ENSEMBLE OF MULTIPLE MODELS FOR ROBUST INTELLIGENT HEART DISEASE PREDICTION SYSTEM

Md. Jamil-Ur Rahman
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
jamilruet13@gmail.com

Rafi Ibn Sultan
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
rafi.ruet13@gmail.com

Firoz Mahmud
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
fmahmud.ruet@gmail.com

Ashadullah Shawon
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
shawonashadullah@gmail.com

Afsana Khan
Department of Computer Science and Engineering
Military Institute of Science and Technology
Dhaka, Bangladesh
afsana29khan@gmail.com

***Abstract*— Recently heart disease has become the most common fatal diseases in the world. Early stage detection and treatment can reduce the number of cardiac failures, mortality of heart disease and cost of diagnosis. The healthcare industry collects a huge amount of these medical data, but unfortunately, these are not mined. Discovery of hidden patterns and relationships from this data can help effective decision making to predict the risk of heart disease. The main objective of this research is to develop a Robust Intelligent Heart Disease Prediction System (RIHDPS) using some classification algorithms namely, Naive Bayes, Logistic Regression and Neural Network. This article reviewed the effectiveness of clinical decision support systems by ensemble methods of these three algorithms.**

***Index Terms*—research objective; data source; related work; naïve Bayes algorithm; logistic regression algorithm; neural network; ensemble of multiple models; experimental analysis**

I. INTRODUCTION

Heart disease is the number one cause of death and holds the title of the most concerning problem in the US. According to the American Heart Association, in the US one in every seven people die because of heart disease. Heart disease is responsible for more than 360,000 deaths in a year. Close to 790,000 people in the US are diagnosed with heart attacks in every year and about 114,000 of them die eventually [1].

Although medical science is at the peak of its advancement there still lies a major problem when it comes to giving a good quality service at an affordable cost. Patients in the queue waiting to be treated want both quality service and affordable cost while doing so. Thus, comes the challenge of both of these requirements to be fulfilled. This paper is proposing a data mining technique that would be enough to solve such alarming matter.

Quality service means diagnosing a patient's current problems and conditions both correctly and accurately. Accuracy is the most important term discussed here as well as the reduced cost. According to the data trend, the cost of quality medical care is not reasonable and the cost will even increase in the upcoming years [2]. A poor analysis of the patient can lead to tragic results which are not acceptable by any means. Decisions based on this analysis are dependent on doctors' skills, intuition and experience regarding the situation. Proper use of the patient database can reduce mistakes as well as extravagant medical costs. Analysis of the patterns of data hidden in the patient medical history may produce knowledge which can be mined from the database.

Currently, most of the medical intuitions already use some kind of information storage system to store, manage and update their patient data i.e. medical history [3]. These data mainly contain the patient name, date of birth, checks, charts, images, results of different tests etc. But these data are only used in for keeping records, some normal decision taking and nothing else. For analyzing and taking intelligent clinical decisions, these data are not used. So, this huge amount of data is quite wasted in a sense. Where with proper data mining of this information, the information can be used to take important and clinical medical decisions just by analyzing the patterns of the data. Thus, for some situations saving us the time and money of going through complex and expensive tests.

Wu, et al presented that rational automated decisions taken from computer-based patient records could help in many sections such as reducing human errors, increasing patient safety in critical operations and ensuring a better patient outcome overall [4]. This type of data analysis tool i.e. data mining technique can be used to create such an environment that can take clinical life-and-death automated decisions. Technically a collection of organized creates a database. The main purpose of database is that the data can be accessed conveniently for updating, managing or replacing at any time. From the definition of Fayyad, data mining is a process of extraction of potentially useful and previously unknown patterns of data from a given dataset [5]. Data mining extracts hidden patterns and relationships from the huge database. It is

a combination of statistical analysis, machine learning, data visualization and expert systems. The method can work on this great amount of data retrospectively in an automated manner. Traditional statistical methods require a definite number of variables. Data mining can include new variables at any time and work with a lot more variables than any previous way [6].

II. Research Objectives

Data mining works with two methods: supervised and unsupervised learning. A training dataset exists in the supervised learning method so that the algorithm has a prior knowledge of how the outputs can be, whereas in unsupervised learning no training set exists to learn parameters in the model [7]. The goals of data mining modeling are both classification and prediction of class labels where classification models classify categorical values from different classes and prediction models predict functions of continuous values [8]. Methods like decision Trees and Neural Networks apply classification algorithms where methods like Regression, Association Rules and Clustering apply prediction algorithms [9].

