Kurdish Studies
August 2024
Volume: 12, No: 5, pp 818-822
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.53555/ks.v12i5.3345

Prediction of Cardiovascular Diseases Through Machine Learning Algorithms: A Supervised Model

Naveed Sheikh[1*], Asma Naeem[1], Sajida Parveen[1], Abdul Rehman[1], Misbah Anjum[1], Muhammad Yasin[2], Raheela Manzoor[3]

[1]Department of Mathematics, University of Balochistan, Quetta, Pakistan.
[2]Department of Mathematics, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan.
[3]Department of Mathematics, Sardar Bahadur Khan (SBK) Women's University, Quetta, Pakistan

***Corresponding Author:** Naveed Sheikh
*Department of Mathematics, University of Balochistan, Quetta, Pakistan. Tel: +923358396878, Email: naveed.maths@um.uob.edu.pk

Abstract
Heart disease is the leading cause of mortality worldwide, ranking as the number one killer of humans. Machine learning (ML) algorithms are not just employed to detect the presence or absence of cardiac disease. However, it is also useful in forecasting the various stages of cardiovascular disease, beginning with stage 1 and progressing to stages 2, 3, and 4 (severe heart disease), respectively. In the current study, three supervised machine algorithms are employed to determine which approach has the greatest accuracy. Models such as Logistic Regression (LR), Super Vector Machine (SVM), and Random Forest (RF) are employed with hyper parameters to optimize classifier performance and determine which one is best for detecting the stage at which the individual is suspected of having disease. The experimental findings suggest that Logistic Regression (LR) has higher Accuracy, Precision, Recall and F-Measure i.e. 82%, 91%, 80%, 85% respectively.

Keywords Classification, Logistic Regression (LR), Machine Learning (ML), Support Vector Machine (SVM)

Introduction

Heart disease is a type of disease in which there is a blockage in the coronary arteries, which are blood vessels that carry blood and oxygen to the heart (Hassan et al. 2022). The coronary heart disease is caused by the development of fatty material and plaque inside the coronary arteries. As deadly as it sounds, heart diseases are not easy to identify in their preliminary stages of development. Researchers have been constantly attempting to identify better detecting techniques that could timely identify the heart disease in a person as the current techniques are not as effective in detecting this disease in the early development stages as one would hope for because of accuracy and computational time (Saboor et al. 2022). Additionally, when professional health experts and advanced technology is not readily available, it can become tremendously challenging to detect any symptoms of a heart disease until the person starts experiencing chest pain or complains about trouble when breathing, by which time, it might be very difficult to cure this disease and save that person's life (Almujally et al. 2023). To prevent an individual from going through this hardship, and to help doctors detect this disease in the preliminary stages, it will be important to have a tool that detects such a life-threatening disease early on (Elbagoury et al. 2023).
Machine Learning (ML) is widely used in healthcare decision support systems to aid clinical diagnoses and disease predictions (Saeed et al. 2018). Many areas of health informatics have recently benefited from the application of data mining and ML techniques (Sherazi et al. 2023). To address the shortcomings of traditional methods based on invasive detection of heart diseases (HD), researchers have attempted to create a non-invasive intelligent health care system based on predictive ML technologies such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), Random Forest (RF), and Decision Tree (DT) (Hussain et al. 2023). Cleveland heart disease datasets are widely used by researchers in the literature. Different results have been obtained from various methods by applying data mining algorithms for diagnosis (Saleem et al. 2023).
(Revathi and Jeevitha 2013) proposed a model for prediction of cardiovascular disease using machine learning algorithm hybrid random forest with linear mode. It was found that the prediction accuracy of 88.7%. The dataset was collected from UCI repository site. Using Cleveland dataset. On the other hand, KNN decision tree, linear regression, support vector machine algorithms were used for prediction of heart disease and their accuracy were compared. Results showed that the best accuracy of 87% were obtained by using k-nearest neighbor algorithm followed by support vector machine 83%, decision tree 79% and linear regression of 78% accuracy among all these algorithms for prediction of heart disease (Soni et al., 2020). A study was conducted to compare statistical, ML and data mining methods in terms of their ability to assist in predicting heart failure risks. The researchers compared the performance of statistical evaluation, Decision Trees, Random Forest, and convolutional neural networks, and they obtained prediction accuracy results of 85%, 80.1%, 85.38%, and 93%, respectively (Saeed et al. 2018) . Furthermore, different varieties of unsupervised clustering algorithms were used to determine their accuracy in terms of cardiac

