Research Journal in Advanced Sciences
ISSN: 2709-281X (Online) ISSN: 2709-2801 (Print)

Advanced Sciences

Comparing predictive performance of k-nearest neighbors and support vector machine for predicting ischemic heart disease

Research article

Muhammad Yaqoob[1*], Farhat Iqbal[1] & Samiha Zahir[2]

[1]Department of Statistics, University of Balochistan, Quetta, Pakistan

[2]*Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan*

Corresponding author: mohammadyaqoob065@gmail.com

Article History:
Received: 2020-08-15
Revised: 2020-09-16
Accepted: 2020-09-01
Published: 2020-10-12

Keywords:
Ischemic heart disease, prediction, support vector machine, k-nearest neighbors

How to cite?
Yaqoob, M., Iqbal, F., & Zahir, S. (2020). Comparing predictive performance of k-nearest neighbors and support vector machine for predicting ischemic heart disease. *Research Journal in Advanced Sciences*, 1(2). Retrieved from https://royalliteglobal.com/rjas/article/view/391

Published in Nairobi, Kenya by Royallite Global in the **Research Journal in Advanced Sciences**

Abstract

This research compared the predictive performance of k-nearest neighbors (k-NN) and support vector machine (SVM) for predicting ischemic heart disease (IHD), based on its important risk factors. For this study, the information on the risk factors of IHD were collected from 300 individuals. Among them, 100 were recruited from the IHD group and 200 from the control group. Furthermore, the entire data set was randomly partitioned into training and testing set by the ratio of 7:3 respectively. The k-NN and SVM models were fitted on the training data set with 10-fold cross-validation. Both models were evaluated based on their accuracy rate, sensitivity, specificity, and area under the receiver operating characteristics (ROC) curve (AUC) on both training and testing datasets. The results from different evaluation methods revealed that SVM outperformed compared to k-NN with a higher value of accuracy (86.67%), sensitivity (80%), specificity (90%), and AUC (94.1%) on testing data set. However, no statistical significant differences were found between SVM and k-NN. In addition, both models showed that blood pressure, cholesterol, physical activity, BMI, diet, family history, and type of oil are the most important risk factors for increasing the chance of IHD. The results indicated that SVM and k-NN models can be used to develop a predictive system for IHD using its important risk factors.

Scan and Read

Public Interest Statement

In this research, we developed the most accurate machine learning (ML) algorithm such as, k-nearest neighbors (k-NN) and support vector machine (SVM) for predating ischemic heart disease (IHD). These two ML algorithms can be used without any strict assumptions about the relationship between the risk factors (features). The predictive performance of k-NN and SVM were compared on both testing and training data set using different evaluation methods. Also, both methods provided important risk factors of IHD among the given population. Moreover, the methods of k-NN and SVM can be used by the physician for accurately predicting patients of IHD based on its important risk factors.

Introduction

During last several decades, the cardiovascular diseases remained a common cause of death all over the world, including Pakistan. It has been estimated that more than 17.9 million demises occurred as a result of cardiovascular diseases in 2016 [59]. One type of cardiovascular disease is ischemic heart disease (IHD), it occurs when the supply of blood reduces into the area of heart, due to the lipid plague in the arteries. It is believed that IHD is mainly caused by some modifiable risk factors such as blood pressure, cholesterol level, physical activity, diet, random blood sugar, socioeconomic status, marital status, and some unmodified risk factors such as family history, age, and gender [59]. Therefore, based on its important risk factors, a valid measurement tool can be used to accurately estimate the state of IHD among individuals.

In past, numerous classical statistical methods have been used by practitioners for prediction of IHD based on its risk factors. The widely used statistical method for predicting different kinds of diseases is logistic regression (LR). The LR method employs log-odd ratio as a response variable to construct a better relationship with the independent variables (risk factors). The method of LR works; based on some assumptions such as, no multicollinearity among independent variables, each observation must be independent within the independent variables (no auto-correlation) and these must not contain any extreme values (outliers) [59]. Therefore, conventional statistical technique of LR provide valid results in the prediction of disease when their assumptions are fulfilled. However, in real-world situations, the observations may not fulfill the assumptions of a particular prediction technique. As a result, different non-parametric prediction methods have been employed by researchers to accurately predict the state of various kinds of diseases.

