Heart Disease Prediction based on External Factors: A Machine Learning Approach

Maruf Ahmed Tamal[1], Md Saiful Islam[2]
Md Jisan Ahmmed[3], Md. Abdul Aziz[4], Pabel Miah[5]
Department of Computer Science and Engineering
Daffodil International University, Dhaka, Bangladesh

Karim Mohammed Rezaul[6]
Faculty of Arts, Science and Technology
Wrexham Glyndŵr University
Wrexham, UK

Abstract—**Technology has immensely changed the world over the last decade. As a consequence, the life of the people is undergoing multiple changes that directly have positive and negative effects on health. Less physical activity and a lot of virtual involvements are pushing people into various health-related issues and heart disease is one of them. Currently, it has gained a great deal of attention among various life-threatening diseases. Heart disease can be detected or diagnosed by different medical tests by considering various internal factors. However, this type of approach is not only time-consuming but also expensive. Concurrently, there are very few studies conducted on heart disease prediction based on external factors. To bridge this gap, we proposed a heart disease prediction model based on the machine learning approach which enables predicting heart disease with 95% accuracy. To acquire the best result, 6 distinct machine learning classifiers (Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, Quadratic Discriminant, and Logistic Regression) were used. At the same time, sklearn.ensemble.ExtraTreesClassifier has been used to extract relevant features to improve predictive accuracy and control over-fitting. Findings reveal that Support Vector Machine (SVM) outperforms the others with greater accuracy (95%).**

Keywords—Heart disease; Risk prediction; Decision Tree (DT); Support Vector Machine (SVM); Naive Bayes (NB); Random Forest (RF); Logistic Regression (LR); Quadratic Discriminant Analysis (QDA); Machine learning

I. Introduction

Recently, heart disease has become the leading cause of human death comparing other life-threatening diseases. As human life itself depends on the heart's function effectively, if it does not function properly, it will affect various parts of the human body. Heart disease (also known as a coronary disease) continues the world's major cause of death for centuries. 1/3 deaths of the world are caused by coronary disease and the death rate is higher than cancer mortality rates [1]. A large number of people around the world are struggling to control the risk factors of cardiovascular disease. Several factors are responsible for heart diseases like a family history of coronary illness, smoking, poor eating methodology, high pulse, cholesterol, high blood cholesterol, obesity, physical inertia, overweight, high blood pressure, stress or hypertension, chest pain, taking a drug, etc. [2, 3]. The diagnosis of cardiac disease is usually based on the patient's signs, external symptoms, and physical tests (Electrocardiogram (ECG), Holter monitoring, Echocardiogram, Stress test, Cardiac catheterization, Cardiac computerized tomography (CT) scan, etc.), but the main challenge facing by medical providers is to provide quality services at manageable cost [5]. At the same time, overall diagnosis processing is time-consuming [6] which has a negative impact on patients, especially those in need of emergency treatment. To remove this barrier, researchers are trying to utilize different Machine Learning (ML) algorithms like Decision Tree (DT) [4], Multi-layer perceptron (MLP), Artificial Neural Network (ANN) [6] [5], Naïve Bayes (NB) , K-closest neighbor (K-NN) and Support vector machine (SVM) [6] [7] [8] for distinguishing and extricating valuable data from the clinical dataset with insignificant client inputs and efforts [9]. Nevertheless, there are still very few effective measures to solve these problems. Detecting heart disease is still now a big challenge because of the overall diagnosis process and cost [10]. Researchers are relentlessly trying to develop an early heart disease detection system to minimize the death rate. Nevertheless, the lack of diversity in previous studies is a matter of great concern. As a result, improvement in this field is being hampered. Table I shows that almost all the past studies [5, 6, 7, 10, 12, 14, 15, 16, 8, 19] were conducted on the secondary data [11] which was published by UCI (Machine Learning Repository) where the features were almost same. At the same time, this dataset is a little bit outdated too (donated in 1988). So, the lack of innovation is a major concern here. Concurrently, most of the existing studies have been performed on both internal and external factors (see Table I) where heart disease diagnosis using internal factors (MRI, ECG, Echocardiogram, Blood Test, etc.) is not only costly but also time-consuming. Another downside of previous studies is less predictive accuracy (see Table I). A limited data set [5, 6, 7, 14, 18] as well as a lack of features extraction [5, 6, 13, 14], are the main reasons behind this poor accuracy. On the other hand, a model has been [5] proposed where a multilayer perceptron neural network with backpropagation was used which gained 100% accuracy. However, they did not clarify the methodologies well. The following specific objectives were followed to resolve all these gaps in order to achieve the main objective of this paper:

1) Detecting heart disease at an initial stage using machine learning based on external factors.

2) Providing less time-consuming services for a more reliable diagnosis of heart disease.

TABLE. I. SUMMARY OF THE PREVIOUS STUDY

Ref	Data Size	Type of data	Type of Approach	No. of features	Features extraction	Algorithm used	Based on	Accuracy
[7]	270	Secondary	Machine Learning	13	Yes	SVM	External + internal factors	Max. 88.34%
[6]	270	Secondary	Machine Learning	13	No	ANN, kNN, SVM, LR, CT	External + internal factors	Max. 83%
[5]	182	Secondary	Data Mining	15	No	MPNN with backpropagation	Internal factors	100%
[12]	1190	Secondary	Statistics	14	Yes	Fuzzy system	Internal + External factors	Av. 92.3%
[10]	N/A	Secondary	Data Mining	14	No	DT, K-mean Clustering	Internal + External factors	N/A
[13]	1000	Secondary	Data Mining	13	No	Decision Support and NB	Internal + External factors	Max. 88.33%
[14]	300	Secondary	Data Mining	14	Yes	DT, SVM, NB	Internal + External factor	Max. 84.85%
[15]	N/A	Secondary	fuzzy decision support system	14	Yes	NN, Clinical Decision support system, RF, J48	External + internal factors	Almost 80%
[16]	4146	Secondary	Machine learning	16	Yes	LR,NN	External + internal factors	81.163%
[17]	370	Primary +Secondary	Data Mining	13	No	KStar, J48, SMO, Bayes Net, MLP	External + internal factors	89%
[18]	303	Secondary	Machine Learning	**14**	Yes	ANN, BNN	External + internal factors	**95%**
[19]	N/A	Secondary	Data Mining	11	Yes	NN, Bayesian Networks, DT, SVM	External + internal factors	93%

N.B. MPNN = Multilayer Perceptron Neural Network, ANN = Artificial Neural Network, DT = Decision Tree, NN = Neural Network, SVM = Support vector Machine, NB = Naïve Bayes, SMO = Sequential Minimal Optimization, BNN = Backpropagation Neural Network.

II. RESEARCH METHODOLOGY

The primary purpose of this study is to explore the best predictive model based on external symptoms for diagnosing heart disease. Fig. 1 represents the methodological framework of the current study. The methodological part was separated into several sections to clearly reflect the overall study.

A. Data Collection

A substantial literature review was performed at the beginning of the study to identify the gap in the existing studies. Simultaneously, significant study issues were found through literature review which assisted to collect relevant data and factors. This study was conducted between July 2018 and September 2019 at Dhaka, Bangladesh. Primary data was collected through both field surveys and online surveys. A web-based questionnaire (Google form) was sent to the targeted audience of various ages.

B. Participants

A total of 3500 questionnaires were distributed and 1247 valid records were gathered (field survey: 952 internet survey: 295), including spontaneous female (43% and male (57%) participants. The ethical factors of the respondent have been closely assured to keep their privacy strictly and confidentially secret.

C. Data Preprocessing

In order to obtain the precise value, the dataset was fully prepossessed until irrelevant, incomplete, inconsistent information was removed. Several python libraries were used to preprocess the raw data. At the same time, we also carried out data transformation from string to numerical value in order to fit the data with the classifiers.

