JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Heart disease Prediction – A Performance analysis of Machine Learning Classifiers

Kumari Anjali Suman[1] and Vandana Bhattacharjee[2]
Birla Institute of Technology, Mesra
Department of Computer Science and Engineering

ABSTRACT

Heart related disease or cardiovascular disease are becoming the major cause for death in worldwide. The major factors that contribute to it are smoking, drinking, diabetes and many more. With the increase of cases, prognosis of this disease is much crucial in medical tasks which helps cardiologist to provide in time and proper treatment. Machine learning methods in medical group having rapid growth as they can recognize patterns, preprocess the data and also improve the percentage of accuracy of trained models. This paper aims to compare the machine learning classifier performances and based on that develop a model with higher accuracy in approach of heart disease prediction. We used five different machine learning methods to train predictive models such as Random Forest, Decision Tree, KNN, Support Vector Machine and ANN. Through a meticulous evaluation process, best classification accuracy, recall and F1 score was obtained from random forest algorithm and highest precision probability achieved from decision tree technique for correct diagnosis of system.

Keywords: Machine learning, Classifier, K-Nearest Neighbour, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, Confusion matrix, Performance metrics

INTRODUCTION

Cardiovascular disease or pertaining to heart disease is a primary and notable disease in medical organization which are causing death., not only in India but also globally. According to EMRI 108, 72573 cardiac emergencies was there in 2023 which is massive 35% higher compared to 2018. Heart disease is disease that affects working of heart such as coronary artery disease (blood vessels disease), arrhythmia (irregular heartbeat) and heart failure. When our heart doesn't work well then it may face problem in sending enough blood, oxygen and nutrients to our body that it can cause the problems like hypoxia, anemia and many more. The main thing that can put a person in cardiovascular disease are our gender, smoking habits, age, family history, poor diet, lipids, lack of physical activity, high blood pressure, weight gain, and drinking alcohol. Some of the technologies like Electrocardiogram (ECG), Echocardiogram(ultrasound), Magnetic resonance imaging (MRI), etc., helps to diagnose the disease related to cardio but it takes much time in result and quite expensive. Thus, expedite, reasonable and accurate predictions of heart related disease is vital important.

Machine learning in medical organization allow us to build models which can quickly clean and helps to gain useful insights from collected data. Many researchers designed and proposed models using different kinds of machine learning methods to diagnose heart disease. However, few papers were concentrated to synthesize the trained models into mobile app or internet so that user can monitor their health status smoothly. In my work, we aim to construct a model using five different machine learning algorithms, examine their performances using performance metrics and comparatively analyzed them. The dataset taken from Kaggle site that holds 303 cases with 14 physical components and circumstances such as fasting sugar level, age, blood pressure and others. To achieve the goal, we employed various technique namely, K Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Decision Tree and Artificial Neural Network (ANN) to build predictive models.

RELATED WORK

The applications of artificial intelligence and machine learning algorithm have gained much popularity in recent years as it can potentially aid to improve accuracy and efficiency to make prediction. These methods are widely used in medical applications and also a number of articles have been published in domain of predicting diseases. Some of them are as follows:

Raniya Rone Sarra (2023) published a paper for predicting a heart disease by using ANN, SVM and DNN with 93.44%, 86.88% and 87% accuracy respectively using 303 samples.

Fahad Mehfooz (2021) used Logistic regression with the accuracy of 88.52%, Gaussian Naive Bayes with the accuracy of 88.52%, Bernoulli naïve bayes with the accuracy of 86.88%, Support Vector Machine with the accuracy of 91.80%, random forest with the accuracy of 85.24% and X gradient boosting with accuracy of 80.32%. Two different values for K were taken in the KNN algorithm and accuracies were calculated. Hence, it was concluded that SVM is performing the best for prediction of heart attack dataset.

Singh et al., (2021) predicted breast cancer using K Nearest Neighbor. The dataset collected from WBC datasets (Wisconsin breast cancer) with 699 instances (Training Set with 468 data values and Testing Set with 231 data points). Dimension reduction techniques is also aimed to apply for classification models.

Madhumita and Smita (2021) experimented a data mining model for predicting heart disease. The data was taken from Kaggle site (303 rows and 14 columns). For implementation of dataset, python programming was used in Jupyter notebook. The author used random forest data mining algorithm was implementation and 10-fold cross validation technique for splitting data. In proposed work, 86.9% classification accuracy is obtained for prediction of heart disease with diagnosis rate 93.3%.