The primary goal of this research is to establish a Robust Intelligent Heart Disease Prediction System (RIHDPS) using ensemble method of three data mining modeling techniques such as Naïve Bayes, Logistic Regression and Neural Network. RIHDPS can mine useful hidden knowledge such as patterns and associations relating to heart disease from a patient heart disease database. This may help doctors to take clinical unbiased decision accurately by providing them with an extra assistance by answering some complex questions. Effective treatments will not only ensure better medical diagnosis but will also help to reduce cost.

III. Data source

The Cleveland Heart Disease database consists of 303 records with 14 medical attributes relating to heart attack [10]. Using this database, patterns related to the prediction of heart attack were extracted. A training dataset of 200 records and a testing dataset of 103 records were created by splitting the main source into these two sections. In figure 1 displays the framework of the heart disease prediction system proposed. The medical data was preprocessed, the missing data in the records were handled [11]. The samples for every dataset were selected in a random process to keep away from creating bias. Then the train set was employed to learn parameters which were used in the test set to predict the patient's condition.

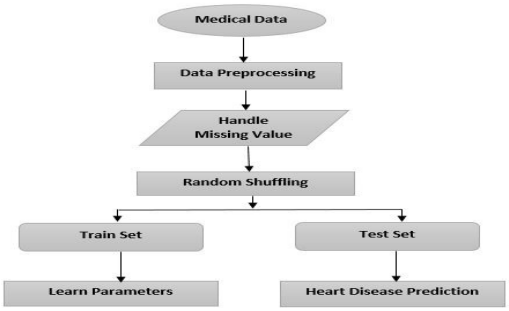

Fig. 1. The framework of Heart Disease Prediction System

III-A. Attributes in the dataset

Features in the dataset are explained below:
1. Age (Year)
2. Sex (Male (1); Female (0))
3. Chest Pain Category (typical type 1 angina (1); typical type angina (2); non-angina pain (3);
value 4: asymptomatic)
4. Blood Pressure (mm Hg)
5. Serum Cholesterol (mg/dl)
6. Blood Sugar (before eating/ fasting) (greater than 120 mg/dl (1); less than 120 mg/dl (0))
7. Electrocardiographic resting measure (normal (0); showing ST-T wave abnormality (1); showing probable or definite left ventricular hypertrophy (2))
8. Heart rate (maximum)
9. Exercise affected angina (yes (1); no (0))
10. Exercise affected ST depression
11. Slope of the ST segment (in time of peak exercise) (unsloping (1); flat (2); down sloping (3))
12. Intensity of colored major vessels by fluoroscopy (0-3)
13. Thal (normal (3); fixed defect (6); reversible defect (7))
14. Diagnosis (less than 50% diameter (heart disease absent); value 1: greater than 50% diameter (heart disease present))

Here, "Diagnosis" acts as the output of values "1" and "0". Where "1" denotes patients diagnosed with heart disease and "0" denotes patients diagnosed with no heart disease. Figure 2 and figure 3 illustrates the box plot and basic statistics respectively of the numeric features which are age, blood pressure, serum cholesterol, Heart rate (maximum), Exercise effected ST depression (oldpeak), Intensity of colored major vessels by fluoroscopy (CA).

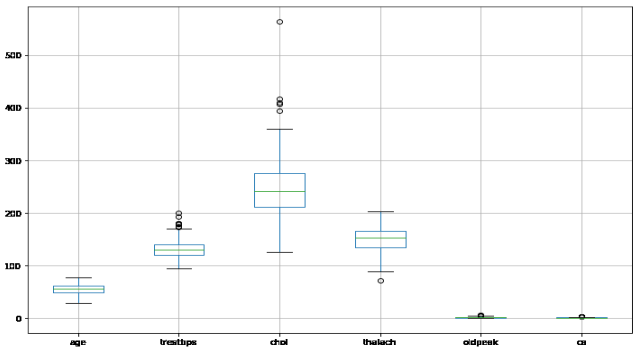

Fig. 2. Boxplot of the numeric features

	age	trestbps	chol	thalach	oldpeak	ca
count	303.000000	303.000000	303.000000	303.000000	303.000000	299.000000
mean	54.438944	131.689769	246.693069	149.607261	1.039604	0.672241
std	9.038662	17.599748	51.776918	22.875003	1.161075	0.937438
min	29.000000	94.000000	126.000000	71.000000	0.000000	0.000000
25%	48.000000	120.000000	211.000000	133.500000	0.000000	0.000000
50%	56.000000	130.000000	241.000000	153.000000	0.800000	0.000000
75%	61.000000	140.000000	275.000000	166.000000	1.600000	1.000000
max	77.000000	200.000000	564.000000	202.000000	6.200000	3.000000

Fig. 3. Basic statistics of the numeric features

In table 1 and figure 4 illustrate the class instances taken for training and testing data respectively. The limitations of the dataset are the total number of samples of both the training and test dataset. This total number is comparatively low as the main source is a small dataset. The dataset classes are unbalanced because of not having the same number of instances. It may introduce bias in the prediction.

