Global Scientific JOURNALS
GSJ: Volume 10, Issue 7, July 2022, Online: ISS 2320-9186
www.globalscientificjournal.com

Heart disease prediction approach using machine learning and Multi-Criteria Decision Making methods.

T.J Malapane
Department of Electrical and Electronic Engineering Technology
University of Johannesburg
Gauteng, South Africa
216088717@student.uj.ac.za

W. Doorsamy and B.S Paul
Institute for Intelligent Systems
University of Johannesburg
Gauteng, South Africa
wdoorsamy@uj.ac.za;babusenapaul@gmail.com

Abstract
Cardiovascular diseases are considered one of the most difficult diseases to treat and many people suffer from this disease in the world including related death due to heart diseases. Prediction of heart diseases is one of the main challenges in the area of medical data investigation. Machine learning has been interesting technology in the healthcare industry as been used to analyze medical datasets and predict diseases. Most researchers in the present era are using Machine Learning techniques to predict heart disease by selecting only one or two ML models for prediction accuracy and comparison. In this paper, we propose the use of the Multi-Criteria decision-making (MCDM) method to select the best machine learning algorithm. In MCDM, we use the TOPSIS method one of the best MCDM techniques which combine both hard and soft technologies for selecting the best algorithm. The number of machine learning algorithms such as support vector regression, k-nearest neighbor (KNN), Random Forest, Decision Trees, and Logistic Regression are tested with the dataset. We have used the cardiovascular disease dataset from the Cleveland University of California Irvine (UCI) Repository which consists of 14 different attributes related to heart disease. The experimental results show that using the MCDM method, the Random Forest algorithm is the best with a performance value of 79.7% compare to the other four algorithms.

Keywords
Machine learning, MCDM, TOPSIS, Cardiovascular disease.

1. Introduction
The heart is one of the vital organs in the human body, with the functionality of pumping blood all over the body. Malfunction of the heart can cause death in most cases. World Health Organization (WHO) estimated that more than 17.3 million people die each year due to heart-related diseases which consist of 31% of all deaths globally. Strokes and heart attacks are the main cause of cardiovascular disease (CVDs) deaths with four out of five CVD deaths. Millions of people globally are having difficulties managing the risk factors that lead to cardiovascular disease, while others are unaware that they are at high risk. The risk of CVD can be identified in individuals through raises in body glucose and blood pressure, including obesity and overweight. These vital signs can be easily measured in healthcare facilities and identify those at the highest risk of CVDs.

Due to emerging technology in the healthcare industry, data collection and storage have become possible. Data investigation is one of the significant characteristics of the medical field. Medical datasets are collected and

analyzed by using different machine learning algorithms that determine certain outlines and correlations. Machine learning algorithm does not determine the root causes of disease but the major contribution is to predict diseases and learn from current data for future prevention. Machine learning has become more popular as a subset of Artificial Intelligence that can study by itself and improve from previous acknowledgments after it makes better conclusions and predictions. Machine learning algorithms are implemented to perform several tasks such as decision making, classification, and prediction.

This paper presents performance analysis by using the technique for order preference by similarity to ideal solution (TOPSIS) as one of the MCDM methods to vote for the best ML technique for CVD prediction. ML techniques in the healthcare sector play a vibrant role to detect hidden discrete patterns and analyze captured datasets. Using only ML model comparisons by only considering one or few evaluation criteria such as precision, accuracy, sensitivity, and AUC cannot replicate real model performance during data imbalances. Lots of effort from researchers have been done to predict the CVDs using different types of ML algorithms, but in this paper different method is been used by selecting ML algorithms which are Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB) and K- nearest Neighbour (KNN). MCDM is used to select the most accurate ML technique. MCDM is one of the influential tools used to discover better accuracy in complicated decision-making environments with various techniques like ELECTRE, Fuzzy AHP, SMART, TOPSIS, etc. MCDM techniques are used for the ranking and evaluation of the criteria to find the best alternative.

The more of this paper is prepared as follows: Section 2 covers the discussion of the recent work of researchers. Section 3 details and describe the proposed methodology. Section 4 clarifies, in brief, the performed experimentations and results based on the proposed technique. Finally, Section 5 delivers the conclusions of this paper.

2. Related Works

There is a lot of work carried out by researchers, using different ML algorithms to predict heart diseases and different percentages of accuracy have been obtained using various methods. The main goal is to classify and predict heart disease diagnoses.