In the research work of KM Jyoti Rani (2020) for diabetes prediction, K Nearest Neighbor, Logistic Regression, Decision Tree, Support Vector machine with several kernels is used and compared their accuracies of both training and testing set. Experimental results determine 99% of accuracy achieving in using decision tree algorithm.

Chintan et.al., (2023) aims to improve the classification accuracy of different machine learning classifiers such as decision tree, random forest, XGBoost, multilayer perceptron. This research develops a model that can correctly predict cardiovascular diseases to reduce the fatality caused by it. Using the dataset consisted of 70,000 rows and 12 attributes and cross validation approach, the study results the highest accuracy rate of 87.28% by using multilayer perceptron and lowest accuracy of

86.37% given by decision tree.

Asif et.al (2021) proposed a machine learning-based model for detection and classifying the COVID19 cases by examining the chest x-ray scans. The dataset is taken from Kaggle containing 550 total scans and splitting 75% as training and 25% as testing set. Two well-known machine learning classifiers support vector machine and random forest were used in this dataset. Support vector machine show an accuracy of 99.27% and random forest show an accuracy of 96.89%. The author also calculated precision, recall and compared accuracy of proposed model with their literature review models.

For the prediction of thyroid disease, Ankita et.al (2018) chosen four machine learning classifiers such as decision tree, ANN (Artificial Neural Network), support vector machine, K-NN (Nearest Neighbor). The author examined their accuracy and mean squared error of each algorithm also.

Chandan et.al., (2021) aims to discover a model for prediction of thyroid disease with higher accuracy. The dataset was taken from UCI (University of California Irvine machine learning repository) containing 215 samples and 5 features. K-Nearest Neighbor, support vector machine, artificial neural network, decision tree and logistic regression were taken and measured their accuracy. Higher accuracy obtained from logistic regression model (96.92%) and further concluded that this model will be considered for their prediction model.

OVERVIEW OF ALGORITHMS

A. K-Nearest neighbor

Most popular machine learning technique that can tackle both classification and regression tasks. It is a non-parametric method that gives an output similar to selected k value between data points. KNN didn't learn immediately from training set instead it stores the dataset and at the time of classification it works on it. So, it's also knowns as "lazy learner" and "Instance-based learner". Various distance metrices are used to calculate distance between query points and other data points such as Euclidean, Manhattan, Minokowski and Hamming distance. B. Support Vector Machine

Supervised max-margin models that analyze and solve classification, regression and outlier task. The fundamental principle of SVM is to create a hyperplane which can also referred as optimal boundary that maximize the distance between two margins and efficiently separate the classes. Depending upon nature of problem, SVM transforms the input features into high-dimensional space and implicitly map it for easy separable but it may lead to overfitting and curse of dimensionality. Mathematical technique i.e., kernel trick is used in this for handling non-linear relationship of data point and enables the computation of dot product. In other word, kernel function in SVM provides a window to manipulate the data. This technique enhances the versatility and applicability of SVM across a wide of problem without incurring cost and provide flexibility in modeling complex patterns. Different kernel functions are used in different problems such as linear, rbf, polynomial, sigmoid and gaussian.

C. Decision Tree

Specific type of flow chart like tree structure that are used in supervised learning approach. Most probably, it is preferred for solving classification problems. The decision tree is classified in three nodes i.e., the internal node that represents decisions on attributes containing one or more branches, the root node that represents initial decision as it is top most node and lastly, the leaf node that represents final decision with no further splits. Various types of decision tree such as CART (classification and Regression trees), ID3(Iterative Dichotomiser3), etc., are extensively used researchers and many others works like predicting models, calculating entropy, etc.

