure is a significant risk factor for IHD in the population of Islamabad (Pakistan). Likewise, Hussain et al. (2013) found that the most common risk factors of IHD among women are hypertension and diabetes mellitus and among men is smoking in one of the remote areas of Pakistan. Iqbal et al. (2014) reviewed the risk factors of IHD using two classification methods (LR and DT) in the native population of Balochistan. Similarly, a study related to risk factors of cardiovascular disease among the population of Pakistan was carried out by Barolia and Sayani (2017). They investigated the possible risk factors of cardiovascular disease in their study, such as high blood pressure, unhealthy diet, lack of physical activity, smoking, psychosocial issues, and obesity.

To date, in Pakistani settings, no such epidemiological study has been conducted to generate a prediction system for IHD and to determine its important risk factor using different ML algorithms. In this study, we compared the performance of various classification methods in ML, such as LDA, ANN, NB, DT, and RF, and determined the most accurate ML algorithm based on different evaluation methods. For this purpose, we evaluated their predictive performance on training and testing data sets for this purpose. The reason for using both training and testing data is to avoid the problem of overfitting in ML algorithms. Because of overfitting, evaluation methods of ML algorithms on training data may provide the best model and perform poorly in the case of test data. Apart from selecting the best classification method, we also aimed to determine the possible risk factors of IHD among the native population of Balochistan, Pakistan.

The rest of the paper is arranged as follows: Section 2 describes the data along with the risk factors of IHD, various ML algorithms used in this study and the evaluation methods. The results from different evaluation methods of ML algorithms and the important risk factors of IHD from different classification methods are reported in Section 3. Section 4 discusses results about the important risk factors of IHD and various evaluation methods of ML algorithms with the related results from prior literature. Finally, the study is concluded in Section 5.

2. Materials and Methods

2.1. Data

The data were collected from different hospitals in Quetta, Balochistan. The data set consists of a total of 300 individuals, of which 100 individuals had the IHD and 200 individuals had not IHD (IHD was absent) for both genders. Of these, 150 were male and 150 were female. The disease cases were classified based on their admission to cardiology wards of the hospitals that had the first onset of Angina and Myocardial Infarction. For the selection of control cases, the age and gender of the participants were matched with the disease cases. From both groups, the demographic information along with risk factors of IHD was obtained through a structured questionnaire. For consent, a form was designed and signed by the participants.

2.2. Risk factors

The risk factors for the study consisted of 10 categorical variables, which are summarized in Table 1. These risk factors were marital status (single or married), physical activity (less than 30 min/day, 30-45 minutes/day and more than 45 minutes/day), socioeconomic position (income of Rs.12,000, between Rs.12,500 and Rs.35,000, more than Rs.35,000), type of oil used for cooking (oil or ghee), diet (balanced diet or unbalanced diet), body mass index (less than 25, between 25 and 30, greater than 30), blood pressure (Normal, prehypertension, Stage One, Stage Two), random blood sugar (normal, early diabetes, established diabetes), history of the known disease (obesity, hypertension, diabetes, myocardial infarction, none) and cholesterol level (Desirable, Borderline, High).

2.3. Machine learning algorithms

2.3.1 Linear discriminant analysis

The LDA algorithm is the first pattern recognition method introduced by Fisher (1936), among other classification methods. This method is very similar to the method of Bayesian. Because it employs the posterior probability distribution function for predicting the class labels of the cases. The class label of the new case will be the one with the highest posterior probability value. For *p*-dimensional space, this procedure assumes that the distribution of the features from each class follows multivariate Gaussian distribution with the identical variance-covariance matrix. In the case of one-dimensional space (one feature) X, the procedure of LDA for classification of categorical random variable Y, with K number of distinct classes, let $f_k(x) = P_r(X = x \mid Y = k)$ be the conditional probability function of data in the k^{th} class, π_k is the prior probability of observations in the k^{th} class, with $\sum_{k=1}^{K} \pi_k = 1$, and $P_r(X)$ is the marginal distribution of data or normalized constant of the posterior distribution. Then posterior function $p_k(x) = P_r(Y = k \mid X = x)$ takes the form (see James et al. 2013)

$$P_r(Y = k \mid X = x) = \frac{\pi_k P_r(X \mid Y = k)}{P_r(X)}, \tag{1}$$

or

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{2}$$

In the procedure of LDA, the data in each class follows normal or Gaussian distribution, i.e. $X \sim N(\mu_k, \sigma_k^2)$. The density function $f_k(x)$ can be written as follows:

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{1}{2\sigma_k^2}(x - \mu_k)^2 \right\}, \tag{3}$$

where $\pi \approx 3.1416$, μ_k is the mean of observations in the k^{th} class and σ_k^2 is the variance of observations in the k^{th} class. The posterior distribution function of the k^{th} class can be obtained by solving the Bayes classifier, which is defined as,

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{1}{2\sigma_k^2}(x - \mu_k)^2 \right\}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left\{ -\frac{1}{2\sigma_l^2}(x - \mu_l)^2 \right\}}. \tag{4}$$

Since the procedure of LDA also assume that the variances of all categories are identical, i.e. $\sigma_1^2 = \sigma_2^2 ... = \sigma_K^2 = \sigma^2$. Then, the posterior function takes the form

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu_k)^2\right\}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu_l)^2\right\}}. \tag{5}$$

After taking the logarithm and rearranging the posterior probability function, we get the following function for decision making

$$\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k), \tag{6}$$

here, π_k can be estimated by dividing the sum of training observations of k^{th} class by the total number of observations in the training data set, i.e.

$$\hat{\pi}_k = \frac{n_k}{n},$$

where n is the total number of training observations and n_k is the total number of training observations in the k^{th} class. The mean (μ_k) can be estimated from the training data of k^{th} class, while the estimate of variance (σ^2) can be obtained from the weighted average of sample variances of each class, which are defined as,

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i,$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} \left(x_i - \hat{\mu}_k\right)^2. \tag{7}$$

The prediction about the observation $X = x$ to the k^{th} class can be made by the largest value of $\delta_k(x)$ (James et al. 2013), which can be estimated as,

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k). \tag{8}$$

2.3.2 Artificial neural network

The concept of ANN is related to the natural interconnected neuron system of the brain, which was introduced by McCulloch and Pitts (1943). In the procedure of ANN, various kinds of nodes, also called neurons, were arranged in different layers. A simple ANN consists of one input layer, one output layer, and one hidden layer, which is located between the input layer and the output layer. Weights of real numbers connect each layer. The input layer corresponds to the different features (independent variables) in the procedure. The hidden layer is the combination of various nodes that get information from features, apply some mathematical operation, and then give the results to the output layer. The output layer is the predicted values of the task (classification or regression), which the ANN model estimates. For the mathematical operation, each node used a perceptron. The perceptron in each node first uses a linear function and then a non-linear function for better prediction of class labels. The linear function is the summation of features with their corresponding weights, and the non-linear function is an activation function (Hastie et al. 2009). Commonly, the sigmoid function is used as an activation function. The activation function is used to estimate the class of observation by using a threshold value. One type of sigmoid function is the logistic function. Let Y be a binary dependent random

variable, which takes values 0 and 1 and $X=(X_1,X_2,...,X_p)$ be the p number of features. Then the logistic function, for $Y=1$ given X, is defined as,

$$P_r(Y=1\mid X)=\frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{p}\beta_i X_i)}}, \tag{9}$$

where the intercept β_0 and slope coefficients $\beta_i\,(i=1,2,...,p)$ can be estimated from observed data using the maximum likelihood method.

2.3.3 Naïve Bayes

The method of NB was proposed by Good (1950) to correctly classify observations according to their class labels based on the conditional probability of given observations belonging to a specific class. Like LDA, the NB also utilizes the Bayesian rules for the prediction of the unknown class label based on features. But the method of NB does not need the assumption of multivariate Gaussian distribution for class-conditional probability. However, the procedure of NB can be carried out with a strong assumption of independence among features. Nevertheless, it was observed that this method also performed very well in real-world applications where the assumption of independence of features is false. For *p*-dimensional space, the procedure of NB is defined as follows:

Let Y be a dependent random variable, which can take categorical values $C_i\,(i=1,2,...,K)$, and $X=(X_1,X_2,...,X_p)$ be the collection of independent features. For class C_i, NB employed the following posterior probability distribution function $P_r(Y=C_i\mid X)$, with components $P_r(X)$, $P_r(Y)$ and $P_r(Y=C_i\mid X)$, to estimate the unknown classes of cases, where $P_r(Y=C_i\mid X)$ is the posterior probability of class C_i; $P_r(Y)$ is the prior probability of class C_i, which can be estimated from training data or expertise of experts; $P_r(X)$ is the probability of the feature values. It is also called the marginal distribution of data; $P_r(X\mid Y=C_i)$ is the class-conditional probability distribution, which can be observed from data. In this case, it can be determined by the product of each class-conditional probability distribution of features, which is given by

$$P_r(X\mid Y=C_i)=\prod_{j=1}^{p}P_r(X_j\mid Y=C_i). \tag{10}$$

Then, the final posterior density function becomes,

$$P_r(Y=C_i\mid X)=\frac{\prod_{j=1}^{p}P_r(X_j\mid Y=C_i)P_r(Y)}{P_r(X)}. \tag{11}$$

In general, the decision about the observation for i^{th} class (C_i) can be made by observing the highest posterior probability of that class. If the posterior probability of the i^{th} class is higher than the posterior probabilities of other classes then the observation belongs to the i^{th} class, i.e., $P_r(Y=C_i\mid X)>P_r(Y=C_l\mid X)$ for $i\neq l$ (Kuhn and Johnson 2013).

2.3.4 Random forest

The RF is an ensemble method of classification trees (Breiman et al. 1984) introduced by Breiman (2001), which is extensively used to handle pattern recognition problems. This method is capable of solving various tasks related to classification and regression. Let there are p number of features in a

given data set and m is the subset of that features (typically the number m is defined as $\sqrt{p}$. The subset of various features can be randomly selected by using a bootstrapped resampling technique on the training data set. After the selection of a subset of features, the RF use to construct different combinations of trees. Therefore, all the trees will be different from each other. The decision about the class of new cases can be made based on the majority voting of trees or the mode of the different tree's outcomes.

2.3.5 Evaluation method

Usually, a 2×2 contingency table that includes true positive (TP) values, true negative (TN) values, false positive (FP) values and false negative (FN) values have been employed for the evaluation of binary classification problems. The TP represents the number of positive cases correctly predicted, TN represents the total number of negative cases correctly predicted, FP represents the number of negative cases incorrectly predicted, and FN represents the number of incorrectly predicted positive cases. From a 2×2 contingency table, we have determined the accuracy rate, sensitivity, specificity, PPV and NPV in order to evaluate various classification methods. For further evaluation, we have employed the receiver operating characteristic (ROC) curve to assess the internal accuracy of the classification methods (Zhou et al. 2002).

In order to implement the classification methods, we partitioned the entire data set into training and testing (unseen) data sets in 7:3. To overcome the problem of overfitting, a standard 10-fold cross-validation is used to fit these four classification methods on the training data set. Later on, all evaluation methods of four ML algorithms were employed by using the testing data set. R-Statistical programming software (version 3.6.1) was employed for statistical analysis (R Core Team 2018).

3. Results

Table 1 represents the cross-tabulation of participants with IHD and without IHD and with a range of other features, including gender, age, marital status, physical activity, socioeconomic status, type of oil, diet, BMI, blood pressure, random blood sugar, family history and cholesterol. A sample comprised of similar proportions of gender, including male (50%), and female (50%). Similarly, equal proportion of age, including age group of 30-44 (5%), 45-54 (20%), 55-64 (46%) and ≥ 65 (29%). Among IHD cases, a small portion belonged to a high level of physical activity category (3%) and in contrast, among healthy cases, a significant portion belonged to a high level of the physical category (23%). The use of ghee as cooking oil in food was greater among IHD cases (69%) than in control cases (55%). Similarly, a significant portion of IHD cases (94%) has taken unbalanced food as compared to control cases (60%). No patient of IHD had normal and high normal blood pressure, while 17% had normal and 33% had high normal blood pressure among the sample of healthy participants. Similarly, a significant proportion of IHD cases were obese (48%) and most healthy cases were overweight (42%) compared to other categories. Among those diagnosed with IHD, at least one of their relative who had myocardial infarction was 34%, obesity was 26%, hypertension 21% and diabetes 10%. Similarly, among the healthy participants, at least one of their relatives with myocardial infarction was 10%, obesity was 26.5%, hypertension 32.5% and diabetes 21%. Most of the IHD cases had high cholesterol levels (63%), while most of the control cases had borderline cholesterol levels (56.5%) as compared to other categories.