disease search and diagnosis. The algorithms were applied to the Cleveland dataset. The study results highlighted k-means as the most appropriate algorithm for cardiac disease diagnosis (Kodati, Vivekanandam, and Ravi 2019). A model was suggested using ensemble approaches (boosting and bagging) with feature extraction algorithms (LDA and PCA) for predicting heart disease. The authors compared ensemble techniques (bagging and boosting) with five classifiers (SVM, KNN, RF, NB, and DT) on selected features from the Cleveland heart disease dataset. The results of the experiments indicated that the bagging ensemble learning method with DT and PCA feature extraction obtained the most outstanding performance (Gao et al. 2021). (Rani et al. 2021) designed a novel hybrid decision support system to diagnose cardiac ailments early. They effectively addressed the missing data challenge by employing multivariate imputations through chained equations. Additionally, their unique approach to feature selection involved a fusion of genetic algorithms (GA) and recursive feature reduction. Notably, the integration of random forest classifiers played a pivotal role in significantly enhancing the accuracy of their system. However, despite these advancements, their hybrid approach's complexity might have posed challenges in terms of interpretability and practical implementation. Machine learning techniques were embraced to forecast cardiac diseases. A hybrid model by incorporating random forest as the base classifier was introduced. This hybridization aimed to enhance prediction accuracy; however, the decision to capture and store user input parameters for future use was intriguing but yielded suboptimal classification performance. This unique approach could be viewed as an innovative attempt to integrate patient-specific information, yet the exact impact on overall performance warrants further investigation (S et al. 2021). Further advanced the in the field were obtained by employing a hybrid model combined with random forest with a linear model to predict cardiovascular diseases. Through this amalgamation of different classification approaches and feature combinations, resulting in a commendable performance with an accuracy of 88.7%. However, it is worth noting that while hybrid models show promise, the trade-offs between complexity and interpretability could influence their practical utility in real-world clinical settings (Prabhakaran et al. 2022).

In the present study, it has been suggested a multiple-stage cardiovascular disease diagnosis model based on three different algorithms while evaluating their performance. The suggested technique has been developed by combining three ensemble learners: Random Forest (RF), Logic Regression (LR), and Support Vector Machine (SVM).

BACKGROUND OF CLEVELAND DATASET

Experiments with the Cleveland dataset have concentrated on simply attempting to distinguish thepresence of heart disease from absence. The 14 attributes that were used are listed in Table 1.

Table 1 Cleveland dataset attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female, 1 = male
Cp	Discrete	Chest pain: 1 = typical angina, 2 = a typical angina, 3 = non-angina pain, 4 =a symptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1 = true, 0 = false
Restecg	Discrete	Resting Electrocardiograph
Thalach	Continuous	Exercise Max Heart Rate Achieved
Exang	Discrete	Exercise Induced Angina: 1=yes, 0=no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1=up sloping, 2=flat, 3=down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that rangebetween 0 and 3
Tha	Discrete	3=normal, 6=fixed defect, 7=reversible defect
Class	Discrete	Diagnosis classes: 0=No Presence, 1=Least likely to have heart disease 2=>1 3=>2 4=More likely have heart disease

The five stages of our prediction of heart disease presented are mapped in Table 2.

Table 2 Five stages prediction

0	No Heart Disease
1	Stage1
2	Stage2
3	Stage3
4	Heart Disease presented

Recommendation of Model Algorithms

A study conducted by Randal S. Olson provides insightful best practice advice for solving bioinformatics problems with Machine Learning (ML), "Data-driven Advice for Applying Machine Learning to Bioinformatics Problems". It has been

analyzed 13 state-of-the-art commonly used machine learning algorithms on a set of 165 publicly available classification problems to provide data-driven algorithm recommendations to current researchers.
From his findings, he was able to provide a recommendation of three algorithms with hyperparameters that maximize classifier performance across the tested problems, as well as general guidelines for applying machine learning to supervised classification problems. The recommendations are shown inn Table 3.