Golande et al. analyzed several machine learning algorithms for predicting cardiac disease. In other to assess better accuracy KNN, DT, and K-Means methods were used that may be employed for classification. After some analysis, it was concluded that DT generated the best accuracy. The authors proposed that using various algorithms and parameter modification can better predict with more efficiency. G. D Kumar et al. proposed the prediction technique for Cardiovascular Disease by implementing supervised machine learning algorithms which are GB, SVM, NB, LR, and RF using a dataset from the UCI directory. After comparing the accuracies of all algorithms, the LR algorithm was providing the best result compared to other algorithms and was considered for CVD prediction. Khennou et al. used a dataset from UCI Repository, KNN was introduced to attribute missing data values in a database. Naïve Bayes and SVM machine learning algorithms are used for classification. As the author's results, 87 % accuracy is obtained from NB which is better compared to SVM. D. Shah et al. proposed cardiovascular disease prediction by using the Cleveland dataset. He used 14 attributes by implementing KKN, DT, NB, and RF algorithms. Modify null and noisy data are filtered by applying data pre-processing. KNN with a k value of 7 is used to obtain the highest accuracy of 90.79%.

Otoom et al. demonstrated an ML system for data analysis and prediction for Coronary artery heart disease. Cleveland Heart data was obtained from UCI which consists of 303 patient data cases and 76 features. Out of 76 features, 13 attributes are used. The author conducted two experiments using three ML techniques. By using WEKA data analysis tool for prediction with SVM, Functional Trees (FT), and Bayes Naïve ML techniques. For the Holdout test, the SVM technique achieved an accuracy of 88.3% and for the cross-validation test, FT achieved an accuracy of 81.5% while both Bayes Naïve and SVM achieved an accuracy of 83.8%. For cross-validation tests, the author selects and applies the 7 best attributes using the Best First selection technique. The results are: SVM provides an accuracy of 85.1%, Bayes Naïve, and FT both with an accuracy of 84.5%. Parthiban et al. used Naïve Bayes and SVM ML techniques to predict cardiovascular disease in diabetic patients. A dataset from Chennai Research Institute is used, with 500 cases of patients. 358 patients were detected with no disease and 142 patients have heart disease. Using the WEKA tool, 74% accuracy was achieved from the Naïve Bayes algorithm, and the highest accuracy of 94.6% was obtained from the SVM algorithm.

Based on the above papers, researchers selected the best ML algorithm based on the accuracy score. Most parameters such as Recall, Precision, F-measure, and Receiver Operating Characteristic – Area Under Curve (ROC-AUC) values are not considered when selecting the best ML algorithm. It is difficult to select the best algorithm

based on single evaluation criteria while other algorithms can perform better with other different evaluation criteria. The use of criteria combination methods such as MCDM is recommended as the best choice to vote for the best prediction ML algorithm.

3. Methodology

In our proposed method we are using the Cleveland UCI Repository dataset which contains several attributes that are used for cardiovascular disease analysis. Figure 1 presents the detailed flow diagram of the proposed process. An amount of 14 features were imported from the dataset. Immediately the data is preprocessed, normalized, and divided into certain percentages of the training dataset and testing dataset. We used ML algorithms which are LR, SVM, FR, NB, and KNN chosen based on their popularity. To select one best ML algorithm out of chosen five, we evaluate the proposed MCDM model by using the TOPSIS technique in terms of F1 score, accuracy, recall, precision, and AUC values. In the following subsections, we outline the dataset description, features extraction, classification ML techniques, and MCDM method used to select the best model.