Table 1 The distribution of IHD and control cases and its risk factors

Variables	Levels	Total cases (%)	IHD cases (%)	Control cases (%)	p-value
Gender	Male	150 (50)	50 (50)	100 (50)	1.00
	Female	150 (50)	50 (50)	100 (50)	
Age	30-44	15 (5)	5 (5)	10 (5)	1.00
	45-54	60 (20)	20 (20)	40 (20)	
	55-64	138 (46)	46 (46)	92 (46)	
	≥ 65	87 (29)	29 (29)	58 (29)	
Marital status	Single	34 (11.33)	15 (15.00)	19 (9.50)	0.2212
	Married	266 (88.67)	85 (85.00)	181 (90.50)	
Physical activity	Mild	127 (42.33)	60 (60.00)	67 (33.50)	< 0.001
	Moderate	123 (41.00)	37 (37.00)	86 (43)	
	High	50 (16.67)	3 (3.00)	47 (23.50)	
Socioeconomic status	Low income	152 (50.67)	50 (50.00)	102 (51.00)	0.9848
	Average income	115 (38.33)	39 (39.00)	76 (38.00)	
	High income	33 (11.00)	11 (11.00)	22 (11.00)	
Type of oil	Oil	121 (40.33)	31 (31.00)	90 (45.00)	0.02743
	Ghee	179 (59.67)	69 (69.00)	110 (55.00)	
Diet	Balanced	86 (28.67)	6 (6.00)	80 (40.00)	< 0.001
	Unbalanced	214 (71.33)	94 (94.00)	120 (60.00)	
BMI	Normal	65 (21.67)	5 (5.00)	60 (30.00)	< 0.001
	Over weight	131 (43.67)	47 (47.00)	84 (42.00)	
	Obese	104 (34.67)	48 (48.00)	56 (28.00)	
Blood pressure	Normal	34 (11.33)	0 (0.00)	34 (17.00)	< 0.001
	High normal	67 (22.33)	0 (0.00)	67 (33.50)	
	Stage one	62 (20.67)	20 (20.00)	42 (21.00)	
	Stage two	98 (32.67)	53 (53.00)	45 (22.50)	
	Stage three	39 (13.00)	27 (27.00)	12 (6.00)	
Random blood sugar	Diabetes	14 (4.67)	2 (2.00)	12 (6.00)	0.1258
	Early diabetes	102 (34.00)	40 (40.00)	62 (34.00)	
	Established diabetes	184 (61.33)	58 (58.00)	126 (61.33)	
Family history	Obesity	79 (26.33)	26 (26.00)	53 (26.50)	< 0.001
	Hypertension	86 (28.67)	21 (21.00)	65 (32.50)	
	Diabetes	52 (17.33)	10 (10.00)	42 (21.00)	
	Myocardial infarction	55 (18.33)	34 (34.00)	21 (10.50)	
	None	28 (9.33)	9 (9.00)	19 (9.50)	
Cholesterol	Desirable	57 (19.00)	6 (6.00)	51 (25.5)	< 0.001
	Borderline	144 (48.00)	31 (31.00)	113 (56.5)	
	High	99 (33.00)	63 (63.00)	36 (18.00)	

Furthermore, associations between the samples of IHD patients and the healthy individuals were determined by using the chi-square test and the results are also reported in Table 1. The result of the chi-square revealed that seven features (risk factors of IHD, i.e. physical activity, type of oil used for cooking, diet, body mass index, blood pressure, history of known disease, and cholesterol) were statistically significant at the 5% level. These risk factors and three insignificant factors (marital status,

socioeconomic position, and random blood sugar) were retained for predictive modeling. The remaining two features (gender and age) were statistically insignificant because of the equal proportion in the group of IHD and control cases.

3.1. Comparison of different classification methods

Figure 1 showed the predictive ability of the four classification methods based on training and testing data sets. The results of the training data set revealed that the RF achieved the highest accuracy rate (98.10%) for predicting disease and non-disease cases than all other classification methods. The next classification method, which yields a decent accuracy rate, was ANN (with an accuracy rate of 89.52%), followed by LDA and NB, with an accuracy rate of 88.10% and 86.19%, respectively. For the testing data set, the accuracy rate of the ANN model (91.11%) was found to be the best among all classification methods, followed by NB, LDA, and RF, resulting in an accuracy rate of 86.67%, 85.56% and 84.44%, respectively.