Table 3 Three Machine Learning Algorithms

Algorithm	Parameters	DatasetsCovered
Random Forest (RF) Classifier	n_estimators=500, max_features = 0.25, criterion=entropy	19
Support Vector Machine (SVM)	C=0.01, gamma=0.1, degree=3, coef0=10.0	16
Logistic Regression (LR)	C=1.5, Penalty=L1, Fit_intercept=true	8

The database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by DL researchersto this date. The "num" field in the figure refers to the presence of heart disease in the patient. It is integer valued from zero (no presence) to four. Experiments with the Cleveland database have concentrated on attempting to distinguish presence (values 1,2,3,4) from absence (value 0) (Smiley S., 2011).

Methodology
The proposed methodology using three classification techniques; Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (L R) to predict heart disease as the proposed methodology shown in Fig1. These classifiers are used to improve the prediction. The classifiers have been applied in to heart disease data that comes from the Cleveland dataset to predict which of five stages a patient has heart problems. The performance of these classifiers is evaluated on the bases of accuracy, precision recall, and F-Measure.

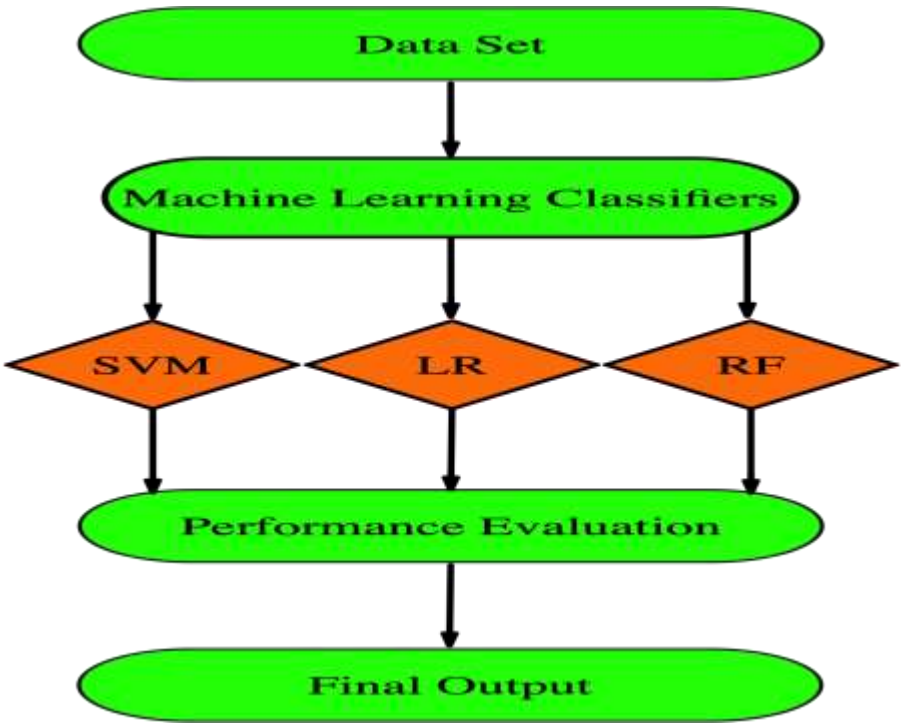

Figure 1 Proposed Methodology

The dataset of heart is taken from UCI repository, the classifier taking it as input for disease prediction. These classifiers are implemented in Python language. Python is a powerful interpreter language and a reliable platform for research. The accuracy of prediction increased by comparing the results of these three classifiers using evaluation parameters. The experimental result describes which classifier is best between them.

Evaluation Parameter
Some evaluation parameters in data mining are accuracy, precision, recall, and F-Measure. WhereTP- True Positive, TN- True Negative, FP- False Positive and FN- False Negative.

