ONLINE FIRST

ORIGINAL ARTICLE

Empirical Study and Statistical Analysis of Risk Factors for Cardiovascular Disease (CVD) in Sindh Province

Khushboo Ishaq[1*], Nazia Parveen Gill[1], Raja Ilyas[1]

[1] *Department of Statistics, University of Sindh Jamshoro, Sindh, Pakistan*
Correspondence: *khushbooishaq34@gmail.com*
doi: 10.22442/jlumhs.2025.01224

ABSTRACT

OBJECTIVE: To gather information from Sindh's Civil Hospitals to discover crucial clinical risk factors for CVD in that state.

METHODOLOGY: Data was gathered between January and December 2022, focusing on 21 possible cardiovascular risk factors. Hypotheses were tested, and essential clinical risk factors were found using logistic regression. Additionally, this study utilizes algorithms such as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM) for classification and prediction purposes by splitting the data into training and validation sets. ROC curves are also utilized to evaluate these machine-learning classifiers.

RESULTS: The study showed that 1358(75.2%) out of 1800 respondents were CVD affected. Moreover, logistic regression results for hypothesis testing showed that multiple variables are statistically significant risk factors for CVD patients at the $\alpha = 0.05$ level of significance for developing cardiovascular disease (CVD). Additionally, the Random Forest model outperforms DT, LR, and SVM as the most accurate predictor of CVD (77% accuracy).

CONCLUSION: Based on the findings, Random Forest (RF) outperformed with 77% accuracy Then, the existing models are used in terms of classification and predictions.

KEYWORDS: Cardiovascular Disease, Clinical Data, Health Informatics, Public Health, Machine Learning

INTRODUCTION

Today's most critical challenge facing the health sector is introducing healthcare systems with relatively high facilities that could detect diseases and provide timely treatment, improving patient outcomes. Cardiovascular disease (CVD) is a class of diseases affecting the heart and blood vessels. It is a leading cause of death and disability worldwide, posing a significant public health challenge. CVD encompasses various conditions, including coronary artery disease, heart failure, stroke, and peripheral arterial disease, among others. Coronary artery disease is the most common type of CVD and occurs when there is a buildup of plaque in the arteries that supply blood to the heart. Heart failure can lead to shortness of breath, fatigue, and fluid retention. Stroke is a condition that arises when the blood supply to the brain is interrupted or reduced, typically due to a blockage or rupture of a blood vessel. This interruption of blood flow can result in brain damage, leading to various neurological deficits such as paralysis, speech difficulties, or cognitive impairments. Several risk factors contribute to the development of cardiovascular disease. Some of the most common ones include Smoking, high blood pressure, high cholesterol levels, obesity, diabetes, a sedentary lifestyle, poor diet, excessive alcohol consumption, and a family history of CVD. Advanced age and certain genetic factors also increase the risk. Several risk factors contribute to the development of cardiovascular disease (CVD). This study uses statistical and machine learning methods for categorization and prediction to examine cardiovascular disease's major clinical risk factors in Sindh, Pakistan. It emphasizes how vital early detection and model comparison are to improving the diagnosis of CVD.

Recent studies have investigated various risk factors associated with cardiovascular disease (CVD), highlighting key findings through logistic regression and machine learning classifiers.

Studies by Mbassi LS 2023[1] and Lin XN 2024[2] focused on examining CVD risk factors, specifically lipids and lipoproteins. They found that Turkish individuals had high levels of hepatic lipase and fasting triglycerides and that physical inactivity and Smoking were prevalent, particularly among men. Another study identified cigarette smoking as a significant risk factor for CVD, especially among young smokers[3].

Jiang Q et al. [4] and Gluvic ZM et al. [5] explored the relationship between heart rate and CVD through a systematic literature review. In another study, it was found that middle-aged adults with a sibling affected by CVD were more likely to develop the disease themselves[6]. The study looked at CVD prevalence in Canada, finding high rates of Smoking, obesity, physical inactivity, low income, hypertension, and diabetes among the population[7].

In another study, obesity was linked to a higher risk of CVD in women, particularly coronary disease, stroke, and heart failure[8]. Nelson RH 2013[9], in their study, examined different types of hyperlipidemia, such as elevated cholesterol and triglycerides, and their connection to CVD.