D. Feature Extraction

Feature extraction is a process of minimizing dimensionality by reducing less effective features from raw data which to helps extract relevant features to improve predictive accuracy and control over-fitting. Fig. 2 indicates more significant features from bottom to top. From 14 features, we took the first 11 features that helped us get more accuracy. To extract the important features “sklearn.ensemble.ExtraTrees Classifier” class has been used.

E. Model Selection

The final data set was split into a training set (80%) and a testing set (20%). The top six common algorithms have been selected to explore the best-performance machine learning classifier for predicting heart disease.

a) Decision Tree (DT): A non-parametric supervised classification and regression learning method. The goal is to create a model that predicts the value of a target variable by learning simple rules of a decision based on data characteristics.

b) Support Vector Machine (SVM): A Vector Support Machine (SVM) is formally defined as a biased classifier by a particular hyper plane. The determination function of SVMs relies on some of the training data sub-sets called support vectors.

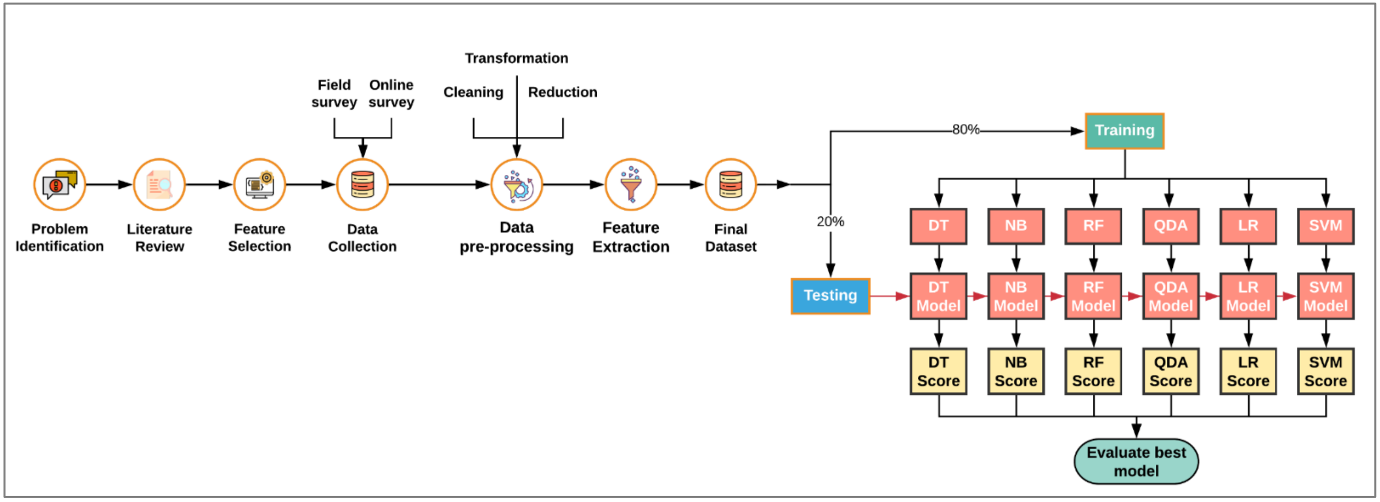
c) Naïve Bayes (NB): It is one of the learning algorithms that are commonly used. The NB classifier is a Bayes rule-based probabilistic model. We used GaussianNB in this current study to create a predictive model.

d) Random Forest (RF): Random forest (RF) is a meta estimator that suits a variety of decision tree classifiers on different data set sub-samples and uses the average to boost predictive precision and over-fitting power.

e) Logistic Regression (LR): It is a probabilistic classifier, generally applied to problems of binary classification.

f) Quadratic Discriminant Analysis (QDA): A quadratic classifier is used to distinguish observations from two or more groups of artifacts or occurrences by a quadric layer in ML and numerical classification.

In this study, all of the above algorithms were implemented using Scikit-learn which is a Python-based open-source machine learning library.



