IProCAD: Intelligent Prognosis of Coronary Artery Disease Excluding Angiogram in Patient with Stable Angina

Md. Shah Jalal Jamil, A. K. M. Muzahidul Islam, Bulbul Ahamed, Mohammad Nurul Huda

Abstract: ***Cardiovascular diseases are one of the main causes of mortality in the world. A proper prediction mechanism system with reasonable cost can significantly reduce this death toll in the low-income countries like Bangladesh. For those countries we propose machine learning backed embedded system that can predict possible cardiac attack effectively by excluding the high cost angiogram and incorporating only twelve (12) low cost features which are age, sex, chest pain, blood pressure, cholesterol, blood sugar, ECG results, heart rate, exercise induced angina, old peak, slope, and history of heart disease. Here, two heart disease datasets of own built NICVD (National Institute of Cardiovascular Disease, Bangladesh) patients', and UCI (University of California Irvin) are used. The overall process comprises into four phases: Comprehensive literature review, collection of stable angina patients' data through survey questionnaires from NICVD, feature vector dimensionality is reduced manually (from 14 to 12 dimensions), and the reduced feature vector is fed to machine learning based classifiers to obtain a prediction model for the heart disease. From the experiments, it is observed that the proposed investigation using NICVD patient's data with 12 features without incorporating angiographic disease status to Artificial Neural Network (ANN) shows better classification accuracy of 92.80% compared to the other classifiers Decision Tree (82.50%), Naïve Bayes (85%), Support Vector Machine (SVM) (75%), Logistic Regression (77.50%), and Random Forest (75%) using the 10-fold cross validation. To accommodate small scale training and test data in our experimental environment we have observed the accuracy of ANN, Decision Tree, Naïve Bayes, SVM, Logistic Regression and Random Forest using Jackknife method, which are 84.80%, 71%, 75.10%, 75%, 75.33% and 71.42% respectively. On the other hand, the classification accuracies of the corresponding classifiers are 91.7%, 76.90%, 86.50%, 76.3%, 67.0% and 67.3%, respectively for the UCI dataset with 12 attributes. Whereas the same dataset with 14 attributes including angiographic status shows the accuracies 93.5%, 76.7%, 86.50%, 76.8%, 67.7% and 69.6% for the respective classifiers.***

Keywords: Machine learning, Data mining, Coronary artery disease, Artificial neural network, Heart disease.

I. INTRODUCTION

At present, more than seventeen (17) million people die globally every year by cardiovascular diseases, out of them 30% die due to the cardiac attack and 80% of them happen in the under developing or low-income countries [1]. Bangladesh is one of them despite the exact number of incidences of the disease in Bangladesh is not known. But the identification of heart disease depends on costly and composite of biomedical, pathological and clinical data [2].

In the low-income countries, most of the people avoid medical checkup because of ignorance and awareness due to unaffordability of costly diagnosis for example, angiogram. Besides, heart disease or coronary artery disease is most common and hazardous among all the diseases universally because of blockage of blood stream to the brain. Moreover, the angiogram is costly, complex, time consuming, and also not available in rural area especially in Bangladesh. Therefore, an intelligent system that predicts heart status by measuring simple information of low cost is inevitable, which assists medical practitioners or non-cardiologists.

For intelligent prediction system it is needed to integrate supervised (Artificial Neural Network, Support Vector Machine, Naïve Bayesian Classifier, Decision Tree, Logistic regression, Random Forest) or unsupervised machine learning (ML) algorithms may be used to detect the status of cardiac disease based on the feature of the clinical data [3], [4], [5].

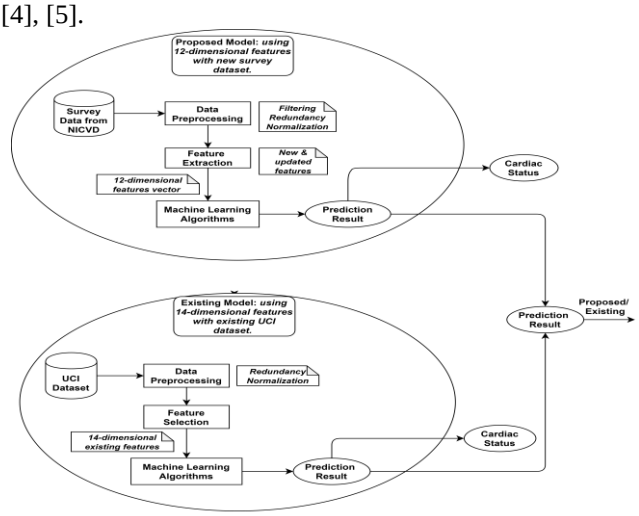

Fig. 1. Overview of this research.

Revised Manuscript Received on March 05, 2020.

***** Correspondence Author

Md. Shah Jalal Jamil*, Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh. E-mail: mjalal162017@mscse.uiu.ac.bd, Mob: +880 1718287450.

A.K.M. Muzahidul Islam, Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh. E-mail: muzahid@cse.uiu.ac.bd

Bulbul Ahamed, Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. E-mail: bulbul_cse@su.edu.bd.

Mohammad Nurul Huda, Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh. E-mail: mnh@cse.uiu.ac.bd.

In this study the proposed system suggests a low-cost prediction of cardiac status that comprises three stages where we collect dataset through survey questionnaires from National Institute of Cardiovascular Disease (NICVD), Bangladesh [6] to construct corpus at first stage. At second stage we have reduced the noisy data that are integrated in the survey and the dimension of the data from standard 14 features to 12 features by excluding the angiographic status. Finally, we have injected 12-dimensional features vector (age, sex, chest pain, blood pressure, cholesterol, blood sugar, ECG results, heart rate, exercise induced angina, old peak, slope, and history of heart disease or Stenosis) to machine learning tools to predict the cardiac status. The overall procedure of our research is depicted in the "Fig. 1".