Mukherji D 2013[10] used machine learning techniques, including decision trees and logistic regression, to predict heart disease, finding that combined models improved prediction accuracy. Yaqub FN 2019[11] studied the link between chronic kidney disease (CKD) and risk factors like CVD, diabetes, and hypertension, concluding that these conditions were significant predictors of CKD.

The study by Iqbal R 2012[12] explored how blood pressure and other factors, such as age and lifestyle, contribute to CVD. The authors examined medical risk factors for myocardial infarction (heart attack), identifying atherosclerosis, ischemic heart disease, hypertension, and obesity as significant contributors[13].

Ayub M et al.[14] emphasized the importance of education in increasing awareness of modifiable CVD risk factors in Karachi, Pakistan. Sajid MR 2020[15] found that sleep satisfaction, diet, and physical activity were key risk factors for CVD, using a statistical approach that examined linear and non-linear factors. Nusinovici S et al.[16] compared the machine learning models to logistic regression, finding that machine learning performed well in CVD prediction.

Muhammad S 2022[17] conducted a study based on diet style and the relation of lifestyle. The results found that out of 400 participants, 53.6% of the patients who use open spices were affected by cardio disease, and the remaining were not. Zulfiqar N 2019[18] conducted a study to explore the risk factors related to cardiovascular disease in Islamabad. The results revealed that cholesterol is a significant factor in developing cardiovascular disease.

Koene RJ 2016[19] reviewed the shared risk factors between CVD and cancer, such as obesity, diabetes, hypertension, and dyslipidemia. Ajeganova S 2021[20] performed a systematic review of existing literature based on systemic lupus erythematosus (SLE) and atherosclerotic cardiovascular disease (ACVD). This study follows the PRISMA procedure to collect data from the PubMed database. This study found that patients who are affected with SLE have a high chance of being CVD patients. Van Bussel EF 2021[21] conducted a study based on a literature review on traditional risk factors (TRF) related to CVD. This study includes 12 previous studies with 11 cohorts. The study found that factors like sex, diabetes and, SBP, HDL are significant predictors of CVD with increasing age.

Al-Jafar R et al.[22] investigated the effect of fasting during Ramadan on CVD, finding no significant difference in CVD development between those who fasted and those who did not. The study by authors Aburto NJ et al.[23] found that higher potassium intake could lower the risk of CVD based on randomized trials.

Niedhammer I 2021[24] explored the relationship between job insecurity and the development of CVD based on meta-analysis and systematic review. This study investigated how this insecurity and CVD affect employees, varying by age, sex, unemployment rate, and type of welfare regime. The studies by Malambo P et al.[25] and Huang YQ et al.[26] conducted a systematic review on the influence of built-in environmental attributes that affect cardiovascular disease. The study found a strong relationship between environmental characteristics and CVD risk factors.

Nag T 2013[27] presented a literature search-based study on CVD risk factors. The authors used the data from 1968 to 2012 from PubMed sources, and this study found that hypertension and diabetes are highly spread among Asians and multiple factors are cause of CVD in patients.

Akioyamen LE et al.[28] conducted a systematic and meta-analysis-based study to explore the relationship between cardiac risk factors and cardiac in familial hypercholesterolemia (FH). The result showed that genetic affection is related to CVD (FH), and other factors like Smoking, hypertension, and diabetes, if present in the history of patients more than one-fourth of the year, are found significant in CVD (FH).

Worrall-Carter L 2011[29] conducted a systematic research-based study on CVD-affected women. The data was taken from the health database system (HDS), which indicates that females are more likely affected by comorbid disease, which leads to CVD spread in women.

Gasevic D 2015[30] highlighted that hypertension, diabetes, obesity, and Smoking in black people of Canada are more likely to be present and highly affected than Chinese, Arabs and Hispanics. Statistical and machine learning techniques have been used in previous studies to investigate CVD risk variables; however, Sindh-specific research is still a small number[40-43].

This study compares model performance to close that gap and identifies important indicators for better CVD prediction.