N.B. DT= Decision Tree, SVM=Support Vector Machine, NB=Naive Bayes, RF=Random Forest, LR=Logistic Regression, QDA= Quadratic Discriminant Analysis

Fig. 1. Methodological Framework.

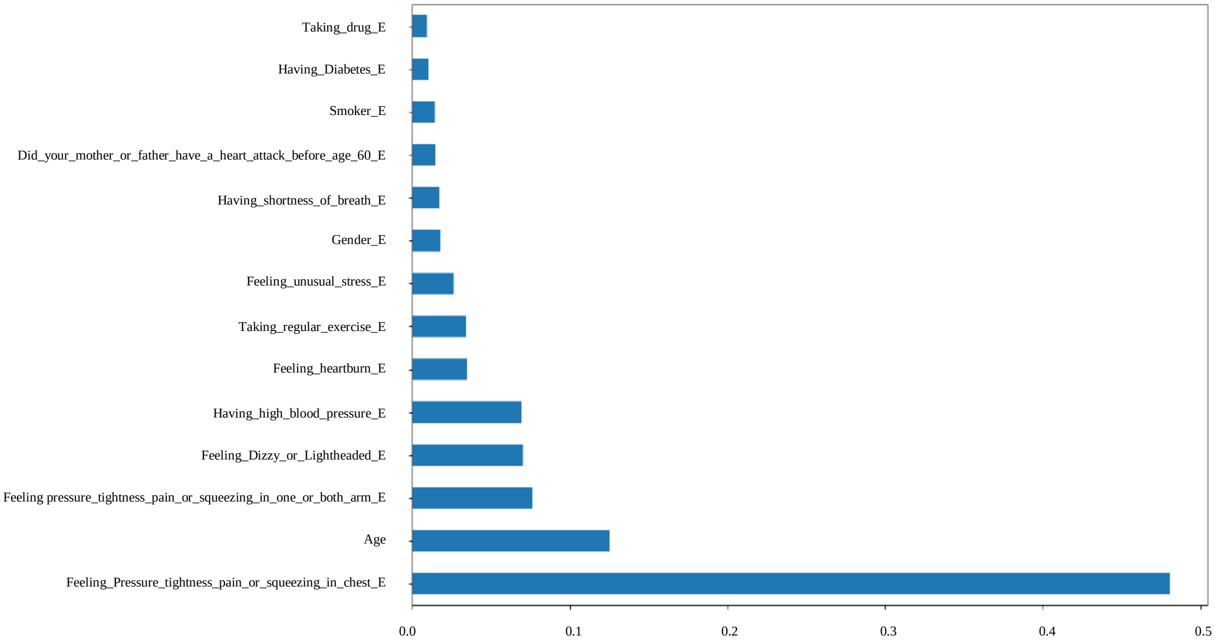


Fig. 2. Feature Importance.

III. Data Analysis and Result

A. Respondents based on Gender

3,500 survey questionnaires had been distributed to collect the primary data, over the heart patients and normal people. Total 1247 responses were collected and preprocessed. After removing the irrelevant, incomplete, inconsistent records, 1201 records were selected for analysis. The proportion of participants among male and female were 57% and 43% respectively. Where the participants ' age ranged from 10 to 95. Table II represents the frequency distribution of the overall dataset according to gender.

B. Materials and Feature Selections

As our key objective is to early predict heart disease without performing any medical test, we focused on the external symptoms of heart disease to design survey questionnaires. 14 individual factors were identified and selected for survey questionnaires to achieve the research goal. To improve prediction accuracy and control over-fitting, feature selection algorithm was performed and 12 out of 14 most important features were selected for final analysis (see Table III and Fig. 2).