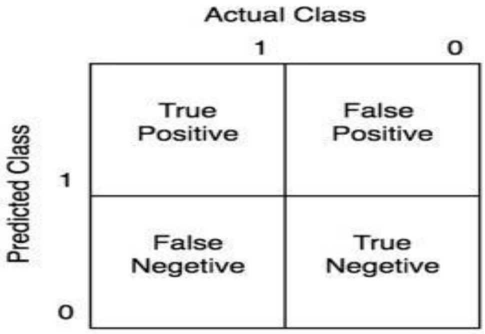
D. Random Forest

Ensemble learning model that belongs to supervised learning technique but more probably it is used in classification problem. In this, given dataset is further divided into different bootstrap sample(subdataset) and each sample having different data of given dataset means it focuses not to repeat the data. By using bootstrap samples, deep classification tree is constructed using classification tree algorithm. In last step, majority voting technique is used in classification whereas average of resultant answer is used in regression problem. Out of bag (OOB) behavior and splitting dataset into training and testing set and apply are two methods to evaluate the error in random forest.

E. ANN (Artificial Neural Network)

Data processing paradigm in machine learning approach that is derived from the concept of biological neural network. Architecture of it is composed of three layers i.e., the initial one is input layer where the data/information is captured, second is hidden layer where the requisite techniques is done (like preprocessing, feature extraction, etc.) and the last one is output layer where the network's prediction is accomplished. ANN can be trained in qualitative, quantitative as well as vector format dataset and works like parallel computational system. Backpropagation method is frequently accustomed in solving error by changing/updating weights.

For analyze the performance of all ml techniques; four different parameters i.e., accuracy, precision, recall and F1 are calculated.



		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negetive	True Negetive

Accuracy

Ratio of accurately predicted prediction of learning model is known as accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

Ratio that gives the value of true positive among all positive prediction.

$$Precision = \frac{TP}{TP + FP}$$

Recall

Same as TPR (True positive rate) that shows the ratio of true positive in total positive.

$$Recall = \frac{TP}{TP + FN}$$

F1 score

Harmonic mean of precision and recall which performs well in imbalanced dataset.

$$F1 = \frac{2 .Precision .Recall}{Precision + Recall}$$

EXPERIMENTAL SETUP / METHODOLOGY

The dataset was taken from Kaggle site https://www.kaggle.com/datasets/rashikrahmanpritom/heartattack-analysis-prediction-dataset?select=heart.csv containing 303 cases of heart attack patients with 14 feature attributes. For this experiment, selected algorithms code is done using python programming language in Anaconda Jupyter Notebook with default kernel Python 3(ipykernel) and many useful python libraries. The main objective is to analyze and predict that whether a patient will suffer from heart attack or not.