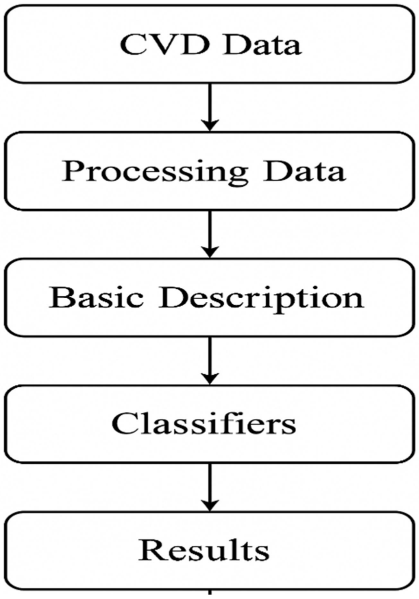
METHODOLOGY

Data is gathered from the civil hospitals of six Sindh province districts: Hyderabad, Nawabshah, Mirpurkhas, Sukkur, Thatta and Larkana. A sample of 300 samples is selected from each district of Sindh province through the hospital's cardiac ward by using the simple random sample for the year 2022. The six civil hospitals in Sindh Province's Human Ethical Committees are consulted for ethical clearance before heart disease patients can be included. Collecting the sample 300 patients from each hospital's cardiac ward consists of a total cumulative sample of 1800 individuals in this cardio risk factor study[39].

Study Variables

The individual output, along with relevant risk factors, are included in the CVD data. The variables included in the comprehensive dataset include age, gender, height, weight, systolic and diastolic blood pressure, cholesterol and its types, alcohol usage, Smoking, exercise, education level, income, cardiovascular disease, and body mass index (BMI). There are two categories for the response variable, CVD: "affected" and "not affected". In addition, noise, inconsistencies, and any missing observations were removed from the data by the specialist's advice and using statistical methods, i.e. mode or median for the attribute data. The flow chart of the work is given below.

Figure I: Flow Chart for Data Processing



Classification

Classification is organizing data into uniform classes based on similarities found in the data. Data that is both structured and unstructured can be classified into different classes[34]. The initial stage in the process is to classify the class of given data points accurately. These classes' common names are target, label, and categories. Various statistical and mathematical techniques include categorization, including linear programming, decision trees, and artificial neural networks. This study also uses the classification technique because the response variable is classified into two categories: "Affected" or "Not Affected" of CVD.

Random Forest (RF)

The random forest technique was first developed by Ho TK 1995[35]. This technique employs a probabilistic domain approach and was re-introduced by Iftikhar H et al.[36]. It is a popular machine-learning method for classification and regression tasks. It works by building a collection (or "forest") of decision trees, where each tree is constructed using a random subset of the data.

RF uses a method called bootstrapping, which randomly selects parts of the dataset for training, while the remaining data (called the out-of-bag or OOB data) is used to check how well the model performs. The trees in the forest are grown by splitting the data into smaller groups at each step based on randomly selected features. The trees are not pruned, so they grow as large as possible. Each tree gives a prediction, and the final result is determined by a majority vote from all the trees, ensuring a more accurate prediction through randomness[37].

Decision Tree (DT)

A decision tree (DT) is a popular predictive modeling and classification method in supervised learning. It works by splitting data into groups based on specific conditions. For classification, the target variable has distinct values. Decision trees help solve decision-making problems by building models that analyze data and make predictions. The process involves five main steps: First, the dataset is split into training and validation sets. The training data is treated as the root of the tree. Features are then selected to divide the data, and if the features have continuous values, they are categorized. The data is further split into smaller subsets based on these features. This process continues until the tree reaches its end, or "leaves." When making predictions, the decision tree moves through its branches by comparing the features, eventually reaching a decision at the leaves.

Logistic Regression (LR)

The logistic regression (LR) model is one of the most used models when the response variable is dichotomous and categorical; this model yields efficient results. In formal terms, binary logistic regression involves binary dependent variables with two classes represented by the letters "0" and "1". In comparison, the independent variables can be either continuous variables with any real value or binary variables with two classes represented by the letters "0" and "1." This model can widely used in the parameter estimation of the medical data sets which are based on binary response such as CVD and related[6-18].