The originalities of the research are to construct the corpus using the Bangladeshi patients' data from survey and to reduce the dimensionality of features vector by excluding the high cost angiographic status. Although the data collection process is inconvenient and risky, it seems that we are the first to collect heart disease data in patient with stable angina in Bangladesh so far.

The paper is organized as follows. Sections II explains literature review. The methodology of this research has been explicated in Section III with diagrams, whereas Sections IV and V analyze the experiments with corpus collection procedure and set up, and results. Finally, Sections VI concludes the paper, and Section VII highlights the limitations with future remarks.

II. LITERATURE REVIEW

In this section, we have reviewed some related works given in the Table-I with architectures (authors name with reference, classifiers name, number of features in input vector) and comments, where most of the authors used UCI dataset [7] with 14 features by incorporating Fuzzy Logic, Genetic Algorithms and various ML algorithms.

Table- I: Reviewed some related works with analysis.

Authors Name & Reference No.	Classifiers	Input Vector	Comments
E. P. Ephzibah and Dr. V. Sundarapandian [8]	Genetic Algorithm and Fuzzy Expert System	6	The system can only implement the rules and cannot learn as it goes along. They have not considered important attributes.
MoloudAbdar, Sharareh R. NiakanKalhori, ToleSutikno, Imam Much IbnuSubroto, and GoliArji [9]	C5.0, Neural Network, SVM, and K-Nearest Neighborhood (KNN)	13	They have observed that decision tree is an outperformer, but the accuracy result of ANN is very low.
Asha Rajkumar, and G. Sophia Reena [10]	KNN, Naïve Bayes, and Decision List	14	The accuracy (below 55%) of these algorithms is very low.
M. Anbarasi, E. Anupriya, and N.Ch.S.N Iyengr [11]	Genetic Algorithm, Decision Tree (DT), Naive Bayes, and Clustering	13	For effective heart disease prediction important attributes such as age, sex, resting ECG, fasting blood sugar, cholesterol, the slope of the peak exercise ST segment is totally ignored.
Sellappan Palaniappan and RafiahAwang [12]	ANN, Naïve Bayes and DT	15	A prototype-based system, where Neural Network is a lower performer than Naïve Bayes and Decision Trees.
K. Srinivas, Dr.G.Raghavendra Rao, and Dr. A.Govardhan [13]	Decision Trees, Naïve Bayes and Neural Network	15	ANN is a low performer than Decision Tree and Naive Bayes.
Yanwei Xing, JieWang, Zhihong Zhao, and YonghongGao [14]	SVM, ANN, and Decision Tree	11	They have used survival data for 11 attributes (e.g. TNF, IL6, IL8, HICRP, MPO1, TNI2, Sex, Age, Smoke, Hypertension, and Diabetes) as any occurrence of CHD.
HninWintKhaing [16]	K-means clustering	14	The author has proposed clustering based MAFIA (Maximal Frequent Itemset Algorithm) approach, but more efficient prediction system has not considered.
Atul Kumar Pandey, Prabhat Pandey,K.L Jaiswal, and Ashish Kumar Sen [19]	Decision Tree (pruned, un-pruned, and reduced error pruning approach)	14	In this model have used different approaches of Decision Tree, where fasting blood sugar is not provided good accuracy.
Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, and M Essam Khalifa, [24]	Fast decision tree and pruned C4.5 tree	14	The authors have observed different datasets with 14 attributes.
Mai Shouman, Tim Turner, and Rob Stocker [26]	K-Nearest Neighborhood (KNN)	13	The authors have implemented only KNN algorithm with voting and without voting approach, but have not considered larger dataset and efficiency.
Markos G. Tsipouras, Dimitrios I. Fotiadis, Katerina K. Naka, and Lampros K. Michalis [27]	Decision Tree and Fuzzy Model	19	This study has implemented for significant or high-risk CAD patients, not for all.

Currently, ML is rapidly growing attention to the research community, where ML based method can produce reliable results, learn from earlier pattern and predict coronary heart disease status. We have mentioned various related ML algorithms, reference no., accuracy, precision, recall, number of features, and dataset [9]-[26] in Table- II.

Here, most of the researchers conducted UCI dataset with 14 or 15 attributes including ML algorithms for the prediction of Coronary Heart Diseases (CHDs) by mining data from the different patterns of heart at different situations.

Consequently, we are finding most reliable machine learning algorithms with high precision to train and test our model for predicting the CAD.

Table- II: Findings of various previous studies [9]-[26] incorporating ML algorithms.

Techniques/ Methods		Reference No.	Accuracy (%)	Precision (%)	Recall (%)	Number of Features	Dataset
Supervised Learning (SL)	*Unsupervised Learning (USL)*						
Decision Tree/List	-	[9]	93.02	90.90	-	14	UCI
		[10]	52.00	48.55	48.97	14	Sensing Data
		[11]	99.20	99.78	99.78	13	Sellapanetal
		[12]	89.00	-	-	15	Cleveland,UCI
		[13]	89.00	-	-	14	APRHI Data
		[14]	89.60	-	-	11	Survival Data
		[15]	72.93	82.60	82.20	14	Cleveland,UCI
		[20]	92.50	92.50	92.50	14	UCI
Naïve Bayes	-	[10]	52.33	-	-	14	Sensing Data
		[11]	96.50	96.60	96.80	13	Sellapanetal
		[12]	86.53	-	-	15	Cleveland,UCI
		[13]	83.53	-	-	14	APRHI Data
		[18]	82.31	93.10	85.70	14	UCI
		[20]	91.20	92.30	91.20	14	UCI
Random Forest	-	[20]	88.70	-	-	14	UCI
SVM	-	[9]	86.05	89.47	-	14	UCI
		[14]	92.10	-	-	11	Survival Data
		[20]	68.80	72.70	68.80	14	UCI
ANN	-	[9]	80.23	83.78	-	14	UCI
		[12]	85.53	-	-	15	Cleveland,UCI
		[13]	83.00	-	-	14	APRHI Data
		[14]	91.00	-	-	11	Survival Data
KNN	-	[9]	88.37	88.09	-	14	UCI
		[10]	45.67	54.79	45.52	14	Sensing Data
		[26]	97.40	-	-	13	UCI
-	Clustering	[11]	88.30	83.26	95.20	13	Sellapanetal
-	k-mean based	[16]	74.00	78.00	67.00	-	Cleveland,UCI
Decision Tree	-	[18]	84.35	86.20	97.20	14	UCI
		[19]	75.73	-	-	14	UCI
Bagging	-	[18]	85.03	86.10	98.40	14	UCI