Features label and their description

Label	Description
age	age of the patient
sex	sex of the patient

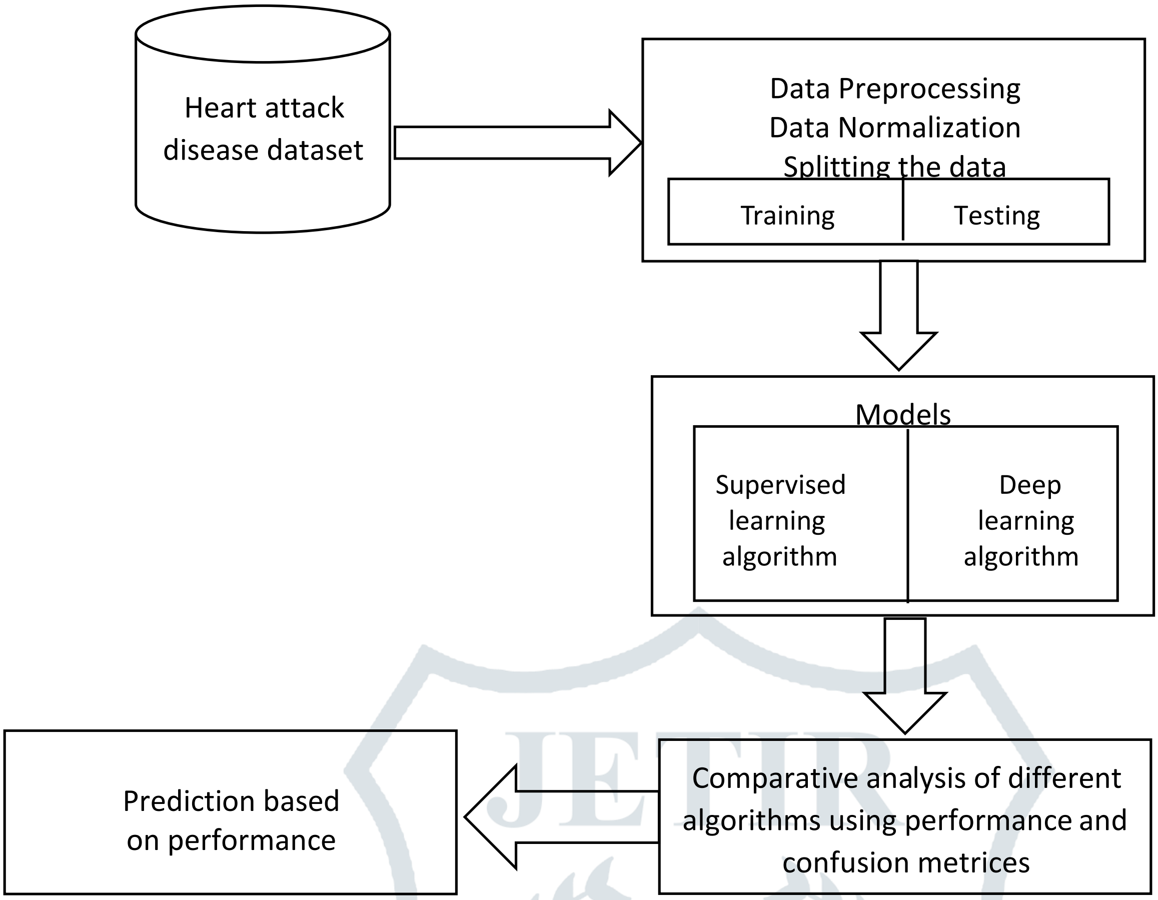
cp	Chest Pain type chest pain type ○ Value 1: typical angina ○ Value 2: atypical angina ○ Value 3: non-anginal pain ○ Value 4: asymptomatic
trtbps	resting blood pressure (in mm Hg)
chol	cholesterol in mg/dl fetched via BMI sensor
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
rest-ecg	resting electrocardiographic results ○ Value 0: normal ○ Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) ○ Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalachh	maximum heart rate achieved
exng	exercise induced angina (1 = yes; 0 = no)
oldpeak	previous peak
slp	the slope of the peak exercise ST segment
caa	number of major vessels (0-3)
thall	1= normal; 2 = fixed defect; 3 = reversable defect
output	0= less chance of heart attack 1= more chance of heart attack

After the collection of dataset, first part is to check garbage data. It may be possible that input can contain unnecessary or null data which becomes a reason to deteriorate an accuracy, thus it should be encountered in data cleaning to take off such unwanted data. In our chose dataset, no garbage data was presented so we didn’t follow this step.

Cleaned data is used in both training and testing phase which are feds as input in the algorithm.

Data preprocessing like standardization or normalization of numerical feature is done in order to set them in common scale. Label Encoding can also do (if necessary) when we want to convert categorical variable to numerical variable. Scikit-learn library is foremost prevalent library used for various purposes like implementing machine learning models, statistical modelling, many more. We used this library for two purposes i.e., splitting the dataset into training and testing part and for standardize the data.

Some well-known machine learning and deep learning algorithms are implemented upon dataset to learn and classify/predict that whether a patient is having less chance of heart attack or more chance of heart attack. For examine the performance, analysis of resultant part is done for all proposed models in terms of four different parameters of performance metrices.


Heart attack disease dataset
Data Preprocessing
Data Normalization
Splitting the data
Training
Testing
Models
Supervised learning algorithm
Deep learning algorithm
Comparative analysis of different algorithms using performance and confusion metrices
Prediction based on performance