Mathematically, the model can be represented as

$$p = P(Y = 1|X = x) = \frac{e^{\beta_0 + B_1 x}}{1 + e^{\beta_0 + B_1 x}} \qquad (1)$$

Additionally, the LR model may be utilized for classification in the machine learning (ML) process[38]. This study used the LR model to classify the problem and satisfy the respondents with cardiovascular disease. By classifying observations into distinct groups and using logistic regression, this model is based on probability. This model utilizes the exponential logit function for the output transformation. The logistic regression hypothesis often limits the cost function to 0 to 1.

In this work, we use the function below to classify and predict the CVD patients in the machine learning logistic regression term.

$$f(x) = \begin{cases} 1, & \text{CVD Affected} \\ 0, & \text{CVD Not Affected} \end{cases} \qquad (2)$$

Support Vector Machine (SVM)

The support vector machine (SVM) is a popular tool for classification because of its ability to discriminate between distinct classes. Assuming binary classification for our response variable, cardiovascular disease (CVD), with a convention of linear separability for training samples, we formulate our dataset as follows:

$$S = \{(x_1, y_1), (x_2, y_2), \dots , (x_n, y_n) \quad (3)$$

Where $x_i \in R^d$, representing the design matrix X in the d-dimensional feature space, and y_i denotes the binary class label for CVD, with $y_i \in \{0,1\}$ in our study. The discriminative function is defined as:

$$f(x) = sgn(z, x) + \beta \quad (4)$$

Here, β are parameters to be learned.

RESULTS

In this study, a sample of 1800 respondents empirically analyzed whether CVD affected or not. **Table I** shows the descriptive statistics of the quantitative variables for the CVD respondents.

Table I: Descriptive Statistic of quantitative CVD respondents

Variables	Mean	Standard Deviation	Minimum	Maximum
Age	56.00	21.40	20.00	95.00
BMI	27.00	1.69	18.40	31.00
Systolic	153.00	13.57	109.00	189.00
Diastolic	84.00	3.32	66.00	89.00
Total Cholesterol	211.00	20.90	115.00	284.00

Table II comprises clinical risk factors statistically significant for CVD patients. This study set the level of significance value $\alpha = 0.05$ and applied the binary multiple logistic regression to the data set. The results found that the risk factors whose p-value is less than 0.05 are considered statistically significant risk factors for CVD patients.

The variables such as gender, age, Family history, body mass index (BMI), systolic, bad cholesterol, and other socio-demographic variables such as education, employment and other variables are responsible for developing cardiovascular disease (CVD) given in **Table II** below.

Table II: Parameter Estimation for the clinical risk factors of CVD-affected patients using multiple logistic regression (LR) model

Variables	B	S.E	Wald	Df	SIG	Exp$^{(B)}$	95% C.I. for EXP(B)	
							Lower	Upper
Gender (1)	-0.435	0.117	13.800	1	0.000	1.067	0.515	1.081
Education (1)	-0.151	0.120	1.585	1	0.208	0.859	0.679	1.088
Employment (1)	-0.031	0.124	0.064	1	0.801	0.969	0.760	1.236
Income (1)	0.394	0.119	11.009	1	0.001	1.482	1.175	1.870
FH(1)	-0.659	0.117	31.450	1	0.000	1.024	0.411	1.026
CKD (1)	-0.271	0.117	5.354	1	0.021	1.176	0.606	1.295
POP(1)	-0.944	0.491	3.688	1	0.055	0.389	0.149	1.020
TOP (1)	-0.280	0.143	3.864	1	0.049	0.756	0.571	0.999
Smoking(1)	-0.647	0.116	31.252	1	0.000	1.052	0.417	1.065
AI(1)	0.019	0.136	0.019	1	0.891	1.019	0.781	1.330
PA (1)	0.015	0.117	0.017	1	0.896	1.015	0.807	1.277
BMI	0.378	0.053	50.853	1	0.000	1.460	1.316	1.620
Age	0.129	0.055	5.450	1	0.020	1.138	1.021	1.269
SYS BP	0.660	0.122	29.066	1	0.000	1.935	1.522	2.460
DYS BP	0.039	0.126	0.096	1	0.757	1.040	0.812	1.332
TC	0.471	0.124	14.501	1	0.000	1.602	1.257	2.042
RBS	0.126	0.125	1.012	1	0.0314	1.134	0.888	1.448
Diabetic (1)	-0.134	0.125	1.152	1	0.0283	0.874	0.684	1.117
LDL	0.374	0.190	3.870	1	0.049	1.454	1.001	2.111
TG	-0.023	0.139	0.027	1	0.869	0.977	0.745	1.283
HDL	-0.012	0.122	0.009	1	0.924	0.988	0.779	1.255