The recent developed algorithms of machine learning (ML) provides more powerful tools to accurately predict different kinds of disease in epidemiological studies. The ML is a branch of artificial intelligence that uses prior knowledge (data) from a given problem and make a delectable pattern for the future prediction without any particular programming. Two common ML algorithms for disease prediction are k-nearest neighbors (k-NN) [59] and support vector machine (SVM) [59]. Both algorithms work without strict assumptions regarding the structure of observations and any kind of relationship between the independent variables.

Thus, these two non-parametric methods has ability to deliver more accurate prediction system compared to classical statistical methods.

Moreover, the method of k-NN, SVM and random forest (RF) were employed for the detection of skin cancer by Murugan et al [59]. In their study, SVM provided better result compared to k-NN and RF for skin cancer detection. Moreover, the methods of k-NN and SVM was implemented by Pereira et al [59] for arterial pulse waveform recognition. The predictive performance of the SVM classifier was better than k-NN in their study. But, for the prediction of diabetes cases k-NN showed better performance than SVM [59]. Likewise, the SVM model was found better than the artificial neural network (ANN) model for predicting coronary heart disease (CHD) [59].In the Pakistani setting, [59].investigated the possible risk factors of cardiovascular disease such as high blood pressure, unhealthy diet, lack of physical activity, smoking, psycho-social issues, and obesity. Among the population of Baluchistan, the use of ghee for cooking, unbalanced diet, and high BMI, high cholesterol, family history of myocardial infarction were positively associated risk factors of IHD but high physical activity was negatively associated risk factor of IHD [59].

Till date, to the best of our knowledge, no such study has been conducted by researchers to utilize SVM and k-NN algorithms for the prediction of IHD in the Pakistani context. Therefore, the aim of this study is to compare the performance of two machine learning algorithms such as k-NN and SVM to generate an accurate prediction system for IHD based on its risk factors. In order to select the best prediction system between k-NN and SVM, the results of these different evaluation methods have been compared on both testing and training datasets. The comparison based on both testing and training data set can reduce the problem of over-fitting of the models.

Material & Methods

Data & Risk factors

In the present study, a total of 300 participants were recruited (100 IHD cases and 200 control cases). The admission at the cardiology ward and the report card from a physician were the inclusion criteria of IHD cases and normal cases were selected by matching their age and gender with IHD cases. The information about the risk factors were obtained from both group (IHD cases and control cases) such as, blood pressure (normal, pre hypertension, stage one, stage two), cholesterol level (desirable, borderline, high), history of the related disease (obesity, hypertension, diabetes, myocardial Infarction, none), random blood sugar (normal, early diabetes, established diabetes), body mass index (less than 25, between 25 and 30, greater than 30), type of oil used for cooking (oil or ghee), diet (balanced diet or unbalanced diet), socioeconomic position (income of Rs.12,000, between Rs.12,500 and Rs.35,000, more than Rs.35,000), physical activity (less than 30 min/day, 30-45 minutes/day and more than 45 minutes/day) and marital status (single or married).

Machine Learning Algorithms

K-Nearest Neighbors

The method of k-NN was developed by [59] in order to efficiently classify observations in accordance to their corresponding classes. The rule of k-NN is a non-parametric pattern recognition method, which is used in these situations where sample information about observations and their corresponding categories are correctly defined. The procedure of k-NN works on the principle that for each category; the values of features are similar. For classification problems, the rule of the k-NN algorithm employs the k training points, which are close to each other in distances. The prediction of the class of new data points is based on more values of the nearest neighbor with different classes (mode value) from new data points out of k data point in each feature [59]. Similarly, for regression problems, the process of k-NN also uses k nearest neighbor values for prediction of a new data point but, in this case; it uses mean or median value of dependent variable from k-most similar values.