III. METHODOLOGY

Angiogram, a standard test for identifying the CAD and costly diagnosis, is mostly unavailable in rural area like Bangladesh creates urgencies to predict model without angiographic features, where we have collected patient's clinical data from cardiologist report. We have used 12 most important clinical features in this model, which are responsible for heart diseases to avoid coronary angiogram hassle and to predict CAD. Earlier most of the researchers conducted UCI dataset for 14 features shown in the Table- III including angiographic disease status and applied different data mining techniques [8]-[27], however in our proposed research we have applied most important 12 clinical features for the NICVD dataset excluding angiographic status without compromising the accuracy of prediction by thinking the financial status of rural people of low incoming countries.

Table- III: Previous mostly used clinical features collected from UCI repository [7].

SL No.	Features Name	Description
1	Age	Instance age in years (33-72)
2	Sex	Instance gender (1 = male; 0 = female)
3	CP	Chest pain type (1= typical angina, 2=atypical angina, 3=non-angina pain,4=asymptomatic)
4	RBP	Resting blood pressure (Systolic in mmHg on admission to the hospital), normal 120/80 mm HG
5	Cholesterol	Total serum cholesterol in mg/dl (<200 mg/dl)
6	FBS	Fasting blood sugar > 120 mg/dl (1= T; 0 = F)
7	ECG	Resting electro cardio graphic results, results (0=normal, 1=ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2 = showing probable or definite left ventricular hypertrophy)
8	HR	Maximum heart rate achieved
9	EIangina	Exercise induced angina (1 = yes, 0 = no)
10	Old peak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment (value1= up sloping, 2=flat, 3=down sloping)
12	Thal	1 = normal; 2 = defect; 3 = reversible defect
13	CA	Number of major vessels (0-3) colored by fluoroscopy (Angiographic disease status)
14	Num	Diagnosis of heart disease, Value 0 =< 50% diameter narrowing, 1 = > 50% diameter narrowing (Angiographic disease status)

For our research, we have discarded 2 angiographic features (Num and CA) from existing most used features of UCI heart disease dataset of 14 features, and have updated one feature 'Stenosis'- history of heart disease status (proposed in NICVD dataset) instead of feature 'Thal' (existing in UCI dataset).

The updated 12 clinical features' name and description for NICVD dataset are described in Table- IV. The proposed system comprises the following steps that are also shown in "Fig. 2".

- Collection of actual valid raw survey data from NICVD, Bangladesh.
- Development of a system to predict heart disease from cardiologist report using minimum features with low error.
- Analysis of the cardiac data using various machine learning methods/data mining techniques for better prediction and accuracy of cardiac status.
- Analysis of the performance of different datasets with different machine learning algorithms.
- Validation of the heart disease prediction result by medical professionals or cardiologist.

Table- IV: Name and description of clinical features of NICVD dataset [Proposed].

SL No.	Features Name	Description
1	Age	Instance age in years (30-74)
2	Sex	Instance gender (1 = male; 0 = female)
3	CPT	Chest pain type (CPT) (1= typical angina, 2=atypical angina, 3=non-angina pain, 4=asymptomatic)
4	RBP	Resting blood pressure (Systolic in mm Hg on admission to the hospital), normal 120/80 mm HG
5	Cholesterol	Total cholesterol in mg/dl (<200 mg/dl)
6	FBS	Fasting blood sugar (FBS)>120 mg/dl or 6.5mmol/L (1= T;0 =F)
7	ECG	Resting electrocardiographic (ECG) results, results (0=normal, 1=abnormal)
8	HR	Maximum heart rate (MHR)
9	Elangina	Exercise induced angina (Elangina) (1 = yes, 0 = no)
10	Old peak	ST depression/elevation induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment (value 1= up sloping, 2=flat, 3=down sloping)
12	Stenosis	1= normal, 2 = yes, 3 = reversible defect;

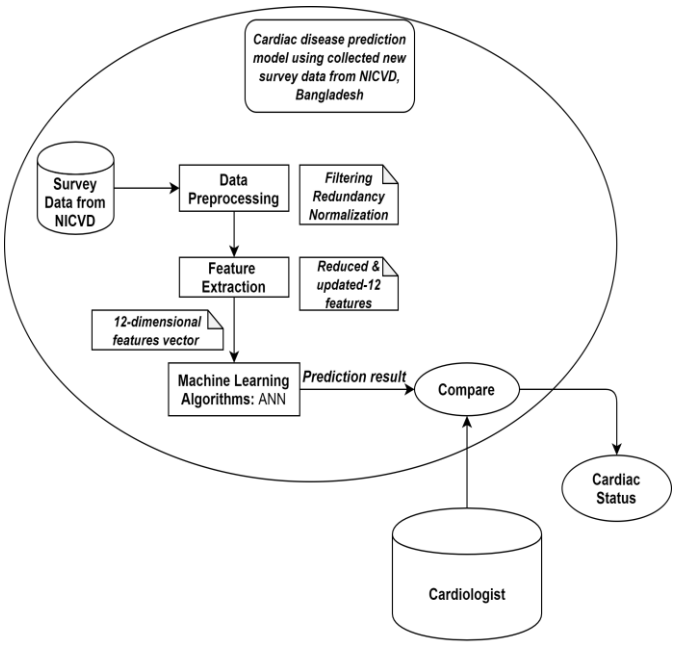

Fig. 2. Proposed model of this research.