DISCUSSION

To accomplish our task, this study uses the binary classifier based on a supervised machine learning algorithm, as suggested by[31] to predict the relation for the relevant class of patients. The results obtained from the predictive models utilized to forecast CVD are presented in **Table III**. DT, SVM, LR and RF are the four ML techniques used to construct the CVD prediction model in two phases. The model is trained using 75% of the data set in the first step, and the remaining 25% is utilized to validate the model in the second stage. According to the results in **Table III,** it is found that with a 95% confidence interval of (0.7224, 0.812), the random forest (RF) model results in a high accuracy of 77%, followed by decision tree (DT), which results in 76% of accuracy with the confidence interval of (0.728, 0.811) additionally the logistic regression (LR) achieves the accuracy of 76.% with the confidence interval of (0.713, 0.804) and in last the support vector machine (SVM) results in 76.% accuracy with the confidence interval of (0.716, 0.804). The results in **Table III** demonstrate that the RF method is the most efficient predictor for CVD patients. Moreover, it is noteworthy that our findings are consistent with the authors' conclusions[31-33].

Table III: Accuracy Outputs of Machine Learning Classifiers for CVD Respondents and MCE on 25% Testing Data Set

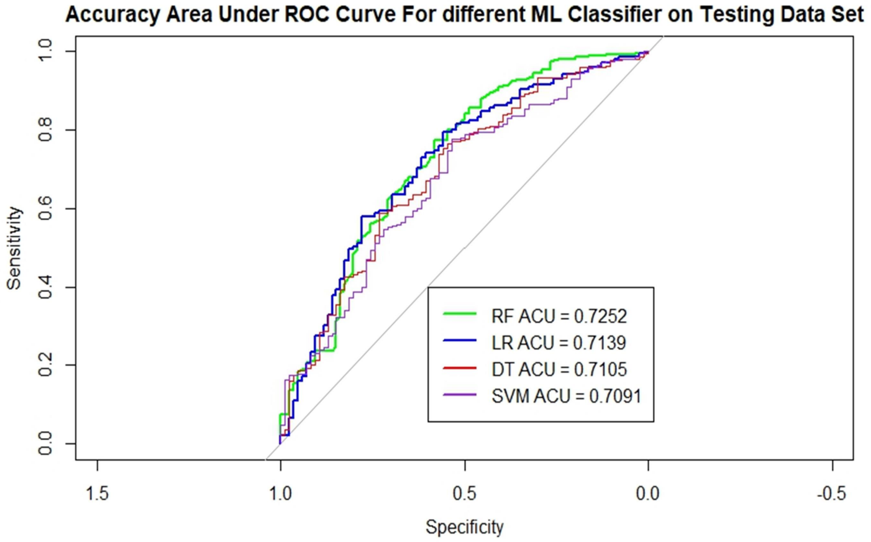
Output	RF	DT	LR	SVM
Accuracy	0.7751	0.7620	0.7608	0.7619
95% C. I	(0.722, 0.812)	(0.728, 0.811)	(0.713, 0.804)	(0.716, 0.804)
Sensitivity	0.9891	0.9599	0.9635	0.9891
Specificity	0.0930	0.1977	0.1395	0.0814
+Predicted value	0.7765	0.7922	0.7811	0.7743
−Predicted value	0.6071	0.789	0.5455	0.7000
F1 Score	0.8700	0.8680	0.8627	0.8686
MCE on 25% Testing Data	0.2166	0.2222	0.2333	0.2250

The RF algorithm calculated sensitivity and specificity to be 98.40% and 10%, respectively. That is to say, 98% of the patients were accurately identified by the algorithm used as having CVD in this study, whereas 10% were incorrect.

Results show that RF predicts a misclassification error rate of 0.2166 for the affected respondents, while the DT model predicts 0.2222, LR model predicts 0.2333, and SVM predicts 0.2250, respectively.