Comparison of predictive ability among the different classification methods based on sensitivity, specificity, positive predictive values, and negative predictive values on both training and testing data sets are reported in Table 2. For the training data set, the values of sensitivity were found to be the highest for RF (98.57%), and the lowest for ANN (87.14%) and LDA (87.14%). Also, the value of specificity was found to be the highest for RF (97.86%), but the lowest for NB (82.86%). The PPV ranged from 73.03% (NB) to 95.83% (RF) and NPV ranged from 93.23.65% (LDA) to 99.28% (RF).

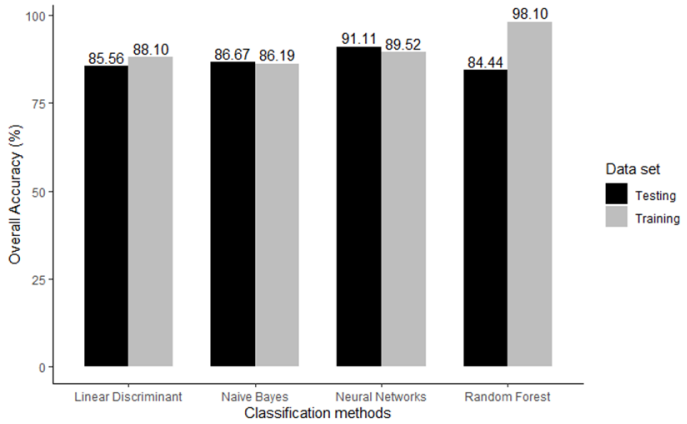

Figure 1 The total accuracy rate of four classification methods

Table 2 The comparisons of the predictive ability among four classification methods

Methods	Training data set (70%)				Testing data set (30%)			
	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
LDA	0.8714	0.8857	0.7922	0.9323	0.7667	0.9000	0.7931	0.8852
ANN	0.8714	0.9000	0.8133	0.9333	0.7667	0.9500	0.8846	0.8906
NB	0.9286	0.8286	0.7303	0.9587	0.8667	0.8667	0.7647	0.9286
RF	0.9857	0.9786	0.9583	0.9928	0.7000	0.9167	0.8077	0.8594

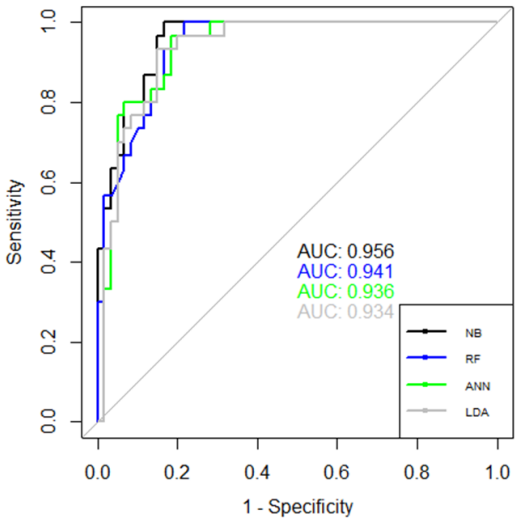

Figure 2 The ROC curve of four different classification methods and their AUC values

Based on the testing data set, the sensitivity value was 86.67% for NB, 76.67% for LDA, 76.67% for ANN and 70% for RF. According to the result, ANN had the highest specificity value (95%), followed by RF (91.67%), LDA (90%), and NB (86.67%). Furthermore, the positive predictive values of the four classification methods ranged from the lowest of 76.47% for the NB to the highest of 88.46% for the ANN and negative predictive values of the four classification methods ranged from the lowest of 85.94% for the RF and to the highest of 92.86% for the NB.

The performance of the four classification methods was also assessed using the area under the ROC curve (for testing data set), and the results showed in Figure 2. The AUC showed that NB achieved a higher accuracy (with an AUC value of 0.95) than other classification methods, followed by RF (with an AUC value of 0.941), ANN (with an AUC value of 0.936), and LDA (with an AUC value of 0.934). Moreover, the result of DeLong's statistical test of AUC showed no significant difference among each classification method.

