Heart Block Prediction using Data mining and Machine Learning

Risul Islam Rasel[1], Phayung Meesad[2]
[1]Chittagong Independent University, Bangladesh
[2]King Mongkut's University of Technology North Bangkok, Thailand
risul@ciu.edu.bd; phayung.m@it.kmutnb.ac.th

Anupam Chowdhury[3], Meherab Hossain[3]
[3]International Islamic University Chittagong, Chittagong-4318, Bangladesh
anu8chy@gmail.com; h.meherabhossain@gmail.com

***Abstract*— Heart block occurs when the flow of electricity interrupted or partially delayed between the top chamber and bottom chamber of the heart. People are now more often affected by this kind of disease. However, early prediction of heart block can reduce the diagnosis complexity and treatment cost. In this study, a data mining and machine learning model is proposed to predict three types of heart blocks, such as 1st degree A-V block, Left Bundle Branch Block (LBBB), and Right Bundle Branch Block (RBBB). Experiment data samples are collected from the cardiology department of Chittagong Medical College Hospital (CMCH), Bangladesh. The dataset contains 32 types of numeric and categorical features about the patient's ECG report, daily activities, and food habits. The prediction model is designed, trained, and tested with some empirical machine learning algorithms namely Decision Tree, Random Forest, K-Nearest Neighbor, and Support Vector Machine. Finally, the experimentation shows that Decision Tree and Random Forest models outperform the other algorithms in overall heart block prediction with an accuracy of more than 92%.**

Keywords- Heart block prediction; 1st degree A-V block; Left Bundle Branch Block (LBBB); Right Bundle Branch Block (RBBB.

I. Introduction

In human history, people have been affected by life-threatening diseases. Cardio Vascular Diseases (CVDs) are the number one cause of death globally because more people die annually from CVDs than from any other cause. The World Health Organization has estimated that over 56.9 million deaths worldwide and 17.9 million people died from CVDs in 2016 which is 31% of all global deaths. On the whole, it is regarded as the primary reason behind deaths in adults [1].

The heart is the second most important organ in the human body after the brain. [3] If people suffer from any kind of Cardiovascular disease, they cannot lead their normal life and become mentally depressed. If the disease is heart block, then it becomes more object of thought. Heart block patients may have a high risk of heart attacks and other CVDs and lead to death. Heart block can be defined as a blockage in the conduction of the normal electrical impulses in the heart. Heart block occurs from degeneration or scarring of the electrical pathways in the heart muscle, either naturally or as a result of disease [2].

The purpose of this study is to build a prediction model to predict three different types of heart blocks. We have used patients' ECG (Electrocardiogram) reports' data, daily activities' data, and food habits' data to prepare a combined experimental dataset for our experimentation purpose. These data are collected from Chittagong medical college hospital (CMCH), one of the largest public hospitals in Bangladesh.

II. Literature Review

Palaniappan and Awang [1] had developed a Web-based prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining and machine learning techniques, namely, Decision Trees, Naive Bayes, and Neural Network. They showed in their study that IHDPS can answer complex "what if" queries which traditional decision support systems cannot. Their developed prototype can predict the likelihood of patients getting heart disease using patients' medical profiles such as age, sex, blood pressure, and blood sugar.

Ordonez *et al.* [5] used association rule mining and decision tree techniques for heart disease prediction. They had undertaken a comparative analysis between these two algorithms and showed that rules produced from decision trees were not as powerful as association rule mining. But the decision tree had at least 50% prediction accuracy and generally above 80% accuracy for binary classification. On the other hand, association rule mining needed very specific trial and error runs to find an acceptable threshold, which was time-consuming.

Nahar *et al.* [4] had used association rule mining with the apriori algorithm for factor analysis of heart disease. The rules were separated for both male and female patients. They had also used a predictive apriori algorithm to increase the expected accuracy of an association rule. It is because; predictive apriori takes confidence and support both criteria for analyzing a dataset, whereas apriori makes the rules only based on the confidence.

Srinivas *et al.* [6] had presented an intelligent and effective heart attack prediction method in their work. They had applied Rule-based, Decision tree, Naïve Bayes, and Artificial Neural Network for extracting meaningful patterns from the patient's healthcare dataset for early prediction of a heart attack. For data preprocessing and effective decision making they had used One Dependency Augmented Naïve Bayes classifier (ODANB) and Naive Credal Classifier 2 (NCC2).

Barhatte *et al.* [7] had worked on the QRS complex of the heart patients. They had used a support vector machine approach to classify Left bundle branch block, Right bundle branch block, and premature ventricular contraction. They applied the Gaussian Radial Basis Function kernel (RBF) and scaling factor sigma for tuning their model with appropriate parameters.