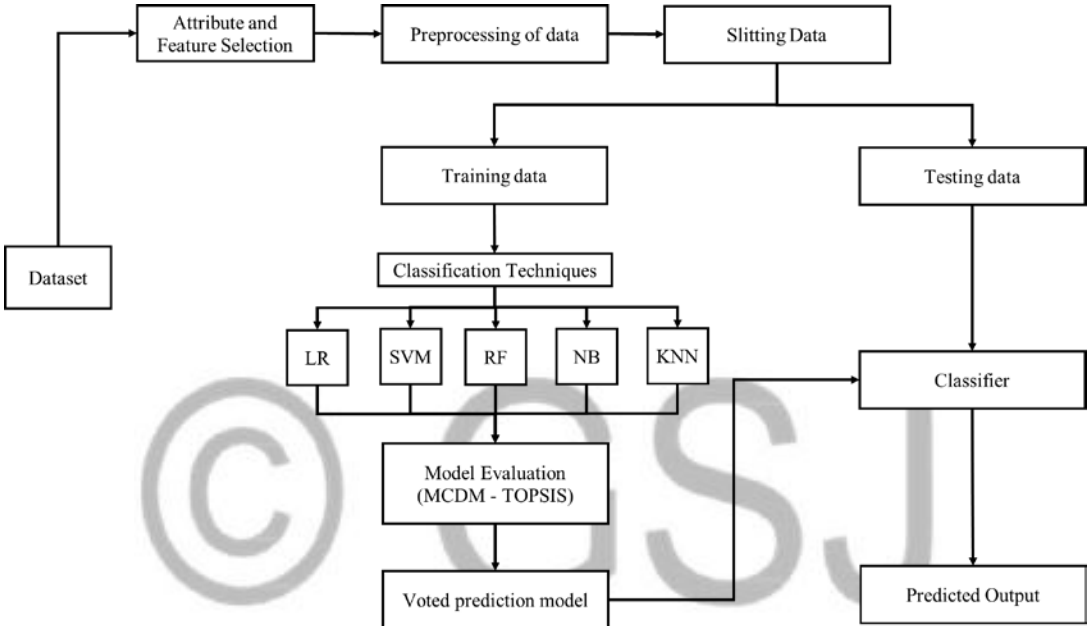

Figure 1. Proposed method flow diagram

3.1 Dataset and Attributes

In this paper, the dataset provided by Cleveland Clinic Foundation is used. Disease datasets can be found in the UCI ML repository, which covers a diverse and enormous number of datasets from different institutions. the UCI dataset consists of 303 records and 76 attributes, in our experiment only 14 attributes are used for this study as shown in Table 1.

Table 1. The attribute of the heart disease dataset

Attribute	Description	Information Attribute Type
Age	Patient's age in years (between 29 to 77 years)	Numerical
Sex	Gender of Patient (Female = 1 and Male = 0)	Nominal
Cp	Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non anginal pain, 4 = asymptomatic)	Nominal
Trestbps	Resting blood pressure in mm Hg [94, 200]	Numerical
Chol	Serum cholesterol in mg/dl [126, 564]	Numerical

Restecg	Resting ECG results (0= Normal,1=ST-Twave abnormality,2=LV hypertrophy)	Nominal
Thalach	Maximum heart rate achieved [71, 202]	Numerical
Exang	Exercise induced angina (0 = No, 1 = Yes)	Nominal
Oldpeak	ST depression induced by exercise relative to rest [0.0, 62.0]	Numerical
Slope	Slope of the peak exercise ST segment (1 = up-sloping, 2 = flat, 3 = down-sloping)	Nominal
Ca	Number of major vessels (0 – 3)	Numerical
Thal	Defect types: 3 = normal, 6 = fixed, 7 = irreversible)	Nominal
Target	Diagnosis of heart disease (1 = high chances of heart disease, 0 = less possibility of heart disease)	Nominal

3.2 Trained classifiers

We have considered five ML algorithms in our classification proposed method, i.e., Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Naïve Bayes (NB). In the following, we will briefly describe each of the different classifiers evaluated in our proposed system.

Logistic Regression

Logistic regression is a well-known supervised machine learning algorithm mostly used for binary classification tasks. The variables of logistic regression are defined by the class of categorical dependency, the result must be a categorical or discrete feature. Instead of fitting a hyperplane or straight line, logistic regression deploys the logistic function to make the result of a linear equation in a variety of 0 to 1. LR algorithm may overfit if the amount of observations is less than the number of attributes in the input dataset.

Support Vector Machine

Support Vector Machine is also a classification algorithm that manages both non-linear and linear data. SVM uses kernel functions to classify the instance then it classifies the most suitable solution based on these modifications. SVM consists of a hyperplane that divides data points with the largest margin, also known as a discrimination classifier. To limit the possibility of misclassification, SVM looks for the optimization of the margin to define a distance between the two closest data points from each respective class and hyperplane.

Random Forest

Random Forest (RF) is a supervised machine learning technique used in both regression and classification problems. Its method consists of a collection of decision trees by applying bagging or bootstrap aggregation which are less likely to be overfitting. The most advantage of RF is that it performs well with large datasets and high dimensionality. In regression problems, RF results on the average of all the outputs of each of the decision trees, while in classification problems it uses a majority voting system.