In our experiments, 12 features from NICVD dataset are inserted into six (06) well known supervised machine learning algorithms: Decision Tree (J48), Naïve Bayes, Random Forest, SVM, ANN, and Logistic Regression for classifying and predicting heart disease based on these clinical features vector that is shown in "Fig. 3".

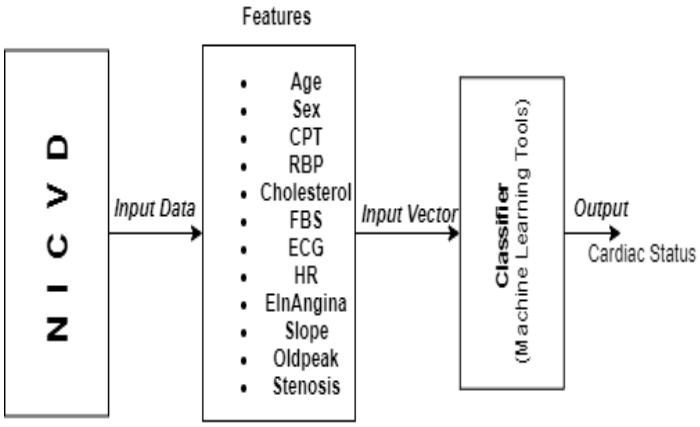

Fig. 3. Machine learning backed cardiac status system.

IV. EXPERIMENTS

A. Corpus Construction

In this study, we have conducted survey at NICVD, Bangladesh where the patients with stable angina were admitted to the hospital. Patients who are echocardiogram (ECG) and exercise tolerance test (ETT) positive were examined by cardiologist or cardiac doctor. It is observed that among 200 stable angina patients there were 132 male and 68 females. From the observation a survey questionnaire report for data collection is also prepared and where data are collected from the biomedical report, ECG and ETT that have been stored in our dataset. The data survey with collection process is shown in "Fig. 4".

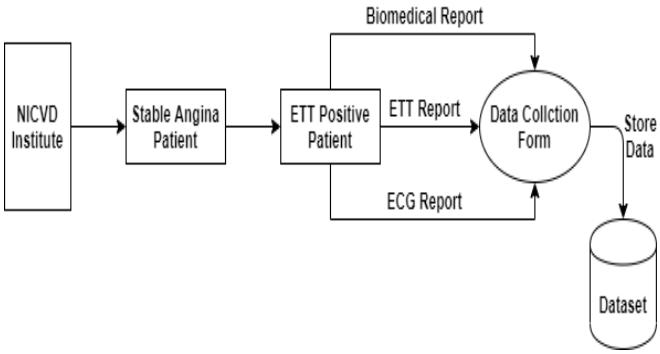

Fig. 4. Data survey and collection process.

In the 40 ETT+ (ETT positive) patients through survey questioners out of 200 stable angina patients in 6 (six) months from NICVD, Bangladesh are observed where 23 patients have heart disease and 17 patients have no heart disease. From ETT+ patients, 16 clinical features of coronary artery disease are collected and then it is reduced to 12 important clinical features manually, there are some collected data shown in Table- V.

The University of California, Irvine (UCI) with 14-dimensional data is collected where 303 patients with 164 instances are having no heart disease and 139 instances containing heart disease.

Table- V: Some sample of heart disease collected data from NICVD, Bangladesh.

Age	Sex	CPT	RBP	CholesT	FBS	ECG	HR	EIangina	Old Peak	Slope	Stenosis
50	0	3	140	220	0	0	155	0	1.4	2	0
59	1	1	130	222	0	0	110	1	0.5	3	3
47	1	4	120	231	1	2	127	1	1.8	3	1
60	1	1	110	212	1	2	160	1	0	1	2
31	1	1	140	198	2	2	127	0	1.4	3	0
50	1	4	190	170	1	0	136	0	2.6	3	1
63	1	3	130	204	0	0	140	1	0	2	0
62	1	4	110	226	0	2	127	1	1.6	3	1
43	0	3	110	153	0	0	156	1	1.8	3	0
55	0	4	140	232	1	0	139	0	1	3	1
64	1	3	130	330	0	0	155	0	0	1	0
63	0	4	145	220	0	2	117	0	0	2	0
39	1	3	130	245	0	0	182	0	3.2	3	0
54	1	4	135	223	1	2	161	1	3.1	3	1
57	1	3	130	250	1	2	137	1	0.8	2	1
49	1	2	120	230	0	0	152	1	1	3	2
59	0	4	140	320	0	2	162	0	0.4	1	1
61	1	4	140	330	0	2	115	1	0.6	3	3
58	1	4	120	221	1	0	158	1	1.6	1	1
44	0	2	110	202	0	0	171	0	1	2	0
73	1	2	150	277	1	0	169	0	0.5	1	0
58	1	4	145	288	0	0	118	1	0.8	3	3
51	1	1	120	159	0	2	145	0	0.6	1	0

B. Experimental Set up

Input feature Set for machine learning Tools:

i) 12 features without angiogram from NICVD dataset, Bangladesh [Proposed]
ii) 12 features without and 14 features with angiogram from UCI dataset, USA

Output status of machine learning Tools:

i) 1 (Positive or Cardiac disease exists)
ii) 0 (Negative or Cardiac disease not exists)

Total 40 patients in NICVD dataset:

i) 23 patients of heart disease
ii) 17 patients of no heart disease

Investigated ML tools:

i) Decision Tree (J48)
ii) Naïve Bayes
iii) Random Forest
iv) SVM
v) ANN
vi) Logistic Regression

Input data vector is normalized to get better accuracy for the classifiers implemented by supervised machine learning algorithms in the environment of MATLAB and WEKA 3.8. We have applied 10-fold cross validation on the datasets and applied leave one out cross validation (Jackknife method) to NICVD data for finding more accurate accuracy.