Ali *et al.* [8] had proposed and investigated a Hybrid Multilayered Perceptron (HMLP) network for the classification of bundle branch block arrhythmias. Five morphological features were then extracted through the median threshold method from patients' ECG data and these features were used for training a single hidden layer HMLP network. Four different learning algorithms were applied to train their model with four hidden nodes in a single hidden layer.

Kelwade and Salankar [9] proposed a model to predict eight types of arrhythmias such as Normal Sinus Rhythm (NSR), Premature Ventricular Contraction (PVC), Right Bundle Branch Block (RBBB), Left Bundle Branch Block (LBBB), Atrial Fibrillation (AF), Sinus bradycardia (SB), second degree Block (B) and Atrial flutter (A) using Particle Swarm Optimization (PSO) and Multi-Layer Perceptron (MLP). Some non-linear and linear parameters such as Largest Lyapunov Exponent (LLE), Spectral Entropy (SE), Hurst exponent (H), SD1/SD2 ratio, normalized Low Frequency (nLF) and High Frequency (nHF) were used in their experiments which were obtained from eight types of heart arrhythmias. The experiment dataset had acquired from the MIT-BIH arrhythmia database.

Shouman, Turner, and Stocker [10] had investigated the decision Trees algorithm applying a range of techniques such as the Gini Index, Information Gain, discretization techniques, voting method, and pruning in heart disease diagnosis. The dataset was collected from Cleveland that contained 76 attributes with 303 numbers of rows of information. However, only 13 attributes from all were selected to use in their experiment. The highest accuracy they had achieved was 84.1% from the equal frequency discretization Gain Ratio Decision Tree.

Ozcift [11] had investigated a correlation-based feature selection technique to select pertinent features from cardiac arrhythmia and Random Forests (RF) ensemble classifiers to improve the diagnosis performance of cardiac arrhythmia. His experimental dataset contained 452 samples in fourteen types of arrhythmias and eleven of these classes had sample sizes less than 15 samples. Finally, he recommended a data resampling strategy to improve the diagnosis of cardiac arrhythmia using the random forest algorithm.

Chaurasia *et al.* [12] had used three popular data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision tree (DT) to develop the prediction models using a large dataset. They used 10-fold cross-validation methods to measure the unbiased estimate. According to their work CART classifier had the potential to significantly improve the conventional classification methods results.

Ahmed, Elsayed, and Syed [13] had proposed an automated Framingham scoring model to classify the early risk factors of hard coronary heart diseases. They had used K-Nearest Neighbor and Random Forests algorithms to design their experimental model and compared the accuracy. Their experimental dataset was collected from a hospital of Saudi Arabia that contained attributes about patients' age, cholesterol level, High-Density Lipoproteins (HDL), Systolic Blood Pressure (SBP), treatment for hypertension, smoking status and diabetes.

Bialy, Salama, and Karim [14] had built an intelligent decision support system for heart disease diagnosis using majority vote based ensemble classification methods such as stacking, bagging and boosting. Two types of the dataset such as patients' heart disease dataset and heart sound signal dataset were collected and used to train and test the six classification algorithms for having comparatively better accuracy for any kind of heart disease data.

Srinivas, Rao, and Govardhan [15] had presented an effective way of heart attack prediction using some predictive data mining techniques such as decision tree, neural network, Bayesian model, and support vector machine. Different types of data pre-processing techniques like wavelet transformation, random projection, principal component analysis (PCA) were used as feature reduction and selection techniques. Their experimental dataset had contained 13 input attributes to perform binary classification. From their work they had shown that neural network can perform better than the decision tree, Bayesian model, and support vector machine.

Kelwade, and Salankar [16] had suggested methodologies to predict cardiac arrhythmias classes such as Left Bundle Branch Block (LBBB), Atrial Fibrillation (AFIB), Normal Sinus Rhythm (NSR), Right Bundle Branch Block (RBBB), Sinus Bradycardia (SBR), Atrial Flutter (AFL), Premature Ventricular Contraction (PVC), and Second-degree block (BII) using artificial neural networks (ANNs). The networks were trained and tested using both linear and nonlinear features of the dataset. They had shown that radial basis function neural network (RBFN) outperformed traditional Multi-layer perceptron (MLP) neural network with an accuracy of 95.9%.

III. METHODOLOGY

A. Data collection and Preprocessing

We have collected a total of 208 patients' data with 32 attributes from the Cardiology department of Chittagong Medical College. Three types of information such as the patient's food habit, daily activities, and ECG data are gathered to prepare a combined dataset. Food habit data and daily activities data are collected through interviewing the patients' and ECG data are compiled from patients' ECG graph reports. Table I shows the attributes' list of our experimental dataset. Most of the attributes are categorical. So, we convert the categorical data into numeric data for making the dataset convenient for computation.