Naïve Bayes

Naïve Bayes (NB) is an easy and mostly used classification technique that is derived from the mathematical Bayes Theorem. Based on this theorem, the probability of the presence of any variant is independent of the absence or presence of any variant. This algorithm is accomplished through the Gaussian function with prior probability and defining every occurrence of a dataset selected to the class of maximum successive probability.

K-Nearest Neighbour

K-Nearest Neighbour is a simple classification technique that uses an imaginary border to categorize data. It captures all available new data and predicts the numerical target based on distance functions. It works based on a distance between the location of data and constructed on this discrete data are classified with each other. The Euclidean distance technique is used to detect the closest the training dataset is to the target.

3.3 Evaluation of Classification Algorithms

In the selection of the best ML algorithm out of five used machine learning algorithms, we have employed the MCDM method that considers different evaluation criteria. For each of the ML techniques, performance is analyzed and computed based on different evaluation metrics used such as accuracy, precision, recall, F1-score, and ROC-

AUC metrics. In our experiment, the TOPSIS method is used to estimate the weight of each valuation criteria and produce the results that are based on the best-performing algorithm in multi-criteria decision-making. The perception of TOPSIS is that the favored alternative must take the shortest distance from the positive ideal solution (PIS) and the farthest from the negative ideal solution (NIS). The algorithm of the TOPSIS method is as follows:

Algorithm 1: Steps of the TOPSIS method

1: ***Input***: Decision matrix, $\mathbf{X} = (x_{ij})_{m*n}$

2: ***Output***: Rank of each model, R_i

3: $r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}, i = 1,\dots,m, \quad j = 1,\dots,n$ ➔Normalization of the Decision Matrix

4: $v_{ij} = w_j r_{ij}, \quad i = 1,\dots,m, \quad j = 1,\dots,n$ ➔Weighted Normalized Decision Matrix

5: $A^* = \{v_1^*, v_2^*, \dots, v_n^*\}, \; A^- = \{v_1^-, v_2^-, \dots, v_n^-\}$ ➔Determine ideal worst and ideal best

6: $D_i^* = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^*)^2}, \quad D_i^- = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^-)^2}$, ➔Calculation of the Separation Measures

7: $C_i^* = \frac{D_i^-}{D_i^* + D_i^-}, \quad i = 1,2,\dots,m$ ➔Measure relative closeness

8: $R_i = Rank(C_i^*)$ ➔Ranking of relative closeness

9: ***Return closeness score and ranks***

Evaluation Metrics

In other to assess the performance of the model, we use seven standard evaluation metrics which are: F1-score, False Negative Rate (FNR), False Positive Rate (FPR), Specificity,Accuracy, Precision, and Recall. The equations are shown below:

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive\ (FP)}{Negative} \quad (1)$$

$$False\ Negative\ Rate\ (FNR) = \frac{False\ Negative\ (FN)}{Positive} \quad (2)$$

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

$$Specificity = \frac{TN}{TN+FP} \quad (7)$$

4. Experiments and Results

This section presents the outcomes of selecting the best model for heart disease through MCDM. The experiment is performed by training 80% of the dataset consisting of 242 instances with 14 different parameters and the rest 20% dataset with 61 instances are for the test. Different evaluation criteria were calculated and obtained outcomes are summarised in Table 2. Using only one performance criteria like accuracy may lead to errors in conclusions and will be more challenging to select the best algorithm, as accuracy results for LR and SVM are similar with 84.71% and NB with an accuracy of 84.3% it will be difficult to choose between them based on one evaluation criteria

Table 2. Evaluation Criteria

Dataset	Model	Accuracy (%)	F1-score(%)	False Positive Rate (%)	False Negative Rate (%)	Precision(%)	Specificity(%)	Recall(%)
Heart Dataset	Logistic Regression	84,71	86,64	12,37	17,24	90,91	77,27	82,76
	Support Vector Machine	84,71	86,44	13,86	16,31	89,39	79,09	83,69
	Random Forest	94,21	94,81	3,85	7,25	88,64	90,91	92,75
	Naïve Bayes	84,3	86,03	14,71	16,43	88,64	79,09	83,57
	K-Nearest Neighbour	75,21	75,81	30,16	18,96	71,21	80	81,03

In figure 2, we have demonstrated a correlation heatmap that visualizes the high and low percentage values of all alternatives.