Support Vector Machine (SVM)

The method of SVM developed by [59] in order to accurately classify values of binary dependent variables by employing a unique hyperplane. The hyperplane is a boundary line between two classes, which can be constructed by maximizing the margin between support vectors (data points under the convex hulls) of both classes. In SVM method, the structure of hyperplane is not uniform for different dimensions of the features. For one dimension, the hyperplane would be a point, for two-dimension the hyperplane would be a straight line, and for more than two dimensions the hyperplane would be a plane. Generally, the optimal hyperplane for p-dimensional features, from observations $x_1, \dots, x_n$ with class labels$y_i i = 1,2, \dots, n \in \{1, -1\}$ can be determined by employing a vector of weights (β^t) of the feature and a scalar of intercept (β_0) defined as,

$$\beta^t x_i + \beta_0 = 0$$

From this hyperplane, the decision boundary between two class labels would be estimated by using the following inequalities,

$$y_i(\beta t x_i + \beta_0) \geq +1 if y_i = 1$$
$$y_i(\beta t x_i + \beta_0) \leq -1 if y_i = -1 \qquad \forall i = 1,2, \dots, n,$$

where the vector of weights of the features(βt), and the scalar of intercept (β_0) can be estimated by maximizing margin of hyperplane (gap between support vectors).

The inequalities indicates that if the value of $f(x_i) = \beta^t x_i + \beta_0$ is positive, then the features belongs to first class (1), and if the value of $f(x_i)$ is negative then the features must go to the second class (-1) [59].

In this study, the entire data set was randomly partitioned into training (70%) and testing (30%) datasets. Both models were trained using training datasets, and their performance were

evaluated on both training and testing datasets. Usually, accuracy rate, sensitivity, specificity and area under the receiver operating characteristics (ROC) curve (AUC) are employed in order to assess the best model. In binary classification problem, these evaluation methods can be constructed by using following quantities: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) from the 2×2 contingency table. Model having high accuracy rate, sensitivity, specificity and AUC is considered best predictive model [59]. R-Statistical programming software version (3.6.1) was employed for fitting both algorithms on training data sets and describing other statistical measures [59].

Results

Associations of Risk Factors with IHD:

Associations between the samples of IHD patients and the healthy individual were determined by using the chi-square test and the results were reported in Table 1. The results of the chi-square association test revealed that out of ten features seven features (risk factors of IHD) are statistically significant at a 5% level, which were retained for predictive modeling. Moreover, the three insignificant risk factor are marital status (p-value = 0.2212), socioeconomic position (p-value = 0.9848) and random blood sugar (p-value = 0.1258).The statistically significant risk factors of IHD are physical activity (p-value < 0.001), type of oil used for cooking (p-value = 0.02743), diet ((p-value < 0.001)), body mass index (p-value < 0.001), blood pressure (p-value < 0.001), history of known disease (p-value < 0.001), and cholesterol (p-value < 0.001). In addition, the remaining two features (gender and age) were statistically insignificant due to the equal proportion in the group of IHD and control cases.

Table 1: Distribution of IHD cases and control cases.

Variables	Levels	IHD cases (%)	Control cases (%)	P-value
Gender	Male	50 (50)	100 (50)	1.00
	Female	50 (50)	100 (50)	
Age	30-44	5 (5)	10 (5)	1.00
	45-54	20 (20)	40 (20)	
	55-64	46 (46)	92 (46)	
	≥65	29 (29)	58 (29)	
Marital status	Single	15 (15.00)	19 (9.50)	0.2212
	Married	85 (85.00)	181 (90.50)	
Physical activity	Mild	60 (60.00)	67 (33.50)	< 0.001
	Moderate	37 (37.00)	86 (43)	
	High	3 (3.00)	47 (23.50)	