TABLE I. DESCRIPTION OF DATASET'S ATTRIBUTES

Sl. No.	Attributes	Descriptions
1	Age	Young = less than or equal 30 yrs., Middle = 31-49 yrs., Old = equal or greater than 50 yrs.
2	Sex	Gender of patients
3	Work	Activity of occupation
4	Smoking	Smoking level
5	ChTobacco	Chewing Tobacco
6	BpSys	Systolic blood pressure
7	BpDias	Diastolic blood pressure
8	Pulse	Pulse rate
9	HTN	Hypertension
10	Diabetes	Diabetes Mellitus
11	Angina	Chest pain (CP)
12	Cough	Cough/Squeezing
13	Salt Intake	Extra Salt in daily food
14	R.A	Rheumatoid arthritis
15	Dyspnea	Shortness of breath
16	Kidney Prob.	Kidney Disease
17	Heart Problem	Other heart diseases
18	Anemia	Less Blood Cell
19	Exercise	Walking regularly
20	Rice	Quantity per day
21	Bread	Fried/ Normal
22	Read meat	Beef, Mutton
23	Interpolation	Hear Rate variation
24	AxisDev.	Axis Deviation
25	T-wave	Normality
26	DCM	Dilated
27	ST slope	ST-segment
28	Pulse	Ranges from 60 to 100 beats per minute for general people
29	Troponin	used for heart muscles damages checking
29	QRS	used for Bundle Branch Block
31	PR	used for A-V block checking
32	Block	Heart block types

Table II shoes the sample ratio of our experimental dataset. A total of 208 numbers of patients' information are collected. Among them, 142 are male patients (68% of total data), and 66 are female patients (32% of total data). The Male and Female patients' ratio is not considered to be equal or likely to be equal as we concentrate on finding the heart blocks for general patients; for both males and females. But, the sample size or the number of patients is the same for each type of class.

TABLE II. SAMPLE RATIO OF THE TOTAL DATASET

Class	Male	Female	Total
No Block	30	26	56
1stDegree A-V Block	35	18	53
LBBB	40	14	54
RBBB	37	8	45
Total	=142 (68.26%)	=66 (31.74%)	=208

Three types of information such as patients' daily activities, food habits, and ECG data are combined to create the experimental dataset. The first two type's data are collected through interviewing the patients directly and ECG data is collected from patients' ECG reports. Data pre-processing is done to make the raw data meaningful and useful for the further experimental process. Next, the dataset is divided into two groups; training dataset and test dataset. 70% data of each class are considered to train the model and the rest of the 30% of each class data is used to test the model. The sampling of these training data and teat data is done with a meticulous cross-validation process. Four types of computational intelligent algorithms such as decision tree, K-nearest neighbor, Support vector machine, and random forest (ensemble learning) are used to build the training model. All the models are trained with optimized model parameters. The cross-validation approach is used to identify the optimized values of models' parameters for each model. Finally, the outcomes from each model are evaluated by applying test data to propose the best possible model for predicting heart blocks.

IV. EXPERIMENT RESULTS

TABLE III to TABLE VI show the test results of our experiments for the decision tree, random forest, k-nearest neighbor, and support vector machine models consecutively.

TABLE III. HEART BLOCK PREDICTION RESULTS FOR THE DECISION TREE MODEL

	No Block	1st Degree Block	LBBB	RBBB	Class precision	Average prediction Accuracy
No Block	21	0	1	0	1.00	
1st Degree Block	0	17	0	0	1.00	
LBBB	0	0	12	2	0.80	0.92
RBBB	0	0	2	8	0.80	

TABLE IV. HEART BLOCK PREDICTION RESULTS FOR THE RANDOM FOREST MODEL

	No Block	1st Degree Block	LBBB	RBBB	Class precision	Average prediction Accuracy
No Block	22	0	0	0	1.00	

1st Degree Block	0	16	0	1	1.00	0.92
LBBB	0	0	12	2	0.86	
RBBB	0	0	2	8	0.73	

TABLE V. HEART BLOCK PREDICTION RESULTS FOR THE K-NEAREST NEIGHBOR MODEL

	No Block	1st Degree Block	LBBB	RBBB	Class precision	Average prediction Accuracy
No Block	17	0	0	0	0.74	0.74
1st Degree Block	1	17	0	0	0.85	
LBBB	3	1	6	5	0.75	
RBBB	2	2	2	7	0.58	