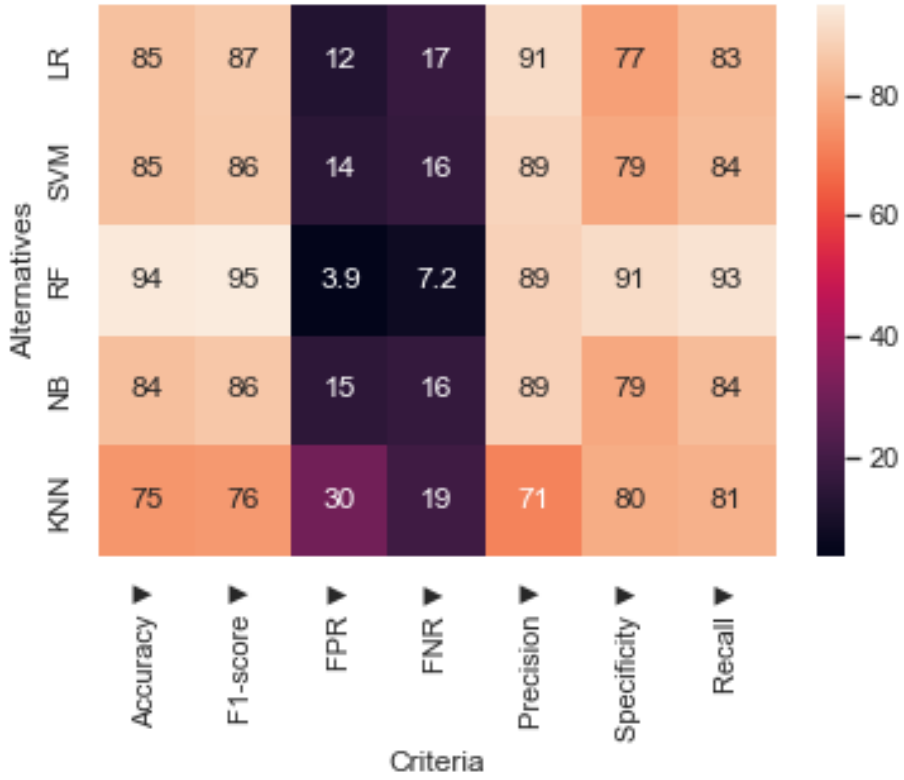

Figure 2. Correlation heatmap

Figure 3 provides the ROC (Reciever Operating Characteristics) curves for comparison of five algorithms which are created by plotting the True Positive Rate (TPR) as a vertical coordinate compared to the False Positive Rate (FPR) as a horizontal coordinate. Area Under Curve (AUC) is a measurement of the overall performance of each model and the percentage of how well the algorithm works.