C. Performance Measurement of Classification Algorithms

a) Confusion matrix: A confusion matrix is represented by a table (see Table IV) that measures the performance of a classification model. By using certain terminologies (TP, TN, FP, FN), it summarizes a classifier's correct and incorrect predictions. In a confusion matrix, True Positive (TP) represents the correctly predicted positive values, True Negative (TN) represents correctly predicted negative values, False Positive (FP) represents that the classifier predicted the value as positive but it was false, False Negative (FN) represents that the classifier predicted the value as negative but it was false. Fig. 3 represents the summary of the confusion matrix of the selected classifiers.

b) Precision: Precision represents the ratio of correctly predicted positive observations of the total predicted positive observations. High precision indicates that the classification model has a low false-positive rate. Table V shows the performance evaluation of the selected classifiers where the Support Vector Machine (SVM) gives high precision.

Precision (P) = TP/ (TP+FP)

c) Recall: Recall which is commonly known as sensitivity represents the ratio of correctly predicted positive observations to all observations in the actual class. Recall measures what proportion of people that actually had heart disease was diagnosed by the classifier as having heart disease. Support Vector Machine (SVM) gives a high recall rate among all other classifiers (see Table V).

Recall (Sensitivity) = TP/TP+FN

d) F1-score: F1-Score represents the weighted average of Precision and Recall that measures a test's accuracy. In our study, Support Vector Machine (SVM) gives a high F1-score rate among all other classifiers (see Table V).

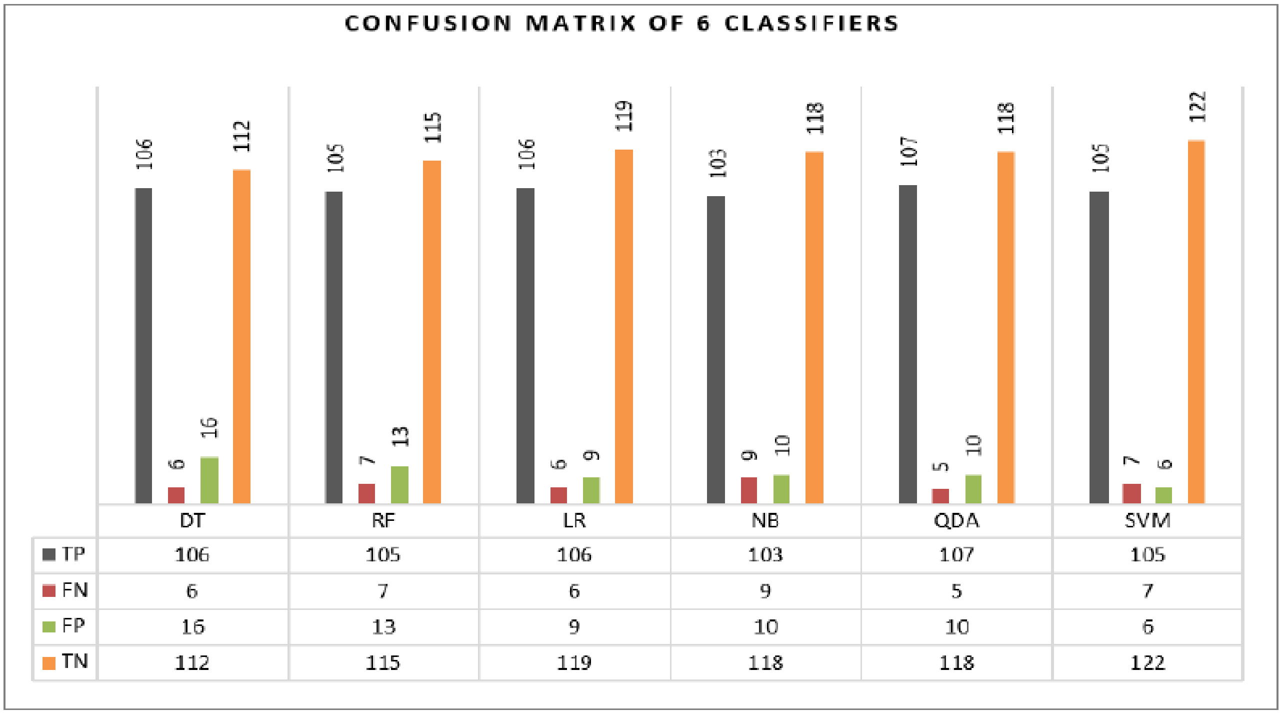
F1-Score = 2*(Recall * Precision) / (Recall + Precision)