We have four Investigations for ML classifiers:

i) Train (36) +Test (4) on NICVD instances (12 features): 10-fold cross validation
ii) Train (39) + Test (1) on NICVD instances (12 features): Leave one out cross validation
iii) Train (272) + Test (31) on UCI instances (14 features): 10-fold cross validation
iv) Train (272) + Test (31) on UCI instances (12 features): 10-fold cross validation

The classifiers' name and various properties of machine learning classifiers described in Table- VI.

Table- VI: Properties of various machine learning classifiers.

SL No.	Classifiers Name	Properties of Classifiers
1	ANN	Input unit= 12, Hidden layer=2, Output layer=1, Activation Function: Sigmoid. Learning rate, η=0.01, Momentum coefficient, μ=0.9 Weight update, ΔWkj= η*ek(n)*xj(n), where error, ek (n)= desired output (dk) – (yk) output signal of neuron K at time n.
2	Decision Tree	max_depth=32, min_samples_leaf=0.1, min_samples_split=0.1 max_features=12
3	Naive Bayes	Normal (Gaussian) distribution, $\text{Posterior}, P(c\|x) = \frac{\text{Likelihood}P(x\|c) * \text{Prior}P(c)}{\text{Evidence}P(x)}$;
4	Random Forest	Total no. of trees=10
5	SVM	Kernel function: Linear, c=1.0; k(x,y)= xTy + c, where c is a tradeoff parameter between error and margin
6	Logistic Regression	Model: Linear with sigmoid function; logistic function(x)= $\frac{1}{1+e^{-(1+x_1+\cdots+x_{12})}}$

Classification, Precision and Recall are calculated for evaluating the classifiers using the following formula, where TP, FP and FN are True Positive, False Positive and False Negative respectively.

Classification Accuracy (CA) = (Correct classification/ Total Classification) * 100% …..... (1)

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{……..……..} \quad (2)$$

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

F1-Score = harmonic mean of Precision and Recall.

V. EXPERIMENTAL RESULTS & ANALYSIS

A. Experimental Result of NICVD Data

In the Investigation of NICVD data we have 10-fold cross validation & leave one out cross validation results where we have obtained results of CA, Precision, Recall, and F1 score by applying different supervised machine learning methods. In Table- VII we have specified the results of classifiers for NICVD instances using 10-fold cross validation. Classification accuracy of NICVD data for 10- fold cross validation has shown in "Fig. 5". In Table- VIII we have specified the results of classifiers for NICVD instances using leave one out cross validation.

- Train (36) +Test (4) on NICVD instances (12 features): 10-fold cross validation

Table- VII: Performance analysis of various supervised methods using NICVD dataset.

Techniques/Methods	CA	Precision	Recall	F1
Decision Tree (J48)	82.50%	85.20%	82.50%	82.60%
Naïve Bayes	85.00%	86.10%	85.00%	85.10%
Random Forest	75.00%	78.60%	75.00%	73.00%
SVM	75.00%	75.60%	75.00%	75.10%
ANN	**92.80 %**	**92.91%**	**92.80 %**	**92.79 %**
Logistic	77.50%	78.60%	77.50%	77.60%

Table-VII shows the investigated experiments' results using the different classifiers for classification accuracy (CA), Precision, Recall, and F1-score, where Train (36) +Test (4) on NICVD instances (12 features) where 10-fold cross validation model is used. From the experiments it is observed that the ANN provides the highest accuracy (92.80%), precision (92.91%), Recall (92.80%) and F1-score (92.79%) in comparison with the other classifiers such as Decision Tree (J48), Naïve Bayes, Random Forest, SVM, ANN, and Logistic Regression. The "Fig. 5" depicts the classification accuracy for investigated classifiers ANN, Decision Tree (J48), Naïve Bayes, Random Forest, SVM, ANN and Logistic Regression, and shows that the ANN provides higher accuracy.

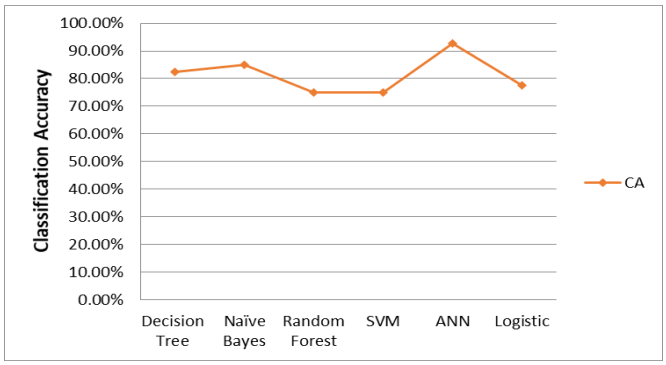

Fig. 5. CA graph of various ML algorithms using NICVD dataset.

- Train (39) + Test (1) on NICVD instances (12 features): Leave one out cross validation

Table- VIII: CA of ML methods using leave-one out method for NICVD dataset.