Socioeconomic status	Low income	50 (50.00)	102 (51.00)	0.9848
	Average income	39 (39.00)	76 (38.00)	
	High income	11 (11.00)	22 (11.00)	
Type of oil	Oil	31 (31.00)	90 (45.00)	0.02743
	Ghee	69 (69.00)	110 (55.00)	
Diet	Balanced	6 (6.00)	80 (40.00)	< 0.001
	Unbalanced	94 (94.00)	120 (60.00)	
BMI	Normal	5 (5.00)	60 (30.00)	< 0.001
	Over weight	47 (47.00)	84 (42.00)	
	Obese	48 (48.00)	56 (28.00)	
Blood pressure	Normal	0 (0.00)	34 (17.00)	< 0.001
	High normal	0 (0.00)	67 (33.50)	
	Stage one	20 (20.00)	42 (21.00)	
	Stage two	53 (53.00)	45 (22.50)	
	Stage three	27 (27.00)	12 (6.00)	
Random blood sugar	Diabetes	2 (2.00)	12 (6.00)	0.1258
	Early diabetes	40 (40.00)	62 (34.00)	
	Established diabetes	58 (58.00)	126 (61.33)	
Family history	Obesity	26 (26.00)	53 (26.50)	< 0.001
	Hypertension	21 (21.00)	65 (32.50)	
	Diabetes	10 (10.00)	42 (21.00)	
	Myocardial infarction	34 (34.00)	21 (10.50)	
	None	9 (9.00)	19 (9.50)	
Cholesterol	Desirable	6 (6.00)	51 (25.5)	< 0.001
	Borderline	31 (31.00)	113 (56.5)	
	High	63 (63.00)	36 (18.00)	

Comparison between k-NN and SVM

The evaluations of SVM and k-NN based on accuracy rate, sensitivity, and specificity are illustrated in Table 2. The SVM algorithm obtained the best accuracy rate for both training and testing datasets (0.8762, 0.8667 respectively), proving that k-NN has slightly lower predictive performance than SVM based on accuracy rate on both training and testing datasets (0.8429, 0.8556 respectively). Moreover, the result showed a higher value of sensitivity achieved by SVM for training data set (0.900) and for testing data set (0.800). In addition, for training data-set SVM resulted in best value of specificity (0.86) in comparison with k-NN (0.84), but for testing data set the higher value of specificity achieved by k-NN.

Table 2 comparison of accuracy, sensitivity and specificity of SVM and k-NN algorithms.

Evaluation methods	SVM			k-NN	
	Training	Testing		Training	Testing
Accuracy	0.8762	0.8667		0.8429	0.8556
Sensitivity	0.9000	0.8000		0.8429	0.7000
Specificity	0.8643	0.9000		0.8429	0.9333

Figure 1 highlights the area under the ROC curves of both SVM and k-NN algorithms on the testing data set. It is clearly visible from figure 1 that SVM performed better than k-NN with AUC of 0.944. However, DeLong's test for significant difference resulted in non-significant difference between these two ML algorithms (p-value= 0.8913).

Figure 2 shows the most important risk factors of IHD form SVM and k-NN algorithms. Both ML algorithms carried similar results of important risk factors. In Figure 2, it is clear that blood pressure is the most important risk factors for predicting IHD. The other most important risk factor of IHD were cholesterol level, physical activity, BMI, diet, family history, and type of oil, respectively. On the other hand, marital status, random blood sugar, and socio-economic status showed zero importance for developing the risk of IHD.