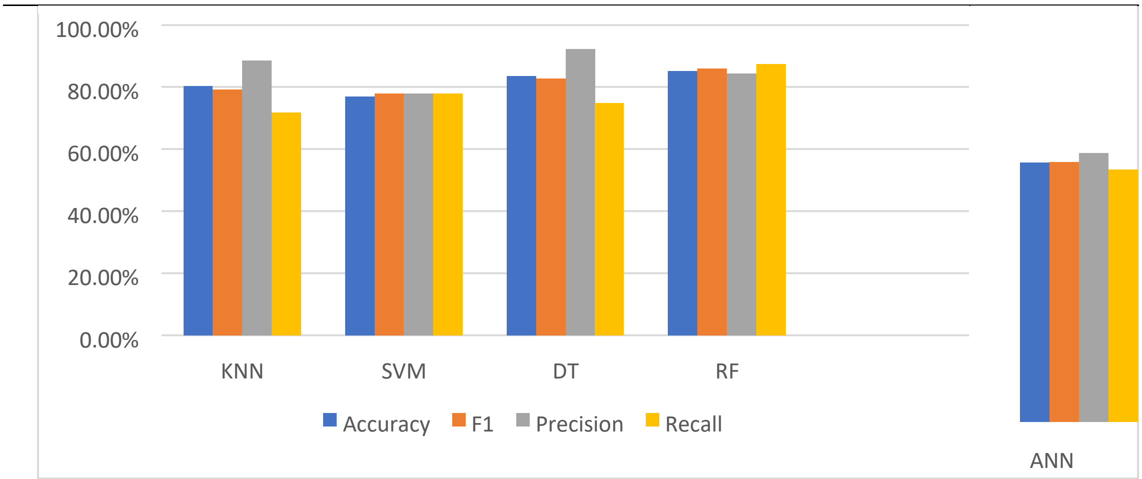
EXPERIMENTS AND RESULTS

In this work, prediction of heart attack is done using five different machine learning techniques. 80% of data is taken as training and left 20% of data is taken as test set. The outcome variable of this model is 0 as less chance of heart attack and 1 as more chance of heart attack.

A. Performance study of all models

In initial evaluation, we performed the experiment and compared their performance metrices of all models as chosen parameters strongly impact the durability and adaptability capability of machine learning algorithms. We used binary crossentropy as loss function in ANN to calculate four different parameters i.e., accuracy, precision, recall and f1 score. Table and its respective 2d chart shows the resultant percentage of techniques.

Algorithm	**Accuracy**	**F1**	**Precision**	**Recall**
KNN	80.32%	79.31%	88.46%	71.875%
SVM	77.04%	78.125%	78.125%	78.125%
DT	83.60%	82.75%	92.30%	75%
RF	85.24%	86.15%	84.44%	87.5%
ANN	83.60%	83.87%	86.66%	81.25%



B. Analysis by change in kernel for SVM

Secondly, we investigated numbers of correct and incorrect prediction by using confusion matrix and also calculated the accuracy of model upon various k value such as., linear, RBF, polynomial and sigmoid. Tested were noted by using performance and confusion metrics as shown in tables below:

	Predicted	
Actual	25(TN)	4(FP)
	3(FN)	29(TP)

Confusion matrix with linear kernel

	Predicted	
Actual	22(TN)	7(FP)
	7(FN)	25(TP)

Confusion matrix with sigmoid kernel

	Predicted	
Actual	25(TN)	4(FP)
	6(FN)	26(TP)

Confusion matrix with rbf kernel

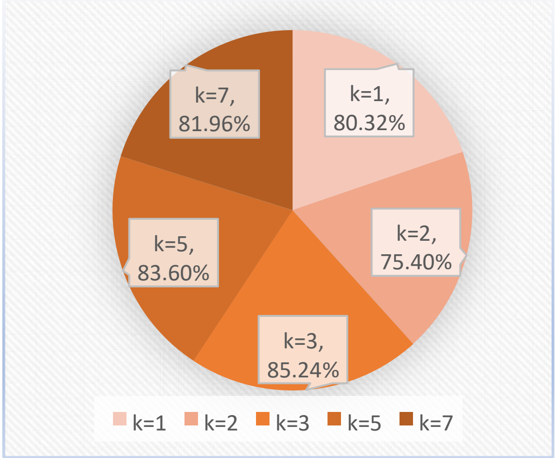
	Predicted	
Actual	24(TN)	5(FP)
	4(FN)	28(TP)

Confusion matrix with polynomial SVM

Kernel	Accuracy	Precision	Recall	F1
Linear	88.52%	87.87%	90.625%	89.23%
RBF	83.60%	86.66%	81.25%	83.87%
Poly	85.24%	84.84%	87.5%	86.153%
Sigmoid	77.04%	78.125%	78.125%	78.125%

C. Comparing accuracy and confusion matrix by changing value of K in KNN

Lastly, we only examined the accuracy of predictive model upon various k value. Four values were selected i.e., 1, 2,3,5 and 7 and tested on proposed heart disease prediction.