Techniques/Methods	CA
Decision Tree (J48)	71.00%
Naïve Bayes	75.10%
Random Forest	71.42%
SVM	75.00%
ANN	**84.80%**
Logistic	75.33%

The Classification Accuracy (CA) for the investigated experiments' results using the different classifiers for leave one out cross validation has been shown in Table- VIII for the NICVD data. It shows more precise results using 12 features and the ANN shows its superiority (84.80%) in comparison with other classifiers.

B. Experimental Result of UCI Dataset

In the Investigation of UCI dataset, 10- fold cross validation results for 14 and 12 features are obtained for CA, Precision, and Recall by applying different classifiers. Table-IX demonstrates the results of different classifiers using UCI data with 14 features including angiogram and 12 features without using angiogram with 10-fold cross validation. The comparison graph between 14 features and 12 features of UCI dataset has been shown in "Fig.6". It is observed that CA of UCI dataset using14 features and 12 features are approximately equal for the experiments Train (272) + Test (31) on UCI instances (14 features): 10-fold cross validation and Train (272) + Test (31) on UCI instances (12 features): 10-fold cross validation.

Table- IX: Comparison of performance between 14 features and 12 features using UCI dataset.

Method Name	CA (%)		Precision (%)		Recall (%)	
	14 features	*12 features*	*14 features*	*12 features*	*14 features*	*12 features*
Decision Tree (J48)	76.7	76.9	76.8	77.3	76.6	76.9
Naïve Bayes	86.5	86.5	86.6	86.3	82.6	86.2
Random Forest	69.6	67.3	69.9	67.3	69.6	67.3
SVM	76.8	76.3	77.4	76.3	76.9	76.3
ANN	93.5	91.7	93.8	91.6	93.7	91.7
Logistic Regression	67.7	67.0	67.6	67.1	67.7	67.0

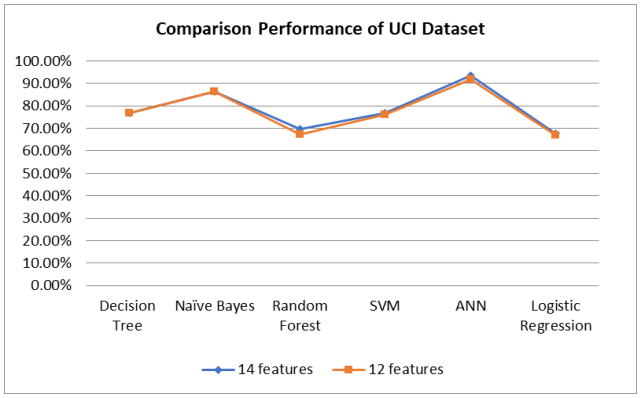

Fig. 6. Comparison of classification accuracy between 14 and 12 features of UCI dataset.

C. Comparison Result of NICVD and UCI Dataset

Comparison of classification accuracies for different classifiers between NICVD and UCI datasets for 10-fold cross validation has been shown in "Fig. 7" and it is observed that ANN performs better compared to other classifiers. The NICVD data in the context of Bangladesh involving noise in some extent makes the necessities of using the ANN in the system and hence it provides the best performance. It is also observed that the accuracy of NICVD instances with 12 features are slightly better in comparison with the accuracy of UCI dataset with 12 features, and NICVD dataset provides good results than UCI dataset for some classifiers.

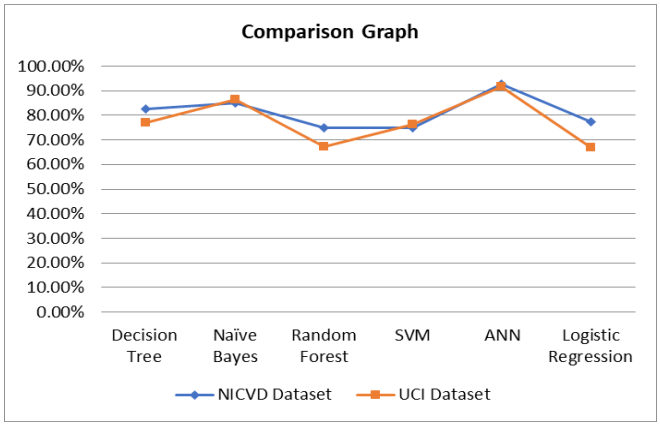

Fig. 7. Comparison graph of classification accuracy between NICVD and UCI dataset.

VI. CONCLUSION

In this research we have proposed intelligent prognosis of coronary artery disease excluding angiogram in patients with stable angina for different machine learning algorithms with 12 features. The paper concludes the following:

i) Two datasets are investigated in the experiments: NICVD and UCI. NICVD dataset is constructed by us with the help of cardiologist for 40 patients, where 23 patients have CAD and 17 have no CAD.

ii) Twelve features instead of 14 are evaluated by dropping high cost angiographic feature by thinking the financial status of rural people in low income countries like Bangladesh without compromising the performance.

iii) Different supervised machine learning algorithms such as Decision Tree (J48), Naïve Bayes, Random Forest, SVM, ANN, and Logistic Regression are incorporated in the investigated experiments, but the ANN shows better results.

iv) Though the UCI dataset with 14 features are slightly better in comparison with the NICVD dataset with 12 features, but high cost features like angiogram can be avoided.

v) UCI dataset with 12 features provides approximately equal performance compared to the NICVD dataset with 12 features. Sometimes, NICVD dataset with 12 features provide better results for some classifiers like Decision Tree, Random Forest, ANN and Logistic Regression.

vi) In low coming countries the general people will get best service for CAD detection without doing high cost diagnosis like angiogram.

vii) People do not need any time-consuming preparation that is required for angiogram.