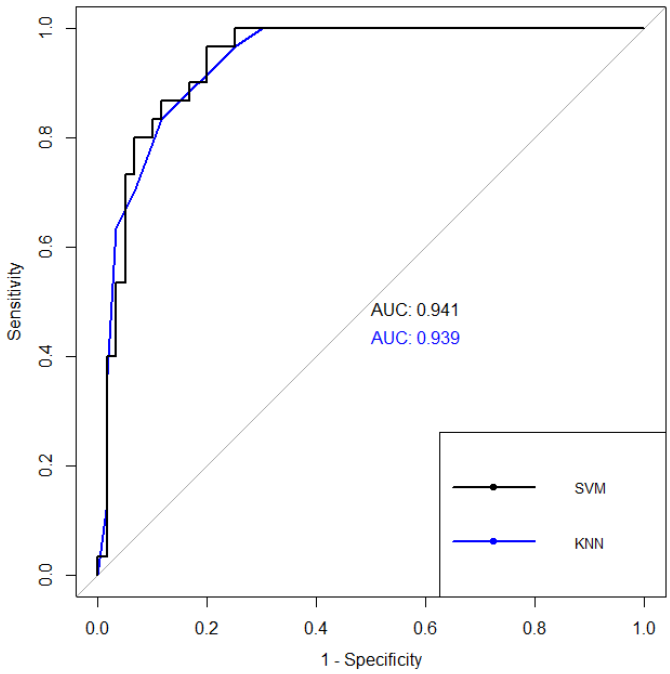

Figure 1: Area under the ROC curves of SVM and k-NN algorithm on testing data set.

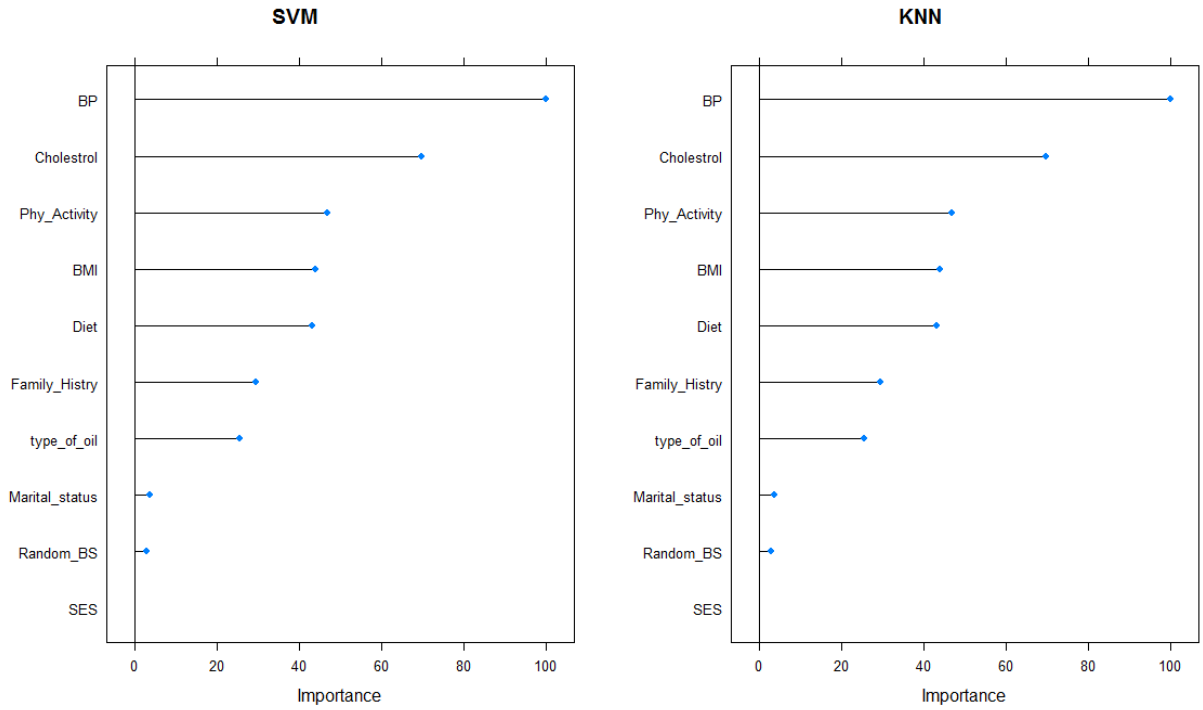

Figure 2: The important risk factors of IHD from SVM and K-NN algorithms.