CONCLUSION

The research on heart disease prediction based on their symptoms by using machine learning classifier is provided in this paper. This model takes the symptom of patients and predict disease by giving output 0 as less chance and 1 as more chance of heart attack. Firstly, we have evaluated different constant of performance metrics and compared them. The outcome shows that random forest predicts output with higher accuracy, recall and F1 score as compared to other machine learning technique. Secondly, we experimented in SVM code by changing the kernels. As the dataset we used is straight-line data with less cases (303), linear kernel performed with higher accuracy and recall. The problems like non-linear nature, complex patterns and use of curve figure, then other kernels can be used for more accurate prediction. At last phase, we investigated by improvement of k value and noted their accurate predicting percentage which are beneficial for preventing underfitting and overfitting.

The proposed work can be re-introduced and enhanced for automation of disease prediction. Real data from health care organizations can be collected, available algorithms can be compared for robustness and generalizability of result.

REFERENCES

Samaa Farhan, Mohammad Alshraideh, Tareq Mahafza, “A Medical Decision Support System for ENT Disease Diagnosis using Artificial Neural Networks”, International Journal of Artificial Intelligence and Mechatronics Volume 4, Issue 2, ISSN 2320 – 5121

Raniya Rone Sarra, Ahmed Musa Dinar, Mazin Abed Mohammed “Enhanced accuracy for heart disease prediction using artificial neural network”, (January 2023), Indonesian Journal of Electrical Engineering and Computer Science, Vol. 29, No.1, pp. 375~383.

Somya Singh, Aditi Sneh, Vandana Bhattacharjee (2021) “A Detailed Analysis of Applying the K Nearest Neighbour Algorithm for Detection of Breast Cancer”, International Journal of Theoretical & Applied Sciences, 13(2)

Madhumita Pal, Smita Parija (2021) “Prediction of Heart Diseases using Random Forest”, Journal of Physics: Conference Series; 1817(2021)012009; doi: 10.1088/1742-6596/1817/1/012009

KM Jyoti Rani (2020) “Diabetes Prediction using Machine learning”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, ISSN: 2456- 3307 (www.ijsrcseit.com)

Chintan M. Bhatt, Parth Patel, Tarang Ghetia, Pier Luigi Mazzeo (2023) “Effective Heart Disease Prediction Using Machine Learning Techniques”, Algorithms 2023, *16*(2), 88; https://doi.org/10.3390/a16020088

Muhammad Asif, Muhammad Ibrar, Shahbaz Ahmad, Muhammad Arslan Farooq, Hamid Ullah, Muhammad Kashif Abbasi and Zeshan Afzal (2021) “Detection of COVID-19 from C-X-Ray Scans Empowered by Machine Learning”, International Journal on Emerging Technologies 12(2): 104109(2021)

Ankita Tyagi, Ritika Mehra, Aditya Saxana (2018) “Interactive thyroid disease prediction system using machine learning technique”, 5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018), 20-22 Dec, 2018, Solan, India 978

Chandan R, Chetan Vasan, Chethan VS, Devikarani H S (2021) “thyroid detection using machine learning”, International Journal of Engineering Applied Sciences and Technology, 2021 Vol. 5, Issue 9, ISSN No. 2455-2143, Pages 173-177

Umarani Nagavelli, Debabrata Samanta, Partha Chakraborty (2022) “Machine Learning TechnologyBased Heart Disease Detection Models”, J Healthc Eng. 2022; 2022: 7351061. Published online 2022 Feb 27, doi: 10.1155/2022/7351061

Yangguang Liu, Yangming Zhou, Shiting Wen, Chaogang Tang(2014) “ A strategy of selecting Performance Metrics for Classifier Evaluation”, International Journal of Mobile Computing and Multimedia Communications, 6(4), 20-35, October- December 2014

Keshav Srivastava, Dilip Kumar Choube(2020) “Heart Disease Prediction using Machine Learning and Data Mining”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 22773878, Volume-9 Issue-1, May 2020

V.V.Ramalingam, Ayantan Dandapath, M Karthik Raja(2018) “Heart disease prediction using machine learning techniques: a survey”, International Journal of Engineering and Technology,7 (2.8) (2018) 684-687

Tesfaye Adugna, Wenbo Xu, Jinlong Fan(2022) “Comparison of Random forest and SVM classifier for regional land over mapping using coarse resolution FY-3C images”, Remote Sens. 2022, 14, 574.https://doi.org/10.3390/rs14030574