VII. LIMITATIONS AND FUTURE WORK

We have collected data from NICVD, but the data collection procedure was not convenient in the country of Bangladesh because of few specialized cardiologists and lack of organizational infrastructure.

In future we would like to concentrate on three things:

- Preparation of a repository database based on Bangladeshi patients
- Preparation of an IoT enabled system to collect data and predict the result from machines ("Fig. 8").
- Development of a mobile based cardiac status application through cloud.

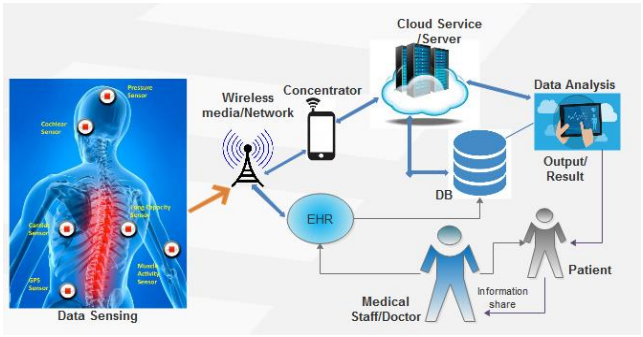

Fig. 8. IoT enabled cardiac disease prediction system.

We have done our experiments in a corpus of small scale. Therefore, many limitations are in our experimental process. Some significant limitations are mentioned below.

- Small scale corpus for NICVD (only 40)
- 14 features for NICVD not collected
- Smoking and Genetic features have not considered
- Troponin -I test feature not considered
- Deep learning not incorporated because of small scale training and test data.

ACKNOWLEDGMENT

We would like to thank Professor Dr. Md. Afzalur Rahman, Director of National Institute of Cardiovascular Disease (NICVD), Associate Professor Dr. Abdul Momen and Dr. Farzana Sultana, Registrar of NICVD, Bangladesh for their assistance and support in this research.

REFERENCES

1. Wong, N. D., "Epidemiological studies of CHD and the evolutionpreventive cardiology",Nature Reviews cardiology,Vol. 11, pp. 276–289,California, Irvine, CA 92697, USA, 2014.
2. Wu R, Peters W, Morgan MW, "The next generation clinical decision support: linking evidence to best practice," Journal of Healthcare Information Management, Vol. 16, No. 4, USA, 2002.
3. Anbarasi M, Anupriya E, and Iyengar NCHSN, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," International Journal of Engineering Science and Technology, Vol. 2, No.10, 2:5370-76, 2010.
4. Palaniappan S, Awang R., "Intelligent heart disease prediction system using data mining techniques," International Journal of Computer Science and Network Security, Vol.8, No.8,8:343-50, August 2008.
5. Cord, Matthieu, Cunningham, Padraig (Eds.), "Machine Learning Techniques for Multimedia Case Studies on Organization and Retrieval," Springer, XVI, 289 p, hardcover, 2008.
6. National Institute of Cardiovascular Diseases, Bangladesh. Available: http://www.nicvd.gov.bd.

7. UC Irvine Machine Learning Repository. Available: https://archive.ics.F.edu/ml/datasets/heart+Disease.
8. E.P.Ephzibah, Dr. V. Sundarapandian, "Framing Fuzzy Rules using Support Sets for Effective Heart Disease Diagnosis," International Journal of Fuzzy Logic Systems (IJFLS), Vol.2, No.1, February 2012.
9. Moloud Abdar, Sharareh R. NiakanKalhori, ToleSutikno, Imam Much IbnuSubroto, GoliArji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," IJECEVol. 5, No. 6, pp. 1569 – 576, ISSN: 2088-8708, December 2015.
10. Asha Rajkumar, G. Sophia Reena, "Diagnosis of Heart Disease Using Data Mining Algorithm," Global Journal of Computer Science and Technology, Page 38, Vol. 10 No. 10, Ver. 1.0, September 2010.
11. M. Anbarasi, E. Anupriya, and N.Ch.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," International Journal of Engineering Science and Technology, Vol. 2, No. 10, 2010.
12. SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," International Journal of Computer Science and Network Security (IJCSNS), Vol. 8, No. 8, August 2008.
13. K.Srinivas, Dr.G.RaghavendraRao, and Dr. A.Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," International Conference on Computer Science and Education (ICCSE), pp. 1344 - 1349, 2010.
14. Yanwei Xing, JieWang, Zhihong Zhao, and YonghongGao "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease," IEEE Transactions on Convergence Information Technology, pp. 868 – 872, November 2007.
15. VikasChaurasia and Saurabh Pal, "Early Prediction of Heart Diseases using Data Mining Techniques," Caribbean Journal of Science and Technology, ISSN 0799-3757, Vol. 1, pp. 208-21,2013.
16. Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data," International Conference on Computer Research and Dvelopment (ICCRD), 2011.
17. Max Bramer, "Principles of Data Mining," 2nd ed. Springer, ISSN1863-7310, 2013.
18. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases," International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 4, pp 56-66, 2013.
19. Atul Kumar Pandey,Prabhat Pandey,K.L. Jaiswal, and Ashish Kumar Sen, "A Heart Disease Prediction Model using Decision Tree",Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p- ISSN 2278-8727,Vol. 12, No. 6, pp. 83-86, August 2013.
20. ThendralPuyalnithi and V. MadhuViswanatham, "Preliminary Cardiac Disease Risk Prediction Based on Medical and Behavioural Data Set Using Supervised Machine Learning Techniques",Indian Journal of Science and Technology, Vol. 9, No. 31,August 2016.
21. ShaiShalev-Shwartz, Shai Ben-David, "Understanding Machine Learning," Cambridge University Press, New York, NY 10013-2473, 2014.
22. Introduction to Support Vector Machines [online]. Open-CV Documentation; 7 March 2017. Available: http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html
23. Ethan Alapydin, "Introduction to Machine Learning," 2nd ed. Cambridge Massachusetts, MIT Press, London, 2010.
24. Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, and M.EssamKhalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets," International Conference on Communication, Management and Information Technology (ICCMIT), 2015.
25. AKM Monwarul Islam, AKM Mohibullah, and Timir Paul "Cardiovascular Disease in Bangladesh: A Review," Bangladesh Heart Journal, Vol. 31, No. 2, July 2016.
26. Mai Shouman, Tim Turner, and Rob Stocker "Applying k-Nearest Neighbor in Diagnosing Heart Disease Patients," International Conference on Knowledge Discover (ICKD), Singapore, 2012.
27. Markos G. Tsipouras, Dimitrios I. Fotiadis, Katerina K. Naka, and Lampros K. Michalis "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling,"IEEE Transactions on Information Technology in Biomedicine, Vol. 12 , Issue: 4, July 2008 .