Discussion

In the prediction of IHD cases, we found that both ML algorithms (SVM and K-NN) show promising potential. However; the level of accuracy, sensitivity, specificity, and AUC of SVM is recorded higher on both training and testing datasets as compared K-NN algorithm. Moreover, similar results were reported in the prior literature for predicting different kinds of diseases. Murugan et al [59], found that the SVM algorithm provided the highest value of accuracy (85.72%), sensitivity (87.68%), and specificity (83.76%) for skin cancer detection as compared to k-NN and RF. Likewise, for cardiovascular risk assessment, a better accuracy rate was obtained by SVM (95.2%) than k-NN [59]. Ayatolahi et al [59], referred that SVM provides significantly higher predictive performance than an ANN for the prediction of coronary heart disease in terms of accuracy rate and area under the ROC curve. Joana et al [59], found that SVM achieved significantly higher accuracy for the recognition of pathological arterial pulse wave. Also, SVM showed better results compared to the naive Bayes algorithm for predicting diabetes in terms of accuracy rate, sensitivity, specificity, and precision.

Conversely, it has been proven in the prior literatures that the k-NN algorithm is also an efficient prediction method for diagnosing various type of disease. For diabetes risk prediction k-NN resulted in higher accuracy (95.80%) than SVM (80.92%) [59]. Rajaguru and Chakravarthy [59], predicted breast cancer using k-NN and decision tree algorithms. In their study, the results showed that k-NN outperformed decision trees in the classification of breast cancer. In the current study, no statistical significant difference were found between the performance of SVM and k-NN.

The findings of the current study, from different evaluation methods of SVM and k-NN, corresponds with the results of Iqbal et al [59]. They employed same data-set for determining risk factors of IHD using LR and classification tree. But the predictive performance of LR and classification tree were evaluated only on training data-set in their study. They reported that blood pressure, cholesterol level, diet, and physical activity are associated with the risk of IHD. In this study, the results of variable importance from SVM and k-NN showed that blood pressure, cholesterol level, physical activity, BMI, diet, family history, and type of oil used for cooking are highly contributing in increasing the chance of IHD. Moreover, the significant impact of these risk factors can be seen in other studies of heart disease. MacMahon et al [60], showed that the chance of CHD is high with a high level of blood pressure as compared with a low level of blood pressure among individuals without a previous history of vascular disease. Yusuf et al [60], evaluated that the cholesterol level is a pathogenic factor for IHD: worldwide. Different studies showed that physical activity is inversely related to the risk of IHD [60-60]. Flint et al [60], revealed that BMI is positively associated with the risk of developing CHD among men and women.

As a result of this study, we found that the utilization of both ML algorithms shall not only enhance the understanding of the practitioners to accurately identify the individuals with a high probability of IHD but also determine its important risk factors. The awareness of

significant risk factors can reduce the risk of IHD among individuals of the given population via controlling blood pressure, getting adequate exercise, avoiding unhealthy food, maintaining a healthy cholesterol level, and maintaining an ideal weight.

Conclusion

The results of this study discovered that both predictive models have similar response for predicting IHD and determining its important risk factors. However, SVM achieved a higher value of accuracy, sensitivity, specificity, and AUC than the k-NN model regarding both testing and training datasets. Despite this, the AUC of k-NN was not statistically significantly different from the AUC of SVM. Both models resulted that blood pressure, cholesterol, physical activity, BMI, diet, family history, and type of oil are most important risk factors for increasing chances of IHD. Therefore, SVM and k-NN models can be used to develop a predictive system for IHD using its important risk factors.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgement: This research was supported by MS. Memoona Zahid for collecting data.

Conflict of Interest: The authors declare no conflict of interest.

Disclaimer Statement: A publication of this research paper is requirement of MPhil degree for submitting my thesis at University of Balochistan, Quetta, Pakistan.

Authors Bionote

Muhammad Yaqoob, is an MPhil student, specializing in Biostatistics Supervised by associate prof. Dr. Farhat Iqbal at department of statistics, University of Balochistan, Quetta, Pakistan. Samiha Zahir, doing her MPhil at department of Statistics,Quaid-i- Azam University in Islamabad. Pakistan.

Authors' contributions: Muhammad Yaqoob: formal analysis, methodology, software, visualization, writing original draft, validation review & editing draft; Farhat Iqbal: conceptualization, supervision, methodology, review & editing draft; Samiha Zahir: writing original draft, validation, visualization, and review & editing draft.