3.2. Risk Factors Importance

The most important risk factors of IHD from each classification method are illustrated in Figure 3. In all classification methods, it can be observed that blood pressure is the most important risk factor for developing the risk of IHD. The second most important risk factor positively associated with the development of IHD was cholesterol level. In most of the classification methods, physical activity was the third most important risk factor (for LDA, ANN, and NB). Similarly, other risk factors that played a significant role in developing an accurate classification method are diet, body mass index, family history, and cooking oil. In contrast, the risk factors, which had quite less impact on developing risk of IHD, were random blood sugar, socioeconomic position, and marital status.

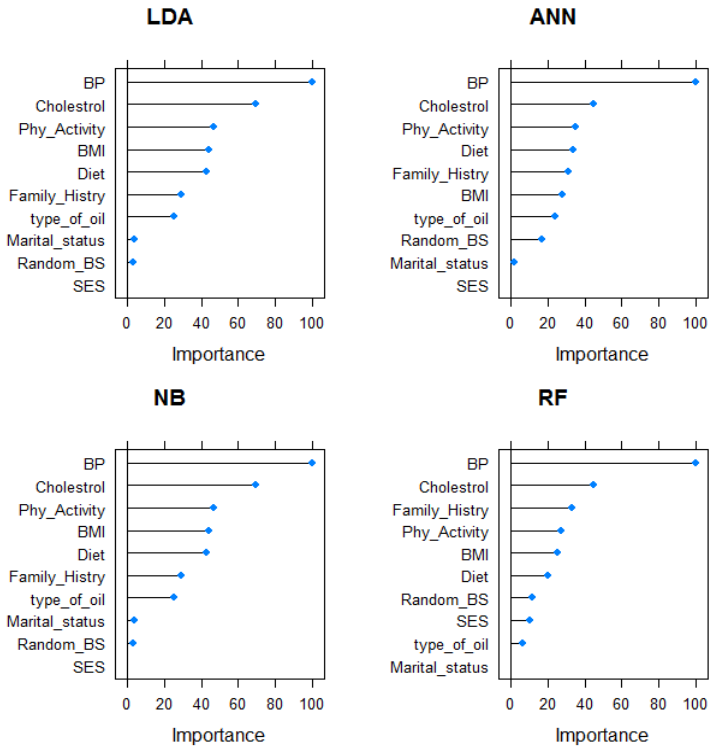

Figure 3 The order of the most important risk factors from each classification methods

4. Discussions

This study's results determined that blood pressure, cholesterol level, physical activity, diet, body mass index, family history, and type of cooking oil contributed to developing the risk of IHD. Iqbal et al. (2014) revealed that physical activity, diet, body mass index, cholesterol level, and family history are significantly related to Balochistan's log-odd ratio of IHD among the native population. Peter et al. (1998) found that blood pressure, total cholesterol, LDL cholesterol, unbalanced diet, high BMI and family history are statistically significant risk factors for developing IHD. Also, these risk factors were significant in other studies of IHD (Keys et al. 1984; Yusuf et al. 2004).

Furthermore, all classification methods retained in this study indicated that blood pressure and cholesterol level are the most important risk factors for IHD. MacMahon et al. (1990) suggested that high level of blood pressure is positively related to the development of IHD. Globally, a high cholesterol level is associated with a risk of IHD (Yusuf et al. 2004). In addition, past studies revealed that blood pressure and total cholesterol could be used to predict the risk of IHD (Anderson et al., 1991; Peter et al. 1998). In the present study, the majority of the classification methods showed that physical activity is the third most important risk factor for IHD. Similar studies claimed that there is an inverse relationship between physical activities with cardiovascular disease in men (Lee et al. 2000), women (Lee et al. 2001), older adults (Batty 2002) and middle-aged individuals (Koolhaas et al. 2016).

In the present study, based on the area under the ROC curve, it is detected that the NB is a more effective classification method for predicting IHD cases. Ross et al. (2016) employed the same evaluation method to identify peripheral artery disease using RF and LR. In their study, RF performed better as compared LR (AUC, 0.87 vs 0.79, respectively; p-value = 0.03). However, Kurt et al. (2008)

suggested that ANN is better for predicting coronary artery disease among the population of Turkey (with an AUC of 0.78) than DT, LR, and radial basis function. In addition, Kim and Kang (2017) claimed that ANN is a better prediction method (AUC = 0.79) than the Framingham risk score (AUC = 0.393) for the prediction of coronary heart disease. This study found no statistically significant difference of AUC values among the classification methods.