TABLE. II. Frequency Distribution According to Sex

Gender	Frequency	Frequency distribution	People type	
			Heart patient	***Normal***
Male	711	57%	51.6%	48.4%
Female	536	43%	52.1%	47.9%

TABLE. III. Feature Details

Features	Feature rank according to the importance	Final feature selection status
Feeeling_pressure_tightness _pain_or_squeezing_in_chest	1	Yes
Age	2	Yes
Feeling_pressure_tightness_pain_or_squeezing_in_one_or_both_arm	3	Yes
Feeling_Dizzy_or_Lightheaded	4	Yes
Having_high_blood_pressure	5	Yes
Feeling_heartburn	6	Yes
Taking_regular_extercise	7	Yes
Feeling_unusual_stress	8	Yes
Gender	9	Yes
Having_shortness_of_breath	10	Yes
Having_parents_heart_attack_before_age_60	11	Yes
Smoker	12	Yes
Having_Diabetes	13	No
Taking_drug	14	No



	DT	RF	LR	NB	QDA	SVM
■ TP	106	105	106	103	107	105
■ FN	6	7	6	9	5	7
■ FP	16	13	9	10	10	6
■ TN	112	115	119	118	118	122

N.B. TP = True Positive, TN = True Negative, FN = False Negative, FP = False Positive, DT = decision tree; SVM = Support Vector Machine, NB = Naive Bayes, RF = Random Forest, LR = Logistic Regression, QDA = Quadratic Discriminant Analysis

Fig. 3. Summary of the Confusion Matrix of the Selected Classifiers.

TABLE. IV. CONFUSION MATRIX

	Predicted **0**	Predicted **1**
Actual **0**	**TN**	**FP**
Actual **1**	**FN**	**TP**

TABLE. V. PERFORMANCE EVALUATION OF THE SELECTED CLASSIFIERS

Classifier	Accuracy	Class	Precision	Recall	F1-score	Support
Naïve Bayes	92%	Heart Patient	0.91	0.92	0.92	112
		Normal	0.93	0.92	0.93	128
Quadratic Discriminant Analysis	94%	Heart Patient	0.91	0.96	0.93	112
		Normal	0.96	0.92	0.94	128
Logistic Regression	94%	Heart Patient	0.92	0.95	0.93	112
		Normal	0.95	0.93	0.94	128
Support Vector Machine	**95%**	**Heart Patient**	**0.95**	**0.94**	**0.94**	**112**
		Normal	**0.95**	**0.95**	**0.95**	**128**
Decision Tree	91%	Heart Patient	0.87	0.95	0.91	112
		Normal	0.95	0.88	0.91	128
Random Forest	92%	Heart Patient	0.89	0.94	0.91	112
		Normal	0.94	0.90	0.92	128

IV. DISCUSSION

The diagnosis of cardiac disease is usually based on the patient's signs, external symptoms, and physical tests. Since the diagnosis of heart disease is time-consuming and expensive, this type of treatment cannot be adopted by everyone. So providing quality services at manageable cost has become a major issue. The purpose of the present study is to find an effective and less expensive way to predict heart disease based on the external risk factors.

The overall study was conducted on 1247 samples where 51% of the sample had heart disease and 49% of the sample was normal (see Table II). A total of 14 external risk factors were identified (see Table III) and 12 factors were eventually selected for further analysis based on the importance of the feature. Six distinct machine learning classifiers (Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, Quadratic Discriminant, and Logistic Regression) were used to obtain the best result. The confusion matrix of prediction results is presented in Fig. 3. The result shows that the Support Vector Machine performed best to identify True negative values. Concurrently, Quadratic Discriminant Analysis did its best to define True Positive values. Overall results (see Table V), however, show that SVM outperformed all other machine learning classifiers with a peak classification accuracy of 95%, whereas LR and QDA (94%) achieved second-highest classification accuracy. Through contrast, the overall recall of 94.5 percent, 94 percent and 94 percent was shown by all three classifiers. At the same time, NB, DT, RF displays 92%, 91%, and 92% accuracy, respectively.