References

[1] WHO, 'cardiovascular diseases: Key Facts. 2017', 2017. http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] S. K. Bhatia, 'Biomaterials for clinical applications'. *Springer Science+Business Media*. 2010.

[3] C. Y. J. Peng, T. S. H. So, F. K. Stage, and E. P. John, 'The use and interpretation of logistic regression in higher education journals: 1988-1999', *Research in Higher Education*, *43*, 259–293, 2002.

[4] T. M. Cover, and P. E. Hart, 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory,* 13, 21-27, 1967.

[5] V. Vapnik, 'Support-Vector Networks', *Machine Leaming*, 20, 273-297, 1995.

[6] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, 'Detection of skin cancer using support vector machine, random forest and k-nearest neighbor classifiers', *Journal of Medical Systems*, 269, 1-9, 2019.

[7] T. Pereira, J. Paiva, and J. Cardoso, 'An automatic method for arterial pulse waveform recognition using k-nearest neighbor classifiers and support vector machine classifiers', *Medical & biological engineering & computing*, 54, 1049-1059, 2015.

[8] S. Bano, and M. N. A. Khan, 'A framework to improve diabetes prediction using k-NN and SVM', *International Journal of Computer Science and Information Security (IJCSIS),* 14, 450-460, 2016.

[9] H. Ayatollahi, L. Gholamhosseini, and M. Salehi, 'Predicting coronary artery disease: A comparison between two data mining algorithms', *BMC Public Health*, *19*, 1-9, 2019.

[10] R. Barolia, and A. H. Sayani, 'Risk factors of cardiovascular disease and its recommendations in Pakistani context', *The Journal of the Pakistan Medical Association*. 67, 1723–1729, 2017.

[11] F. Iqbal, Y. Z. Jafri, A. R. Siddiqi, and M. A. Sabir, 'Determining risk factors for ischemic heart disease using logistic regression and classification tree', *Sylwan*, 158, 69-87, 2014.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, 'An introduction to statistical learning', *New York: Springer*, 2013.

[13] X. Zhou, N. Obuchowski, and D. NcClish, 'Statistical methods in diagnostic medicine' *New York: Wiley-Interscience,* 2002.

[14] R Core Team (2018). R: A language and environment for statistical computing. Vienna, Austria: *R Foundation for Statistical Computing. URL http://www.R-project.org/.*

[15] S. Joana, P. J. Cardoso, and T. Pereira, 'Supervised learning methods for pathological arterial pulse wave differentiation: a support vector machine and neural Networks Approach', *International Journal of Medical Informatics,* 109, 30-38, 2017.

[16] R. H. Harikumar and S. R. S. Chakravarthy, 'Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer', *Asian Pacific Journal of Cancer Prevention,* 20, 3777-3781, 2019.

[17] S. MacMahon 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', *The Lancet,* 335, 765–774, 1990.

[18] S. Yusuf et al., 'Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study', *The Lancet,* 364, 937– 952, 2004.

[19] I. M. Lee, D. Howard, Sesso, S. Ralph, and Paffenbarger, 'Physical activity and coronary heart disease risk in men does the duration of exercise episodes predict risk?' *American Heart Association,* 102, 981-986, 2000.

[20] I. M. Lee, K. M. Rexrode, N. R. Cook, J. E. Manson, and J. E. Buring, 'Physical activity and coronary heart disease in women: Is "no pain, no gain" passe?' *JAMA,* 285, 1447-54, 2001.

[21] G. D. Batty, 'Physical activity and coronary heart disease in older adults: A systematic review of epidemiological studies', *European Journal of Public Health,* 12, 171-176, 2002.

[22] C. Koolhaas, et al., 'Physical activity types and coronary heart disease risk in middle-aged and elderly persons: the Rotterdam study', *American journal of epidemiology,* 183, 729-738, 2016.

[23] A. J. Flint, 'Body mass index, waist circumference, and risk of coronary heart disease: a prospective study among men and women', *Obesity research & clinical practice,* 4, 171-181, 2010.