TABLE I. NUMBER OF CLASS INSTANCES OF TRAINING AND TESTING

Classes	No. of Records	Training data	Test Data
Heart Disease	139	85	54
No Heart Disease	164	115	49
Total	303	200	103

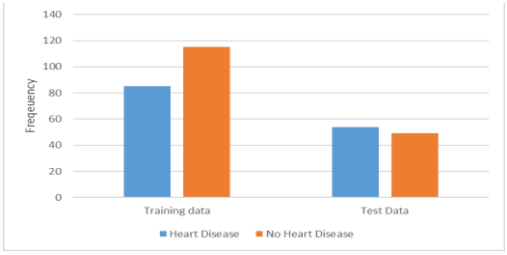

Fig. 4. Number of class instances of training and testing

IV. RELATED WORK

As heart disease prediction is a major challenge, a lot of methodologies have been developed. Palaniappan, et al proposed an intelligent heart disease prediction system by utilizing different data mining methods i.e. naïve Bayes, decision tree, and neural network [12]. Das, et al proposed an effective diagnosis of heart disease through the ensemble of neural networks [13]. In [14] neural network based heart disease prediction was shown.

V. NAÏVE BAYES ALGORITHM

In probability theory and statistics, Bayes' theorem is the description of the probability of an event, based on prior that might be related to the event [15]. According to Bayes theorem, considering two random variables A and B, if conclusion is B and A is the given evidence/observation, where B and A has a dependence relationship exists between them. we can write:

$$P(B|A) = \frac{P(A,B)}{P(A)} = \frac{P(A|B).P(B)}{P(A)};$$

$$P(A) = P(A,B) + P(A,\bar{B}) = \sum_{V_i \in B} P(A|B = V_i)P(B = V_i)$$

Where V is the elements of variable B. We can write,

$$P(Y = V_i|X_1, X_2 \dots X_n) = \frac{P(X_1, X_2 \dots X_n|Y = V_i)P(Y = V_i)}{P(X_1, X_2 \dots X_n)}$$

$$= \frac{P(X_1, X_2 \dots X_n|Y = V_i)P(Y=V_i)}{\sum P(X_1, X_2 \dots X_n|Y = V_j)P(Y=V_j)}$$

Here, $P(Y = V_i|X_1, X_2 \dots X_n)$ is a posteriori probability. And, $P(X_1, X_2 \dots X_n|Y = V_j)$ is a proprietary probability [16]. The denominator is a normalization factor. So, in the optimization problem, this can be neglected. It is equivalent to:

$$\arg\max P(X_1, X_2 \dots X_n|Y = V_i)P(Y = V_i) \quad (1)$$

Now if we estimate a priori probability using naïve distributor estimator it would be a naïve Bayes classifier.

The naïve Bayes classifier is a simple model of a Bayes classifier. In this classifier, an assumption is made that all the attributes are independent (statistically) of each other and contribute equally to the decision. This is known as the "naïve Bayes assumption". Even though this presumption can never be true in a real-life situation, naïve Bayes performs classification tasks considerably well [17]. So, for naïve Bayes theorem the equation (1) will become:

$$\arg\max P(X_1, X_2 \dots X_n|Y = V_i)P(Y = V_i)$$

$$= P(X_1|Y = V_i)(X_2|Y = V_i) \dots (X_n|Y = V_i)P(Y = V_i)$$

$$= \arg\max P(Y = V_i) \prod_{j=1}^{m} P(Xj|Y = V_i) \quad (2)$$

Because of the independence assumption, this paper works with the naïve Bayes classifier. This eases the computational complexity of learning the model, as the dataset is huge.
In our problem,

$$P(diagnosis = yes|evidence) = \frac{P(evidence|diagnosis = yes)P(diagnosis = yes)}{P(Evidence|diagnosis = yes)P(diagnsosis = yes) + P(evidence|diagnosis = no)P(diagnossi = no)}$$