AUTHORS PROFILE

Md. Shah Jalal Jamil, was born in Bhola, district of Bangladesh. **Mr. Jamil** is working as a Lecturer in the Department of Computer Science and Engineering (CSE), Sonargaon University (SU), Bangladesh. He has received M.Sc. degree in Computer Science & Engineering (MSCSE) from United International University, Bangladesh with outstanding result (4.00 out of 4.00). Also, he has completed his B.Sc. degree in CSE program from United International University (UIU), Bangladesh. He has received award and scholarship from UIU. On the other hand, He has 3 international research articles on the field of Artificial Intelligence and Machine Learning. Recently he has submitted another three research papers in the related fields to SCOPUS indexed Journal and IEEE conference. He specially works on Machine Learning, Natural Language Processing (NLP), and Data Mining. Moreover, he is acting as the organizing secretary, Bangladesh Private University Parishad.

Dr. A.K.M. Muzahidul Islam, was born in Shariatpur district of Bangladesh. He has received M.Sc. in Computer Science and Engineering from Kharkiv National University of Radio Electronics, Ukraine, in 1999 and Dr.Eng. in the field of Computer Science and Engineering from Nagoya Institute of Technology, Japan in 2007. He has received Japanese Government Monbusho Scholarship (October 2002 – March 2006). From January 2011 until January 2017, Dr. Muzahid has served as a Senior Lecturer at Malaysia-Japan International Institute of Technology (MJIIT) of Universiti Teknologi Malaysia (UTM), Malaysia. Currently he is a Professor in the CSE department of United International University (UIU), Bangladesh. Dr. Muzahid has published over 75 international research publications (including 21 peer-reviewed Indexed Journals and over 50 Conference Papers). Dr. Muzahid has served as as the Program Chair of ICAICT 2016 Int'l Conference, Bangladesh held in Chittagong in May 2016. He has also served as the Secretariat of ICaTAS 2016 Int'l Conference, Malaysia and the 7TH AUN/SEED-Net 2014 Int'l Conference on EEE. He was an Advisory Member of ICBAPS2015, Malaysia and also played as the Communication Track Chair at ICIEV15, ICAEE 2015, and ICAEE 2017 Int'l Conferences. Dr. Muzahid is serving as an Associate Editor in the International Journal of Computer and Information Technology (IJCIT). He is a Chartered Engineer (CEng), Senior IEEE Member (SMIEEE) and a Fellow IEB (FIEB).

Bulbul Ahamed, was born in 1982 at Munshiganj, Bangladesh. He was graduated from Northern University Bangladesh (NUB) in Computer Science and Engineering (CSE). He has completed his MBA in MIS and Marketing from the same university. He also had his M.Sc. in CSE from United International University (UIU), Bangladesh. He has earned a couple of prestigious awards (Chancellor's Gold Medal & Summa Cum Laude) in recognition of his excellent academic achievements. Now he is pursuing his PhD in Computer Science and Engineering at Jahangirnagar University, Bangladesh. He is now working as an Associate Professor in the Department of Computer Science and Engineering at Sonargaon University (SU), Bangladesh. He had previously worked as Assistant Professor, Senior Lecturer and Lecturer in the Department of CSE at NUB. He also worked as trainer in Partners in Learning (PiL) program, Microsoft Bangladesh Ltd. He has research interest in speech recognition, pattern recognition, artificial intelligence, e-commerce, business and social issues. He has published his articles in prestigious journals and conferences in home and abroad. In addition, he has published one text book on "Information and Communication Technology (ICT) for class XI & XII in Bengali version" Approved by National Curriculum and Textbook Board (NCTB), Bangladesh.

Dr. Mohammad Nurul Huda, was born in Lakshmipur, district of Bangladesh. **Dr. Huda** is working as a Professor and Director of MSCSE Program in the Department of Computer Science and Engineering (CSE), United International University (UIU), Bangladesh. He completed his PhD degree from Toyohashi University of Technology, Aichi, Japan on Automatic Speech Recognition (ASR). He graduated from Bangladesh University of Engineering and Technology (BUET) in Computer Science and Engineering (CSE) department. He specially works on Machine Learning and Natural Language Processing (NLP). He has more than 140 international research articles in the related fields (http://cse.uiu.ac.bd/profiles/mnh/: Publications). Among them more than 78 are SCOPUS indexed articles. Recently, he was invited at the UNESCO Headquarter, Paris, France for the Language Technology for All (LT4All) conference, where NLP experts from 88 countries in the world were invited. Moreover, he is acting as the director, NLP and AI (Artificial Intelligence) of eGeneration Ltd, a renowned Software company in Bangladesh.