- Accuracy is defined as the number of accurately classified instances divided by the totalnumber of instances in the dataset as in equation (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Precision is the average probability of relevant retrieval as described in equation (2).

$$Precision = \frac{TP}{TP+Fp} \quad (2)$$

- The recall is defined as the average probability of complete retrieval as defined in equation (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- F- Measure is the calculated by using both precision and recall as shown in equation (4).

$$F - Measure = \frac{2 * (Precision * Recall)}{Precision+Recall} \quad (4)$$

Dataset

To perform the research, heart disease dataset is used. This heart disease dataset contains 14 attributes and 303 instances. This dataset is taken from UCL repository. It's an online repository that contains 412 diverse datasets. UCI provides data for ML to perform analysis in a different direction. The UCI database is known for its extensiveness in data, its completeness, and accuracy.

In this research, three classification methods are implemented in python using the pandas and keras libraries. These models are used to improve prediction. These classifiers are compared to find out which of the five stages best predicts the chance of heart disease in patients. In the next section, we briefly describe these classification techniques/ classifiers.

i. **Logistics Regression (LR)** is the predictive analysis to conduct on discrete (binary) values based on a specified set of independent variables. LR describes the data and clarifies the relationship between one (binary) dependent variable and independent variables. It predicts the event occurrence probability by fitting the data into a logits function. Therefore, it is also called logistic regression. Input values x is linearly combined using coefficient values β, to calculate an output value P. The output values as predictable lies between 0 and 1. Input data associated with coefficient β (constant value) learned from training data. Where p is the output, b0 is an intercept term and b1 is the coefficient of input value x as shown in equation (5).

$$P = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (5)$$

ii. **Support Vectors Machine (SVM)** is a classification and regression algorithm. In SVM, every data item is plotted in n-dimensional space, a few dimensions are equivalent tothe number of features or attributes. Where n represents the number of attributes. The valueof each attribute being the value of certain coordinates. Once plotting all the data items then performed classification by drawing a line or by finding the optimal hyperplane that separates two classes completely. For example, if we have two features of individual-like hair and height length. First, we plot these two features in two-dimensional space where every point has two coordinates (these co-ordinates are also known as Support Vectors).

iii. **Random Forests (RF)** are ensemble learning technique for regression, regression, classification, and for other tasks. That operate by making a multitude of Decision Tree (DT) at training stint and outputting that is the mean prediction (regression) or mode of classes (classification) of the distinct trees.

Scaling the Data

To accomplish the five stages output prediction for a patient to be diagnosed with one of five stages, it is important to scale the data, so the machine learning algorithms do not overfit to the wrong features. Using the Min Max Scaler () method on Python, the values are scaled per features based on the minimum and maximum between 0 and 1. This keeps the information from being lost but allows the machine learning algorithms to correctly train with the data. The training data and test data are scaled between 0 and 1 and the output data is scaled between 0 and 1 as well. Then, the scaled output value is mapped as shown in Table 4.

Table 4: Five Stages of training data and test data

Output Values	Stages
0	No disease presented
0 < and <= 0.25	Stage1
0.25 < and <= 0.5	Stage2
0.5 < and <= 0.75	Stage3
0.75 < and <= 1	Advance disease presented

Where all evaluation parameters accuracy, precision, recall, and F-Measure are calculated from dataset when splitting the dataset into training data and test data. The Pseudo codes for the evaluation parameters are as follow:

Experimental Results

The experiment is conducted for the prediction of heart disease stages by applying various machine learning classifiers. From the experiment results, it has been identified that Logistic Regression (LR) performs better as compared to the other four ML classifiers in the prediction of these diseases as shown in figure 2,3,4 and 5. The comparison between the classifiers has been shown in Table 5. In this experiment, there has been used multiple stages of heart disease prediction to forecast the stage at

which a person is determined to have heart disease. The study used two outcome predications, either a person has the disease or not; that is represented by (0 ,1) or (true, false).