As an example,
'Age' = 60, 'Sex' = male, 'Chest Pain Type' = asymptomatic, 'Blood Pressure' = 127 mm Hg, 'Serum Cholesterol' = 210 mg/dl, 'Fasting Blood Sugar' = <120 mg/dl, 'Electrocardiographic resting measure' = showing probable or definite left ventricular hypertrophy, 'Heart rate (maximum)' = 125, 'Exercise effected angina' = yes, 'Exercise effected ST depression' = 1.25, 'Slope' = flat, 'CA' = 2, 'Thal' = reversible defect.

$$P(diagnosis = yes|evidence) = 0.9994$$
$$P(diagnosis = no|evidence) = 1 - 0.9994 = 0.0006$$

Here our threshold value was 0.5 as $P(diagnosis = yes|evidence) \geq 0.5$ so we can say for this problem the prediction of diagnosis is 'yes'.

VI. Logistic regression algorithm

In statistics, logistic regression is a statistical method for evaluating a dataset. One or more independent variables in the dataset are used to evaluate an outcome. The outcome is determined as a dichotomous variable which can have only two outcomes [18].

In our dataset the output variable "diagnosis", $y \in \{0,1\}$. Where 0 denotes no heart disease (negative class) and 1 denotes heart disease (positive class). Here y can be either 0 or 1, but we are using linear regression where the hypothesis $(h_{\boldsymbol{\theta}}(x))$ can have values much larger than 1 or less than 0. In logistic regression, we want to satisfy $0 \leq h_{\boldsymbol{\theta}}(x) \leq 1$ where $h_{\boldsymbol{\theta}}(x)$ is a hypothesis function that maps input variables(x's) to output variable(y).

For linear regression, the hypothesis is

$$h_{\boldsymbol{\theta}}(x) = \theta^T x$$

For logistic regression, the hypothesis $h_{\boldsymbol{\theta}}(x)$ needs to be modified as

$$h_{\boldsymbol{\theta}}(x) = g(\theta^T x)$$

Where g is called the sigmoid or logistic function [19] can be defined as

$$g(z) = \frac{1}{1+e^{-z}}$$

So, our hypothesis becomes

$$h_{\boldsymbol{\theta}}(x) = \frac{1}{1+e^{-(\theta^T x)}}$$

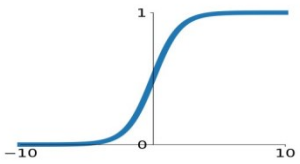

Fig. 5. Sigmoid function

Figure 5 shows the sigmoid function mapping any real value into another value between 0 and 1.

If $h_{\boldsymbol{\theta}}(x) = 0.7$ the model would tell the patient that he/she has the 70% chance of having heart disease.

$$h_{\boldsymbol{\theta}}(x) = p(y = 1|x; \boldsymbol{\theta})$$

Where the probability of $y = 1$, given x is parameterized by $\boldsymbol{\theta}$.

From this, we can get,

$$P(y = 0|x; \boldsymbol{\theta}) = 1 - P(y = 1|x; \boldsymbol{\theta})$$

In our model, if $h_{\boldsymbol{\theta}}(x) \geq 0.5$ or $\theta^T x \geq 0$, it will predict that $y = 1$.

And if $h_{\boldsymbol{\theta}}(x) \leq 0.5$ or $\theta^T x < 0$, it will predict that $y = 0$.

In our problem, there are 13 input variables.

So, it becomes:

$$h_{\boldsymbol{\theta}}(x) = g(\boldsymbol{\theta_0} + \boldsymbol{\theta_1 x_1} + \boldsymbol{\theta_2 x_2} + \cdots + \boldsymbol{\theta_{13} x_{13}})$$

If $(\boldsymbol{\theta_0} + \boldsymbol{\theta_1 x_1} + \boldsymbol{\theta_2 x_2} + \cdots + \boldsymbol{\theta_{13} x_{13}}) \geq 0$ it will predict $y = 1$.

Else it will predict $y = 0$.

$(\boldsymbol{\theta_0} + \boldsymbol{\theta_1 x_1} + \boldsymbol{\theta_2 x_2} + \cdots + \boldsymbol{\theta_{13} x_{13}})$ is our decision boundary.