Furthermore, this study indicates that each classification method shows similar predictive performance for predicting IHD. But, RF and ANN yield decent predictive abilities given the values of accuracy rate, sensitivity, specificity, PPV and NPV than other classification methods. For the training data set, the RF effectively predicted IHD cases with the highest accuracy rate (98.10%). Whereas, for the testing data set, the predictive performance of RF was inferior (accuracy rate = 84.44%) than other classification methods. In contrast, ANN achieved the highest accuracy rate (88.89%) not only on the testing data set but also in an almost similar accuracy rate (89.05%) on the training data set. In addition, past studies suggested that ANN is an effective method in epidemiological studies for disease prediction. HeidarAbadi et al. (2017) found that ANN performed better than Bayesian networks, DT and support vector machines (SVMs) for predicting the type of pain of spinal cord injury with the highest value of accuracy (91%), sensitivity (89%), specificity (95%), positive predictive value (91%) and negative predictive value (96%). Adavi et al. (2016), suggested that ANN has a higher accuracy rate than LR for the prediction of hypertension and diabetes. In another epidemiological study of IHD, Ayatollahi et al. (2019) found that SVMs provide more accurate classification methods compared to ANN in terms of accuracy, sensitivity, positive predictive value and area under the ROC curve.

In the current study, we had several limitations. Firstly, we only considered the age group of middle and older individuals (age $\geq$ 30) and their risk factors, which excluded younger individuals. Secondly, we excluded women with pregnancy from our study. Therefore, the result of this study may not be generalized for younger adults and pregnant women of the given population. Another limitation of this study was that we employed only four ML algorithms, such as LDA, ANN, NB, and RF. As future work, to get more accurate prediction systems for IHD, we intend to employ advanced methods of other ML, such as stochastic gradient boosting and ensemble methods on the same data set.

As a result of this study, we found that the use of classification methods in ML algorithms would not only enhance practitioners' understanding of accurately predicting the state of IHD but also determine its important risk factors. The awareness of the important risk factors can reduce the risk of IHD among the individuals of the given population via controlling blood pressure, getting adequate exercise, avoiding unhealthy food, maintaining cholesterol levels and maintaining an ideal weight.

5. Conclusions

This study demonstrated that all four ML algorithms effectively performed the prediction of IHD. Among them, the RF clearly outperformed all other classification methods for the training data set but failed to maintain the best predictive performance on the testing data set. In contrast, ANN provided the most accurate classification method on the testing data set and resulted in quite a similar accuracy on the training data set. NB was a more accurate classification method for the area under the ROC curve. However, the AUC of ANN was not significantly different from all classification methods. Moreover, in most classification methods, high blood pressure, cholesterol level, physical inactivity, unbalanced diet, BMI, and family history were the most important risk factors for developing IHD. The study's results indicated that ML methods, especially ANN, can be employed for accurately predicting the state of IHD and determining the important risk factors among the native population of Balochistan.

References

Abbas S, Kitchlew AR, Abbas S. Disease burden of ischemic heart disease in Pakistan and its risk factors. Ann Pak Inst Med Sci. 2009;5(3):145-150.

Adavi M, Salehi M, Roudbari M. Artificial neural networks versus bivariate logistic regression in prediction diagnosis of patients with hypertension and diabetes. Med J Islam Repub Iran. 2016; 30: 1-5.

Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. J Am Heart Assoc. 1991; 121(1): 293-298.

Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a comparison between two data mining algorithms. BMC Public Health. 2019; 19(1): 1-9.

Barolia R, Sayani AH. Risk factors of cardiovascular disease and its recommendations in Pakistani context. J Pak Med Assoc. 2017; 67(11): 1723-1729.

Batty GD. Physical activity and coronary heart disease in older adults: a systematic review of epidemiological studies. Eur J Public Health. 2002; 12(3): 171-176.

Bhatia SK. Biomaterials for clinical applications. New York: Springer; 2010.

Bhopal R, Unwin N, White M, Yallop J, Walker L, Alberti KGMM. Heterogeneity of coronary heart disease risk factors in Indian, Pakistani, Bangladeshi, and European origin populations: a cross-sectional study. BMJ. 1999; 319(7204): 215-220.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Florida: Chapman & Hall; 1984.

Breiman L. Random forest. Mach Learn. 2001; 45: 5-32.

Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005; 34(2): 113-127.

Fisher RF. The use of multiple measurements in taxonomic problems. Ann Eugen. 1936; 7(2): 179-188.

Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017; 38(23): 1805-1814.

Good IJ. Probability and the weighing of evidence. London: Charles Griffin; 1950.

Hamraz M, Gul N, Raza M, Khan DM, Khalil U, Zubair S, Khan Z. Robust proportional overlapping analysis for feature selection in binary classification within functional genomic experiments. Peer J Comput. Sci. 2021; 7: e562; https://doi.org/10.7717/peerj-cs.562

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction (2nd edition). New York: Springer; 2009.

HeidarAbadi NN, Hakemi L, Kolivand P, Safdari R, Saeidi MG. Comparing performances of intelligent classifier algorithms for predicting type of pain in patients with spinal cord injury. Electron. Physician. 2017; 9(7): 4847-4852.

Hussain S, Sattar U, Azhar MA. Risk factors for ischemic heart disease in Southern Punjab. Pak Heart J. 2013; 46(4): 232-237.

Iqbal F, Jafri YZ, Siddiqi AR, Sabir MA. Determining risk factors for ischemic heart disease using logistic regression and classification tree. Sylwan. 2014; 158(6): 69-87.

James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013.

Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V. Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak. 2016; 16(1):1-11.

Keys A, Menotti A, Aravanis C, et al. The seven countries study: 2,289 deaths in 15 years. Prev Med. 1984; 13(2): 141-154.

Khan Z, Naeem M, Khalil U, Khan DM, Aldahmani S, Hamraz M. Feature selection for binary classification within functional genomics experiments via interquartile range and clustering. IEEE Access. 2019; 7: 78159-78169.

Kim JK, Kang S. Neural network-based coronary heart disease risk prediction using feature correlation analysis. J Healthc Eng. 2017; 2017: 1-13.

Koolhaas CM, Dhana K, Golubic R, Schoufour JD, Hofman A, Rooij FA, Franco OH. Physical activity types and coronary heart disease risk in middle-aged and elderly persons: the Rotterdam study. Am J Epidemiol. 2016; 183(8): 729-738.

Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.

Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl. 2008; 34(1): 366-374.

Lee IM, Sesso HD, Paffenbarger RS Jr. Physical activity and coronary heart disease risk in men: does the duration of exercise episodes predict risk? Circulation. 2000; 102(9): 981-986.

Lee IM, Rexrode KM, Cook NR, Manson JE, Buring JE. Physical activity and coronary heart disease in women: is “no pain, no gain” passe? JAMA. 2001; 285(11): 1447-54.

Luo XQ, Yan P, Zhang NY, et al. Machine learning for early discrimination between transient and persistent acute kidney injury in critically ill patients with sepsis. Sci Rep. 2021; 11(1): 1-12.

MacMahon S, Peto R, Cutler J, et al. Blood pressure, stroke, and coronary heart disease. part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet. 1990; 335(8692): 765-774.

McCulloch WS, Pitts WH. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943; 5(4): 115-133.

Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Programs Biomed. 2022; 220: 106773; https://doi.org/10.1016/j.cmpb.2022.106773

Peng CYJ, So TSH, Stage FK, John EPS. The use and interpretation of logistic regression in higher education journals: 1988-1999. Stud High Educ. 2002; 43(3): 259-293.

Peter WF, Wilson, Ralph B, et al. prediction of coronary heart disease using risk factor categories. Am Heart J. 1998; 97(18): 1837-1847.

Qasim OS, Algamal ZY. Improving feature selection for credit scoring classification using a novel hybrid algorithm. Thail Stat. 2021; 19(3): 593-605.

R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2018; URL http://www.R-project.org/.

Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. J Vasc Surg. 2016; 64(5): 1515-1522.

Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. 2021; 13(1): 1-7.

Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health. 2017; 5: 307.

Wazir B, Khan DM, Khalil U, Hamraz M, Gul N, Khan, Z. Regulatory genes identification within functional genomics experiments for tissue classification into binary classes via machine learning techniques. J Pak Med Assoc. 2020; 70(12): 2356-2362.

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017; 12: e0174944; https://doi.org/10.1371/journal.pone.0174944

WHO. Cardiovascular diseases: Key Facts; 2017. Available from: http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yusuf S, Hawken S, Ôunpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): a case-control study. Lancet. 2004; 364(9438): 937-952.

Zhou X, Obuchowski N, NcClish D. Statistical methods in diagnostic medicine. New York: Wiley-Interscience; 2002.