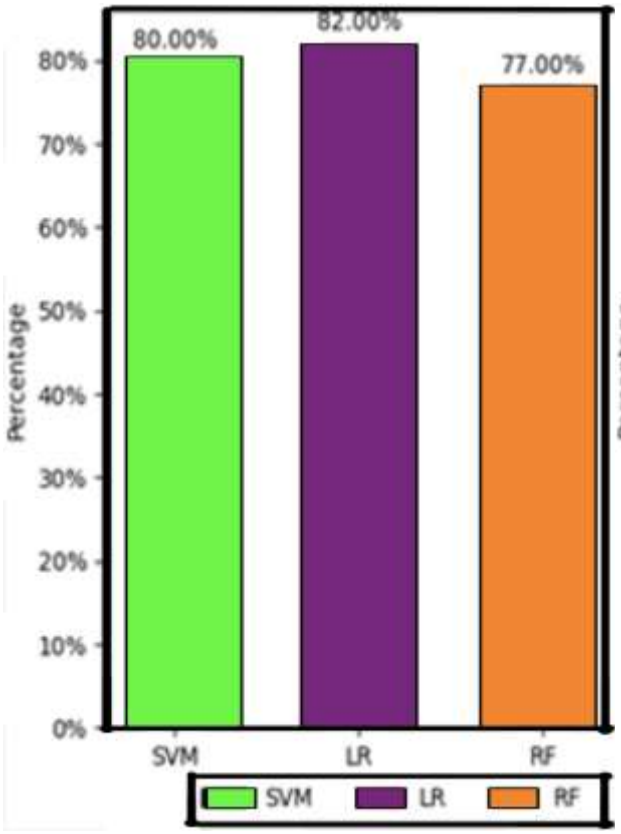

Figure 2 Heart Disease Accuracy

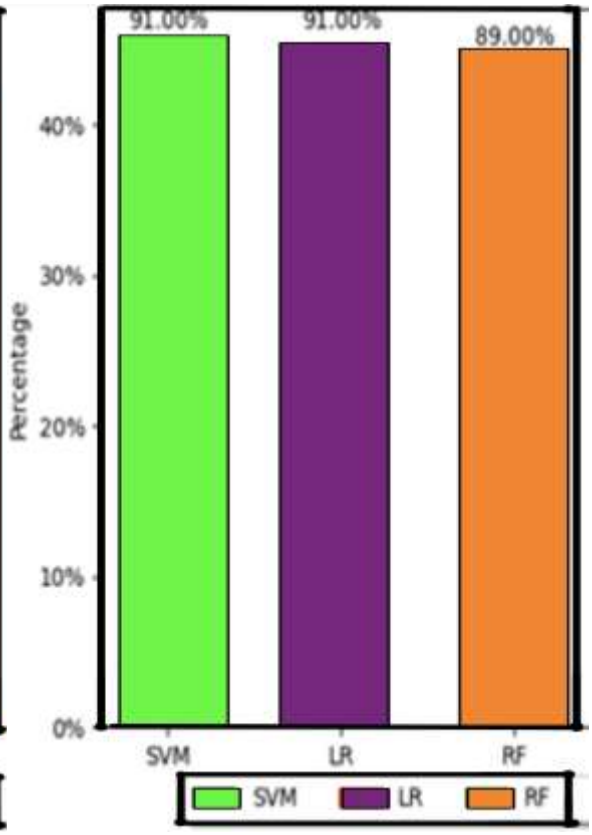

Figure 3 Heart Disease Precision

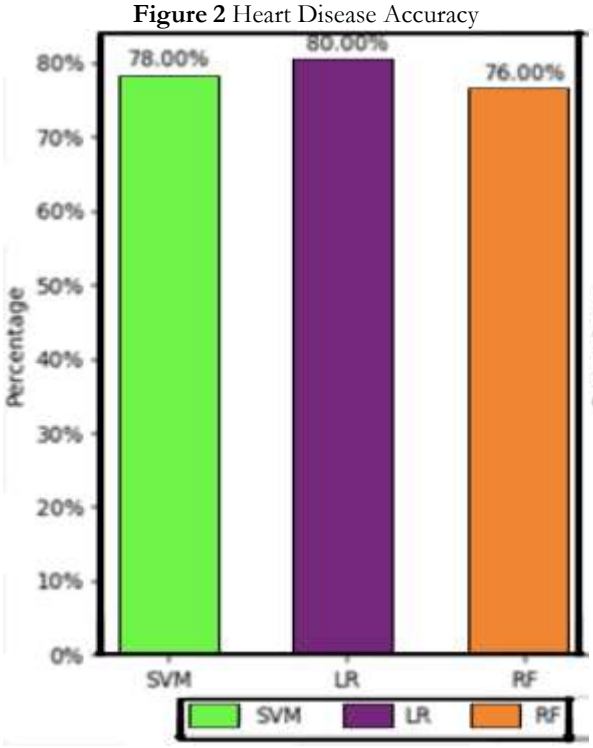

Figure 4 Heart Disease Recall

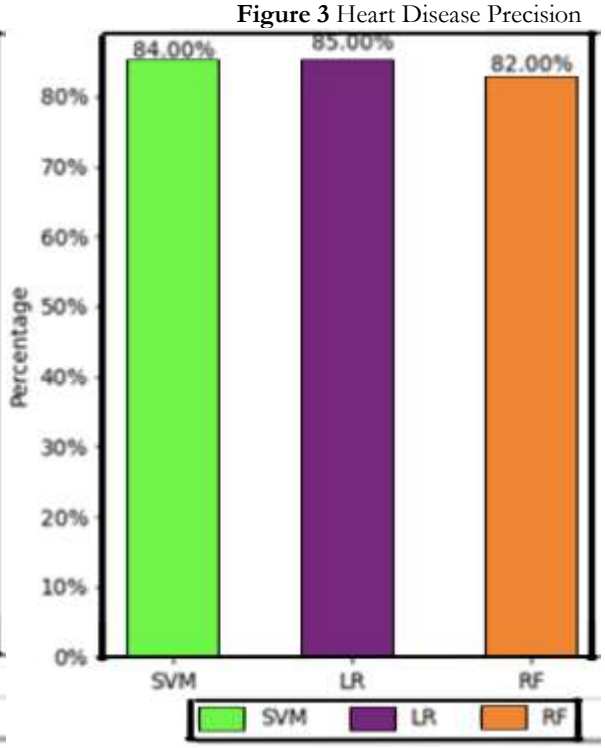

Figure 5 Heart Disease F-Measure

Table 5 Machine Learning Algorithms Comparison

Algorithms	Accuracy	Precision	Recall	F-Measure
SVM	80%	91%	78%	84%
LR	82%	91%	80%	85%
RF	77%	89%	76%	82%

Conclusion

The importance of extracting the valuable information from raw data has very good consequencesin many fields of life such as the medical area, business area, and more. In this study, it has been proposed a multiple stage detection model of heart disease based on three algorithms to compare which one performs better. The proposed method was built by stacking three different ensemble learners, Random Forest, Logic Regression, and Support Vector Machine. The proposed detection model was testedon a well-known Cleveland dataset to provide a fair benchmark against existing studies. Based on the experimental results, The proposed model was able to outperform heart disease detection methods with respect to accuracy, precision, recall and F-Measure. The experimental findings suggest that Logistic Regression (LR) has higher Accuracy, Precision, Recall and F-Measure i.e. 82%, 91%, 80%, 85% respectively.