Need to fit parameter $\boldsymbol{\theta}$'s so that cost function

$$J(\boldsymbol{\theta}) = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} cost\left(h_{\boldsymbol{\theta}}\left(x^{(i)}\right), y^{(i)}\right)$$

becomes minimum. Where m denotes the total number of training examples.

$$cost(h_{\boldsymbol{\theta}}(x), y) = -ylog\left(h_{\boldsymbol{\theta}}(x)\right) - (1 - y)\log(1 - h_{\boldsymbol{\theta}}(x))$$

If $y = 1$ then $cost(h_{\boldsymbol{\theta}}(x), y) = -log\left(h_{\boldsymbol{\theta}}(x)\right)$

And if $y = 0$ then $cost(h_{\boldsymbol{\theta}}(x), y) = -log\left(1 - h_{\boldsymbol{\theta}}(x)\right)$

By gradient descent algorithm we can minimize the cost.

Which repeatedly and simultaneously updates all θ_j.

Where [20]

$$\theta_j = \theta_j - \boldsymbol{\alpha} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} (h_{\boldsymbol{\theta}}\left(x^{(i)}\right) - y^{(i)}) x_j^{(i)}$$

Where $\boldsymbol{\alpha}$ is learning rate and $(x^{(i)}, y^{(i)})$ is i^{th} training data.

VII. Neural network

To develop an artificial intelligence, researchers created a highly interconnected system by using the combination of simple computing elements which mimics the working method of a human brain. At first, researchers were trying to imitate the neurophysiology of the bran [21] and these elements acted as the neurons of a bran. As the modern neural networks now are incorporated with different numerical analysis methods, they can make predictions about different real-world problems [13]. The multilayer perceptron (MLP) is the most used in the supervised neural network [19] [21]. It consists of three or more layers of neurons which are one input layer, at least one hidden layer and an output layer producing the classification results shown in figure 6.

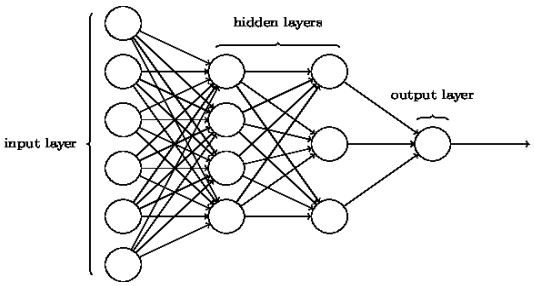

Fig. 6. The architecture of a typical neural network

In our experiment, a neural network was used where the input layer has 13 neurons as the database has 13 attributes. The model includes two hidden layers of 1000 neurons each and in the output layer, two output nodes represent "no heart disease" and "has heart disease" respectively. To train our model the weights were initialized using Xavier initialization [22] and Adam optimizer with standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) [23]. For regularization dropout [24] of probability 0.5, mini-batches of size 16 and an initial learning rate (LR) of 0.001 were used. The LR decays by a factor of 10 when the validation loss doesn't decrease for consecutive two epochs.

VIII. Ensemble of Multiple Models

Ensemble of multiple models is the process of averaging two or more related classifiers' outputs to predict a single outcome. In the majority of cases, it improves the accuracy as well as the generalization performance of data mining applications [13]. In our approach, we produced an output by ensemble of three classifiers i.e. Naïve Bayes, logistic regression and neural network shown in figure 7.

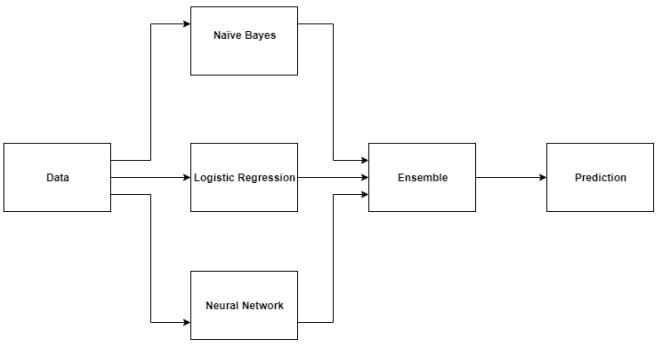

Fig. 7. Ensemble of multiple models

IX. Experimental Analysis

Naïve Bayes, logistic regression and neural network based ensemble method was used to evaluate the proposed system performance.

Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. Table 2 shows a classification matrix.

TABLE II. Classification Matrix

Predicted	1(Actual)	0(Actual)
1	True Positive	False Positive
0	False Negative	True Negative

From Table 2 we can get precision, recall and F1 score. Here:

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (3)$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (4)$$

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} \quad (5)$$

Table 3 shows the classification matrix of different methods that were used.