The ROC for each machine learning classifier is displayed in **Figure 2**. The ROC curve also shows that, out of all the machine learning algorithm classes, the performance of the random forest (RF) algorithm is the most efficient.

Figure II: ROC curve for 25% testing data set for Machine Learning Classifiers



For the random forest (RF) algorithm, the ROC value is 0.7252, which indicates that this machine learning predictive model precisely denotes efficient classification and prediction. Additionally, the ROC values for LR, DT and SVM are 0.7139, 0.7105 and 0.7091 respectively.

Cardiac diseases are a serious concern in healthcare data analysis. Healthcare professionals need to use the power of predictive machine learning algorithms to improve diagnoses and treatments for patients. This study examines how effective machine learning algorithms are in predicting heart disease (CVD) in patients. The study uses data exploration techniques to analyze trends in CVD cases and aims to identify key clinical risk factors for heart disease. The results show that age, gender, and other variables significantly predict CVD.

In the study, several machine learning techniques, such as Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM), were used to classify and predict heart disease. Among these, the Random Forest algorithm performed the best, with the highest prediction accuracy (77.5%), sensitivity (81%), and the best result on the receiver operating characteristic curve (77.5%). It also has the lowest misclassification error (21.66 %) for CVD predictions. This shows that RF is the most reliable model for predicting heart disease.

CONCLUSION

The study concludes that the suggested model helps forecast cardiac illness and can be adjusted to similar data in other domains. Further deep-learning research is advised for increased accuracy.

Ethical permission: University of Sindh, Jamshoro, IRB letter No. DRGS/215.
Conflict of interest: There is no conflict of interest between the authors.
Financial Disclosure / Grant Approval: No Funding agency was involved in the research
Data Sharing Statement: The corresponding author can provide the data proving the findings of this study on request. Privacy or ethical restrictions bound us from sharing the data publicly.

AUTHOR CONTRIBUTION

Ishaq K: Conceived and designed the experiments, performed the experiments, Wrote the paper, Analyzed and interpreted the data, gathered data and materials;
Gill NP: Supervised the whole paper, Analyzed and interpreted the data
Ilyas R: Co-supervise the entire paper, Analyzed and interpreted the data