V. Conclusion and Future Work

Globally, heart disease (also known as cardiovascular disease) has become a major concern due to its destructive behaviour. It can be detected or diagnosed by different medical tests by considering various internal factors. Predicting heart disease (based on internal factors) using Machine Learning is a common approach. However, there are very few studies conducted on heart disease prediction based on external factors. In this study, we proposed a heart disease prediction model (based on external factors) using a machine learning approach that enables predicting heart disease with 95% accuracy. To acquire the best result, six distinct machine learning classifiers (Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, Quadratic Discriminant, and Logistic Regression) were used. Findings reveal that Support Vector Machine (SVM) outperforms the others with greater accuracy (95%). This study's future work involves designing an Android-based application that is based on the results of the current study and helping the general public predict their cardiovascular disease at no cost.

References

[1] Y. Gultepe and S. Rashed, “The Use of Data Mining Techniques in Heart Disease Prediction,” International Journal of Computer Science and Mobile Computing, vol. 8, no. 4, pp. 136–141, Apr. 2019.

[2] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, 2016, pp. 1-5.doi: 10.1109/ICCPCT.2016.7530265.

[3] AK. Peters A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 704-706.

[4] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207.doi: 10.1109/ISCC.2017.8024530.

[5] P. Singh, S. Singh, and G. S. Pandi-Jain, “Effective heart disease prediction system using data mining techniques,” International Journal of Nanomedicine, vol. 13, pp. 121–124, 2018.

[6] A. K. Dwivedi, “Performance evaluation of different machine learning techniques for prediction of heart disease,” Neural Computing and Applications, vol. 29, no. 10, pp. 685–693, 2016.

[7] C. B. Gokulnath and S. P. Shantharajah, “An optimized feature selection based on genetic approach and support vector machine for heart disease,” Cluster Computing, vol. 22, no. S6, pp. 14777–14787, 2018.

[8] 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017.

[9] Srinivas, K., Rani, B.K., Govrdhan, A., 2010. “Applications of data mining techniques in healthcare and prediction of heart attacks”. Int. J. Comput. Sci. Eng. (IJCSE), 2010, Vol. 2, No. 2, pp. 250–255.

[10] S. Babu et al., "Heart disease diagnosis using data mining technique," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 750-753. doi: 10.1109/ICECA.2017.8203643.

[11] Archive.ics.uci.edu. (1988). UCI Machine Learning Repository: Heart Disease Data Set. [online] Available at: https://archive.ics.uci.edu/ml/datasets/heart+Disease [Accessed 26 Dec. 2019].

[12] A. K. Paul, P. C. Shill, M. R. I. Rabin, and K. Murase, “Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease,” Applied Intelligence, vol. 48, no. 7, pp. 1739–1756, Jun. 2017.

[13] Mamatha Alex P and Shaicy P Shaji, “Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique " International Conference on Communication and Signal Processing, April 4-6, 2019, India.

[14] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2019, pp. 619-623. doi: 10.1109/IBCAST.2019.8667106.

[15] A. K. Paul, P. C. Shill, M. R. I. Rabin and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, 2016, pp. 145-150. doi: 10.1109/ICIEV.2016.7759984.

[16] J. K. Kim and S. Kang, “Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis,” Journal of Healthcare Engineering, vol. 2017, pp. 1–13, 2017.

[17] M. Sultana, A. Haider and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, 2016, pp. 1-5. doi: 10.1109/CEEICT.2016.7873142.

[18] T. Karayılan and Ö. Kılıç, "Prediction of heart disease using neural network," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 719-723. doi: 10.1109/UBMK.2017.8093512.

[19] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, “A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease,” Design Automation for Embedded Systems, vol. 22, no. 3, pp. 225–242, Nov. 2018.