TABLE VI. HEART BLOCK PREDICTION RESULTS FOR THE SUPPORT VECTOR MACHINE MODEL

	No Block	1stDegree Block	LBBB	RBBB	Class precision	Average prediction Accuracy
No Block	17	0	2	0	0.89	0.79
1stDegree Block	2	20	0	0	0.95	
LBBB	0	1	6	2	0.43	
RBBB	0	0	6	7	0.78	

V. CONCLUSION

The contribution of this work is that we have collected and analyzed a very high dimensional dataset with 32 attributes. Three types of information such as patients' food habits, daily activities, and ECG report data are merged to prepare the experimental dataset. No other previous work had analyzed such high dimensional data to predict heart block. We have designed and trained a heart block prediction model using four different types of machine learning algorithms such as decision tree, random forest, K-nearest neighbor, and support vector machine. The reason behind choosing these four types of algorithms is to compare and test the prediction performance of ensemble learning, lazy learning, and as well as computational intelligent learning. After a very rigorous and meticulous analysis, we have found that ensemble learning outperforms the others in categorical analysis, and produce very good results in heart block prediction. In this work, we only concentrate on three types of major heart blocks such as 1st degree A-V block, Left Bundle Branch Block (LBBB), and Right Bundle Branch Block (RBBB). Based on this work further research can be done to predict more heart block types other than these three types of heart blocks.

REFERENCES

[1] Palanappan, S., Awang, R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques. In: Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications, pp. 108-115. IEEE Computer Society Washington, DC, USA (2008).

[2] Soni , J., Ansari, U., Sharma, D., Soni, S.: Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications 17(8), 43-48 (2011).

[3] Ordonez, C., Omiecinski, E., Braal, L., Santana, C. A., Ezquerra, N., Taboada, J. A., Cooke, D., Krawczynska, E., Garcia, E. V.: Mining Constrained Association Rules to Predict Heart Disease. In: Proceedings IEEE International Conference on Data Mining, pp. 1-7. IEEE, San Jose, CA, USA (2001).

[4] Ordonez, C.: Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction. IEEE transactions on information technology in biomedicine 10(2), 334-343 (2006).

[5] Nahar, J., Imam, T., Tickle, K. S., Chen, Yi-Ping P.: Association rule mining to detect factors which contribute to heart disease in males and females. Article in Expert Systems with Applications 40, 1086–1093 (2013).

[6] Srinivas, K., Rani, B.K., Govrdhan, A.: Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering 2(2), 250-255 (2010).

[7] Alka S. B., Ghongade, R., Abhishek S.T.: QRS Complex Detection and Arrhythmia Classification using SVM. In: International Conference on Communication, Control and Intelligent Systems (CCIS), pp. 239-243. IEEE, Mathura, India (2015).

[8] Megat Ali, M. S. A., Jahidin, A. H., Norali, A. N.: Hybrid Multilayered Perceptron Network for Classification of Bundle Branch Blocks. In: International Conference on Biomedical Engineering (ICoBE), pp. 27-28, IEEE, Penang, Malaysia (2012).

[9] Kelwade, J. P., Salankar, S. S.: Training of Multilayer Perceptrons with Improved Particle Swarm Optimization for the Heart Diseases Prediction. Int J Swarm Intel Evol Comput 6(2), 1-8 (2016).

[10] Shouman, M., Turner, T., Stocker, R.: Using Decision Tree for Diagnosing Heart Disease Patients. In: Proceedings of the 9-th Australasian Data Mining Conference, pp. 23-29. ustralian Computer Society, Inc. Ballarat, Australia (2016).

[11] Ozcift, A.: Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Computers in Biology and Medicine 41, 265–271, (2011).

[12] Chaurasia, V., Saurabh P.: Early Prediction of Heart Diseases Using Data Mining Techniques. Caribbean Journal of Science and Technology 1, 208-217 (2013).

[13] Ahmed, H., Elsayed, G., Syed, L.: An Automatic Early Risk Classification of Hard Coronary Heart Diseases using Framingham Scoring Model. In: Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing, pp. 1-8. ACM, NY, USA (2017).

[14] El Bialy, R., Mostafa A. S., Omar, K.: An ensemble model for Heart disease data sets: a generalized model. In: Proceedings of the 10th International Conference on Informatics and Systems, pp. 191-196, ACM, NY, USA (2016).

[15] Srinivas, K., Rao, G. R., Govardhan, A.: Survey on prediction of heart morbidity using data mining techniques. International Journal of Data Mining & Knowledge Management Process 1(3), 14-34 (2011).

[16] Kelwade, J. P., Salankar, S. S.: Comparative study of neural networks for prediction of cardiac arrhythmias. In: International Conference on Automatic Control and Dynamic Optimization Techniques, pp. 1062-1066, IEEE, Pune, India (2016).