TABLE III. Classification Matrix of the Methods

	Naïve Bayes (Threshold 0.5)	**Naïve Bayes (Threshold 0.4)**	**Logistic Regression**	**Neural Network**	**Proposed Ensemble RIHDPS**
Actual =1, Predicted =1	46	47	43	47	48
Actual =1, Predicted =0	8	7	12	7	6
Actual =0, Predicted =1	4	4	4	3	3
Actual =0, Predicted =0	45	45	44	46	46

The main target is to detect each and every patient with heart problem so a further pathological investigation can begin. The goal should be increasing the true positive and decreasing the false negative. To achieve this, we decreased the threshold value from the initial value of 0.5 to 0.4 shown in the third column. Figure 8 shows logistic regression decision boundary using principal component analysis (PCA). PCA has been used to plot high dimensional data into a 2D surface [25]. From the figure, we can conclude that the logistic regression cannot classify properly because it uses a linear decision boundary.

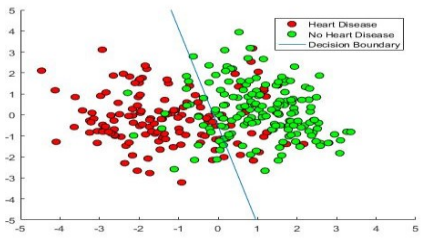

Fig. 8. Decision boundary for logistic regression

Table 4 shows the experimental results of the heart disease dataset where Naïve Bayes (Threshold 0.4), Logistic Regression and Neural Network were used to ensemble the proposed method which gave an accuracy of 91.26%.

TABLE IV. EXPERIMENTAL RESULTS OF THE HEART DISEASE DATASET

	Naïve Bayes (Threshold 0.5)	Naïve Bayes (Threshold 0.4)	Logistic Regression	Neural Network	Proposed Ensemble RIHDPS
Precision	0.92	0.92	0.91	0.94	0.94
Recall	0.85	0.87	0.78	0.87	0.89
F1 Score	0.88	0.88	0.84	0.90	0.91
Accuracy	88.35%	89.32%	84.47%	90.2%	**91.26%**

Comparing with [12] we developed a more robust intelligent heart disease prediction system using ensemble of models. Instead of neural network ensembles [13] [14] we used three different models for this classification task. Our proposed method produced an accuracy of 91.26% which is more than the previous accuracy of 87% [13].

X. CONCLUSION

This paper proposed an ensemble Robust Intelligent Heart Disease Prediction System (RIHDPS) of naive Bayes, logistic regression and neural network. The system analyses and extracts hidden knowledge relating to heart attack from patient heart disease history. This model can predict if a patient has heart disease or not accurately just by analyzing his/her medical data. As a result, it reduces the cost and time of the heart disease detection and it can assist the doctors in decision making. Our further goal is to develop improved methods for decision making by exploring other data mining algorithms and working with greater medical data so the resulting accuracy can be increased.

REFERENCES

[1] "Heart Disease and Stroke Statistics 2017 At-a-Glance" *American Heart Association*, 2017.

[2] "Medical Cost Trend Behind the Numbers 2018", July 25, 2017, The Alliance, *PricewaterhouseCoopers LLP*, 2017.

[3] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[4] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management. 16(4), 50-55, 2002.

[5] Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases", Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[6] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[7] Chaovalit, P. and Zhou, L., 2005, January. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on* (pp. 112c-112c). IEEE.

[8] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[9] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[10] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heartdisease/,2004.

[11] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[12] Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

[13] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36.4 (2009): 7675-7680.

[14] Rani, K. Usha. "Analysis of heart diseases dataset using neural network approach." arXiv preprint arXiv:1110.2626 (2011).

[15] Subbalakshmi, G., K. Ramesh, and M. Chinna Rao. "Decision support in heart disease prediction system using naive bayes." Indian Journal of Computer Science and Engineering (IJCSE) 2.2 (2011): 170-176.

[16] Diez, D.M., Barr, C.D. and Cetinkaya-Rundel, M., 2012. *OpenIntro statistics* (pp. 89-93). CreateSpace.

[17] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." In *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41-48. 1998.

[18] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[19] Hecht-Nielsen, Robert. "Theory of the backpropagation neural network." Neural networks for perception. 1992. 65-93.

[20] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

[21] Bishop, Chris, and Christopher M. Bishop. Neural networks for pattern recognition. Oxford university press, 1995.

[22] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010.

[23] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[24] Srivastava, Nitish, et al. "Dropout: A simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014): 1929-1958.

[25] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2, no. 4 (2010): 433-459