REFERENCES

1. Mbassi LS, Djantou EB, Nguimbou RM, Dicko A, Njintang NY. Improvement of Phenolic Compounds and Antihyperlipidemic Activity of Hibiscus sabdariffa L. Calyxes Powder Using CDS Processing. Am J Life Sci. 2023; 10(3): 24-35. doi: 10.11648/j.ajls.20231102.13.
2. Lin XN, Zeng YJ, Cao S, Jing XB. A real-world pharmacovigilance study of cardiac adverse events induced by sugammadex in the FDA adverse event reporting system. Expert Opinion on Drug Safety. 2024: 1-9.
3. Ramotowski B, Undas A, Budaj A. Altered platelet reactivity, coagulation, endothelial and inflammatory markers early after smoking cessation verified with cotinine plasma concentration. J Thrombosis Thrombolysis. 2023; 56(1): 75-81.
4. Jiang Q, Zhang Q, Wang T, You Q, Liu C, Cao S. Prevalence and risk factors of hypertension among college freshmen in China. Scientific Reports. 2021; 11(1): 23075. doi: 10.1038/s41598-021-02578-4.
5. Gluvic ZM, Zafirovic SS, Obradovic MM, Sudar-Milovanovic EM, Rizzo M, Isenovic ER. Hypothyroidism and risk of cardiovascular disease. Curr Pharmaceut Design. 2022; 28(25): 2065-72. doi: 10.2174/1381612828666220620160516.
6. Murabito JM, Namara PM, Hubert HB, Castelli WP. Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. JAMA. 2005; 294(24): 3117-23. doi: 10.1001/jama.294.24.3117.
7. Wang L, Ardern CI, Chen D. Geographic variation in cardiovascular disease mortality: A study of linking risk factors and built environment at a local health unit in Canada. Geospatial Technologies for Urban Health. 2020: 31-51. doi: 10.1007/978-3-030-19573-1_3.
8. Powell-Wiley TM, Poirier P, Burke LE, Després JP, Gordon-Larsen P, Lavie CJ et al. Obesity and cardiovascular disease: a scientific statement from the American Heart Association. Circulation. 2021; 143(21): e984-1010. doi: 10.1161/CIR.0000000000000973.
9. Nelson RH. Hyperlipidemia as a risk factor for cardiovascular disease. Prim Care Clin Off Pract. 2013; 40(1): 195-211. doi: 10.1016/j.pop.2012.11.003.
10. Mukherji D, Padalia N, Naidu A. A heart disease prediction model using SVM-decision trees-logistic regression (SDL). Int J Comput Appl. 2013; 68(16): 11–5. doi: 10.5120/11662-7250.
11. Yaqub FN, Khan MS. Chronic kidney disease: A statistical analysis. Proc. 17th International Conference on Statistical Sciences; 2019; 33: 385–90.
12. Iqbal R, Zuberi NA, Kamal A, Ahmad S, Ahmed N. A statistical analysis of hypertension as cardiovascular risk factor. Middle East J Sci Res. 2012; 12(1): 19-22. doi: 10.5829/idosi.mejsr.2012. 12.1.1649.
13. Khan MZ, Pervaiz MK, Javed I. Biostatistical study of clinical risk factors of myocardial infarction: A case-control study from Pakistan. Pak Armed Forces Med J. 2022; 66(3): 354-60.
14. Ayub M, Arsalan A, Bajwa S, Hussain F, Umar M, Khizar B et al. Self-reported health and smoking status, and body mass index: a case-control comparison based on GEN SCRIP (GENetics of SChizophReniaIn Pakistan) data. BMJ Open. 2021; 11(4): e042331.
15. Sajid MR, Muhammad N, Zakaria R, Shahbaz A, Nauman A. Associated factors of cardiovascular diseases in Pakistan: Assessment of path analyses using warp partial least squares estimation. Pakistan J Stat Oper Res. 2020; 16(2): 265-77. doi: 10.18187/PJSOR.V16I2.3075.
16. Nusinovici S, Tham YC, Chak Yan MY, Wei Han JP, Wong TY, Cheng CY. Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol. 2020; 122: 56–69. doi: 10.1016/j.jclinepi.2020.03.002.
17. Muhammad S, Iqbal R, Saad M, Khan FA. Relationship of lifestyle and dietary habits of South-East Asian (Pakistani) population with cardiovascular diseases: A case-control study. Pak Heart J. 2022; 55(4): 396-403.