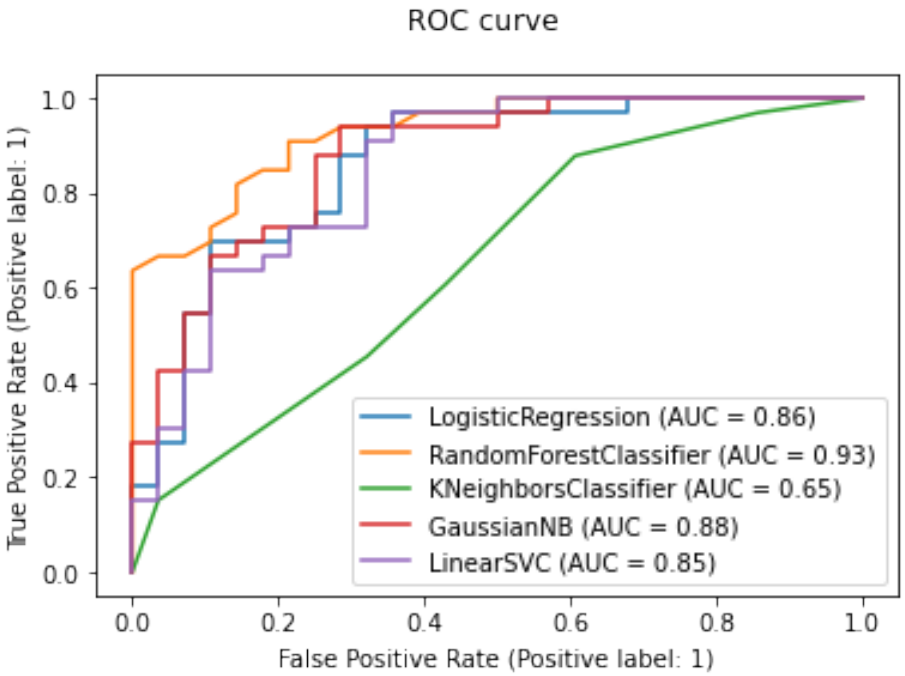

Figure 3. ROC Curve Comparison

After analyzing the results by integrating the TOPSIS method, the Random Forest algorithm became the high ranking with a performance value of 79.7% as shown in Table 3. using the TOPSIS method, algorithms are ranked based on the calculation of relative closeness value and measurement of separations, which reproduce how close each algorithm is to the ideal worst and ideal best as described in Algorithm 1.

Table 3. TOPSIS perfomsnce results.

Model	Perfomance value (%)	Rank
Logistic Regression	55.62	2
Support Vector Machine	53.46	3
Random Forest	79.7	1
Naïve Bayes	51.23	4
K-Nearest Neighbour	20.36	5

In the conclusion of our experiment, results show that MCDM is the best method to select an ML algorithm. The quantitative results of performance measurement metrics resulting from ML algorithms, including ranking are provided in Table 3. The best-performing algorithm is decided based on the TOPSIS performance metric model. Random Forest ML algorithm was the most accurate compared to other machine learning algorithms.

6. Conclusion and future work

In this paper, we proposed and presented the Multi-Criteria Decision-Making (MCDM) method using the TOPSIS model to select the best classification algorithm for the analysis of heart disease prediction. Various classification techniques are distinct in this paper which have arisen in recent years for effectiveness and efficiency in the diagnosis of heart disease. The motive of this paper was to find the most effective ML algorithm by comparing the evaluation criteria performance values of Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbour. Based on our analysis shows that Random Forest has the maximum performance value of 79.7% when compared to the other four ML algorithms.

For future work, more than five different machine learning techniques will be used for better heart disease prediction analysis and different data from multiple medical institutes can be collected and can be used for better MCDM evaluation to validate the weaknesses and strengths of machine learning algorithms.

References

A. Golande and T. P. Kumar, "Heart disease prediction using effective machine learning techniques," *International Journal of Recent Technology and Engineering*, vol. 8, pp. 944–950, 2019.

A.U. Haq, J.P.Li, M.H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," *Hindawi Mobile Information Systems,* Article ID3860146 (2018).

A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease", *International Journal of Software Engineering and Its Applications*, Vol.9, No.1, pp. 143-156, 2015.

B Padmaja, Myneni Madhu Bala, E Krishna Rao Patro (2020), "A Comparison on visual prediction models For MAMO(multi activity multi-object) recognition using Deep Learning," in Journal of Big Data, 7(24), pp.1-15,Springer.

D. Shah, S. Patel and S. Bharati, *"Heart Disease Prediction using Machine Learning Techniques,"* Springer Nature Singapore Pte Ltd, (2020).

F. Khennou, C. Fahim, H. Chaoui, and N.E.H. Chaoui, "A Machine Learning Approach: Using Predictive Analytics to Identify and Analyze High Risks Patients with Heart Disease*," International Journal of Machine Learning and Computing*, " 9 (2019).

G. Dinesh Kumar, D. Santosh Kumar, K. Arumugaraj, V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *International Conference on Current Trends toward Converging Technologies, Proceeding of IEEE*, (2018).

G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients", *International Journal of Applied Information Systems*, Vol.3, No.7, pp.2249-0868, 2012.

I.A. Zriqat, A.M. Altamimi and M. Azzeh, "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods*," International Journal of Computer Science and Information Security 14*,868-879 (2016).

M. Lichman, "UCI Machine Learning Repository".https://archive.ics.uci.edu/, 2013.

Nikoomaram.H, M.Mohammadi, M. JavadTaghipouria and Y. Taghipourian(2009). "Training Performance Evaluation of Administration Sciences Instructors by Fuzzy MCDM Approach". Tehran, Iran.

U.H.Dataset,"UCI Machine Learning Repository",https//archive.ics.uci.edu/ml/machine-learning databases/heartdisease/

World Health Organization, *Cardiovascular Diseases*, WHO, Geneva, Switzerland, 2020, https://www.who.int/healthtopics/ cardiovascular-diseases.