References

1. Almujally, Nouf Abdullah, Turki Aljrees, Oumaima Saidani, Muhammad Umer, Zaid Bin Faheem, Nihal Abuzinadah, Khaled Alnowaiser, and Imran Ashraf. 2023. "Monitoring Acute Heart Failure Patients Using Internet-of-Things-Based Smart Monitoring System." *Sensors* 23 (10). https://doi.org/10.3390/s23104580.
2. Elbagoury, Bassant M., Luige Vladareanu, Victor Vlădăreanu, Abdel Badeeh Salem, Ana Maria Travediu, and Mohamed Ismail Roushdy. 2023. "A Hybrid Stacked CNN and Residual Feedback GMDH-LSTM Deep Learning Model for Stroke Prediction Applied on Mobile AI Smart Hospital Platform." *Sensors* 23 (7). https://doi.org/10.3390/s23073500.
3. Gao, Xiao Yan, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar. 2021. "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method." *Complexity* 2021. https://doi.org/10.1155/2021/6663455.
4. Hassan, Ch Anwar ul, Jawaid Iqbal, Rizwana Irfan, Saddam Hussain, Abeer D. Algarni, Syed Sabir Hussain Bukhari, Nazik Alturki, and Syed Sajid Ullah. 2022. "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers." *Sensors* 22 (19). https://doi.org/10.3390/s22197227.
5. Hussain, Shakira, Naveed Sheikh, Misbah Anjum, Arbab Ghulam Raza, and Rabail Rizvi. 2023. "Mathematical Modelling of COVID-19 Pandemic in Pakistan with Optimal Control." *Journal of Asian Scientific Research* 13 (1): 28–44. https://doi.org/10.55493/5003.v13i1.4721.
6. Kodati, Sarangam, R. Vivekanandam, and G. Ravi. 2019. "Comparative Analysis of Clustering Algorithms with Heart Disease Datasets Using Data Mining Weka Tool." In *Advances in Intelligent Systems and Computing*, 900:111–17. Springer Verlag. https://doi.org/10.1007/978-981-13-3600-3_11.
7. Prabhakaran, Dorairaj, Kavita Singh, Dimple Kondal, Lana Raspail, Bishav Mohan, Toru Kato, Nizal Sarrafzadegan, et al. 2022. "Cardiovascular Risk Factors and Clinical Outcomes among Patients Hospitalized with COVID-19: Findings from the World Heart Federation COVID-19 Study." *Global Heart* 17 (1). https://doi.org/10.5334/GH.1128.
8. Rani, Pooja, Rajneesh Kumar, Nada M.O.Sid Ahmed, and Anurag Jain. 2021. "A Decision Support System for Heart Disease Prediction Based upon Machine Learning." *Journal of Reliable Intelligent Environments* 7 (3): 263–75. https://doi.org/10.1007/s40860-021-00133-6.
9. Revathi, T, and S Jeevitha. 2013. "Comparative Study on Heart Disease Prediction System Using Data Mining Techniques." *International Journal of Science and Research*. Vol. 4. www.ijsr.net.
10. S, Kavitha B, M Siddappa, Engineering Tumakuru, Karnataka Tumakuru, and Karnataka ----------. 2021. "Machine Learning Classifying Approach for Identifying Heart Disease in Medical Field." *International Research Journal of Engineering and Technology*. www.irjet.net.
11. Saboor, Abdul, Muhammad Usman, Sikandar Ali, Ali Samad, Muhmmad Faisal Abrar, and Najeeb Ullah. 2022. "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms." *Mobile Information Systems* 2022. https://doi.org/10.1155/2022/1410169.
12. Saeed, Saeeda, Junaid Baber, Maheen Bakhtyar, Ihsan Ullah, Naveed Sheikh, Imam Dad, and Anwar Ali Sanjrani. 2018. "Empirical Evaluation of SVM for Facial Expression Recognition." *IJACSA) International Journal of Advanced Computer Science and Applications*. Vol. 9. www.ijacsa.thesai.org.
13. Saleem, Muhammad, Naveed Sheikh, Abdul Rehman, Muhammad Rafiq, and Shah Jahan. 2023. "Real-Time Object Identification Through Convolution Neural Network Based on YOLO Algorithm." *Mathematics and Computer Science*, December. https://doi.org/10.11648/j.mcs.20230805.11.
14. Sherazi, Kalsoom, Naveed Sheikh, Misbah Anjum, and Arbab Ghulam Raza. 2023. "Solar Drying Experimental Research and Mathematical Modelling of Wild Mint and Peach Moisture Content." *Journal of Asian Scientific Research* 13 (2): 94–107. https://doi.org/10.55493/5003.v13i2.4814.
15. Soni, S. K., Madan Mohan Malaviya University of Technology, North Dakota State University, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, IEEE Industry Applications Society, and Institute of Electrical and Electronics Engineers. n.d. *ICE3-2020 : International Conference on Electrical and Electronics Engineering : February 14-15, 2020*.