18. Zulfiqar N, Rehman S. Risk factors of cardiac diseases in Pakistan. Proc. 17th International Conference on Statistical Sciences; 2019; 33: 373–84.
19. Koene RJ, Prizment AE, Blaes A, Konety SH. Shared risk factors in cardiovascular disease and cancer. Circulation. 2016; 133(11): 1104-14. doi: 10.1161/CIRCULATIONAHA.115.020406.
20. Ajeganova S, Hafström I, Frostegård J. Patients with SLE have higher risk of cardiovascular events and mortality in comparison with controls with the same levels of traditional risk factors and intima-media measures, which is related to accumulated disease damage and antiphospholipid syndrome: a case–control study over 10 years. Lupus Sci Med. 2021; 8(1): e000454.
21. Van Bussel EF, Richard E, Arts DL, Moll Van Charante EP. Predictive value of traditional risk factors for cardiovascular disease in older people: A systematic review. Prev Med (Baltimore). 2020; 132: 105986. doi: 10.1016/j.ypmed.2020.105986.
22. Al-Jafar R, Zografou-Themeli M, Zaman S, Akbar S, Lhoste V, Khamliche A et al. Effect of religious fasting in Ramadan on blood pressure: results from LORANS (London Ramadan Study) and a meta-analysis. J Am Heart Assoc. 202; 10(20): e021560.
23. Aburto NJ, Hanson S, Gutierrez H, Hooper L, Elliott P, Cappuccio FP. Effect of increased potassium intake on cardiovascular risk factors and disease: Systematic review and meta-analyses. BMJ. 2013; 346: f1378. doi: 10.1136/bmj.f1378.
24. Niedhammer I, Bertrais S, Witt K. Psychosocial work exposures and health outcomes: a meta-review of 72 literature reviews with meta-analysis. Scand J Work Environ Health. 2021; 47(7): 489.
25. Malambo P, Kengne AP, De Villiers A, Lambert EV, Puoane T. Built environment, selected risk factors and major cardiovascular disease outcomes: A systematic review. PLoS One. 2016; 11(11): e0166846. doi: 10.1371/journal.pone.0166846.
26. Huang YQ, Liu L, Huang C, Yu YL, Lo K, Huang JY et al. Impacts of pre-diabetes or prehypertension on subsequent occurrence of cardiovascular and all-cause mortality among population without cardiovascular diseases. Diabet Metabol Syndr Obes. 2020; 1743-52.
27. Nag T, Ghosh A. Cardiovascular disease risk factors in Asian Indian population: A systematic review. J Cardiovasc Dis Res. 2013; 4(4): 222-8. doi: 10.1016/j.jcdr.2014.01.004.
28. Akioyamen LE, Genest J, Chu A, Inibhunu H, Ko DT, Tu JV. Risk factors for cardiovascular disease in heterozygous familial hypercholesterolemia: A systematic review and meta-analysis. J Clin Lipidol. 2019; 13(1): 15-30. doi: 10.1016/j.jacl.2018.10.012.
29. Worrall-Carter L, Ski CF, Scruth E, Campbell M, Page K. Systematic review of cardiovascular disease in women: Assessing the risk. Nurs Health Sci. 2011; 13(4): 529-35. doi: 10.1111/j.1442-2018.2011.00645.x.
30. Gasevic D, Ross ES, Lear SA. Ethnic differences in cardiovascular disease risk factors: A systematic review of North American evidence. Can J Cardiol. 2015; 31(9): 1169-79. doi: 10.1016/j.cjca.2015.06.017.
31. Ramesh TR, Lilhore UK, Poongodi. Predictive analysis of heart diseases with machine learning approaches. Malays J Comput Sci. 2022; 1: 132-48.
32. Reddy KH, Saranya G. Prediction of cardiovascular diseases in diabetic patients using machine learning techniques. Lect Notes Networks Syst. 2021; 130(2): 299-305. doi: 10.1007/978-981-15-5329-5_28.
33. Khan A, Qureshi M, Daniyal M, Tawiah K. A novel study on machine learning algorithm-based cardiovascular disease prediction. Health Soc Care Commun. 2023; 2023: 1-10. doi: 10.1155/2023/1406060.
34. Hussain I, Qureshi M, Ismail M, Iftikhar H, Zywiołek J, López-Gonzales JL. Optimal features selection in the high dimensional data based on robust technique: Application to different health database. Heliyon. 2024; 10(17).
35. Ho TK. Random decision forests. Proc 3rd Int Conf Document Analysis and Recognition. 1995; 278-82.

36. Iftikhar H, Khan M, Khan Z, Khan F, Alshanbari HM, Ahmad Z. A comparative analysis of machine learning models: a case study in predicting chronic kidney disease. Sustainability. 2023; 15(3): 2754.
37. Liu Y, Wang Y, Zhang J. New machine learning algorithm: random forest. Lect Notes Comput Sci. 2012; 246-52.
38. Alshanbari HM, Iftikhar H, Khan F, Rind M, Ahmad Z, El-Bagoury AA. On the implementation of the artificial neural network approach for forecasting different healthcare events. Diagnostics. 2023; 13(7): 1310.
39. Daniël L. Sample size justification. Collabra Psychol. 2022; 8(1).
40. Hussain I, Qureshi M, Ismail M, Iftikhar H, Zywiołek J, López-Gonzales JL. Optimal features selection in the high dimensional data based on robust technique: Application to different health database. Heliyon. 2024; 10(17).
41. Cuba WM, Huaman Alfaro JC, Iftikhar H, López-Gonzales JL. Modeling and analysis of monkeypox outbreak using a new time series ensemble technique. Axioms. 2024; 13(8): 554.
42. Iftikhar H, Khan M, Khan MS, Khan M. Short-term forecasting of monkeypox cases using a novel filtering and combining technique. Diagnostics. 2023; 13(11): 1923.
43. Daniyal M, Qureshi M, Marzo RR, Aljuaid M, Shahid D. Exploring clinical specialists' perspectives on the future role of AI: evaluating replacement perceptions, benefits, and drawbacks. BMC Health Serv Res. 2024; 24(1): 587.