# An Adaptive Heart Disease Behavior-Based Prediction System

O. E. Emam[1], A. Abdo[2], Mona. M. Mahmoud[3]
Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*—Heart disease prediction is a complex process that is influenced by several factors, including the combination of attributes leading to the possibility of heart disease and availability of these attributes in the database, an accurate selection of these attributes and determining the priority and impact of each of them on the prediction model, and finally selecting the appropriate classification technique to build the model. Most of the previous studies have used some heart disease symptoms as major risk factors to build a heart disease prediction system leading to inaccurate prediction results. The main objective of this study is to build an Adaptive Heart Disease Behavior-Based Prediction System (AHDBP) based on risk factors and behaviors that may lead to heart disease. Different classification algorithms will be deployed to get the most accurate results. 18 attributes were used to build the prediction system. The accuracy of the classification techniques was as follows: Decision Tree 90.34%, Naive Bayes 91.54%, and Neural Networks 94.91%. Neural networks can predict heart disease better than other techniques. The Chi square method has also been applied to determine the difference between the expected and the observed results, and the proposed system proved its accuracy at 86.54%.**

***Keywords*—Chest pain; risk factors; coronary; cholesterol; neural networks; decision tree; naive Bayes**

## I. INTRODUCTION

Heart disease is a serious disease that leads to death; the World Health Organization (WHO) announced that more than 12 million people die globally due to coronary diseases every year [1]. There are different types of ad categories of heart disease, such as coronary, cardiovascular and cardiomyopathic. There are large numbers of factors that influence heart performance; smoking, hypertension and obesity rank among the most important factors. Heart disease is associated with functional problems of the heart such as irregular heart rhythms, which increase the risk of heart attack occurrence and other heart problems [2].

The diagnosis of heart disease depends on a complex interaction between the patient's medical data and the doctor's experience in diagnosing the type of heart disease. This combination affects the quality of the offered medical care [3]. Misdiagnosis may harm the patient's health and is associated with financial and moral burdens. Medical patient data is rich with hidden information that not seen by the doctor, but it can used to improve heart diseases diagnosis [4].

Data mining is not just using database software or tools. It is about building a suitable mining model and structure, which can be used to process, identify, and build the needed medical and clinical information [5]. Data mining processes the data from different perspectives and collects the knowledge from the data set. The outputs are useful and used in many different applications such as health care. The basic challenges faced by data mining in medicine are the accuracy of diagnosis and the ability to provide effective treatment to the patient [6].

There is a significant difference between the symptoms of heart disease and the factors leading to it because the symptoms of the disease consist of a set of signs that the person exhibits when affected by heart disease. These symptoms can be used as attributes to build a system that determines the type and degree of a person's heart disease. The factors that lead to the occurrence of heart disease (risk factors) represent a group of diseases that affect the person or some parts of their behavior, which in turn lead to the person suffering from heart disease; these risk factors can used as attributes to build a system that can predict heart disease. Smoking, for example, is a conduct that increases the risk of heart disease and not every smoker with heart disease, while Chest pain is a symptom of heart disease. A patient with a heart disease has a constant feeling of pain in the chest so chest pain cannot be a factor that is used to predict heart disease.

In the period 2012 to 2017, most of the studies in this area used some symptoms of heart disease to build a heart disease predictive system. Building such a prediction systems based on symptoms led to major faults in the results of previously proposed systems. In addition, the efficiency of previous systems has not been tested to determine how accurate the results are.

Therefore, the main objective of this work was to build an Adaptive Heart Disease Behavior-Based Prediction System (AHDBP) using different classification algorithms, to identify and correct all the symptoms of heart disease used in previous studies in this field, and to validate the results using the suitable measuring standards. In this research, a set of new risk factors attributes based on WHO reports for 2015 are used to build the prediction system. The data set is classified by three basic classification techniques: Decision Tree, Naive Bayes and Neural Networks. The accuracy of the system is tested by different evaluation techniques. The results showed

that neural network outperformed other types when new attributes were added. The proposed prediction framework is designed to help doctors in heart diseases prediction, where the accuracy of heart diseases will be improved by using neural network.

The paper is structured as follows: First, we discuss the related work in Section II. This is followed by a description of the clinical data and the phases of proposed framework in Section III. The experimental results are discussed in Section IV. We conclude our paper in Section V and give an outlook to the future work.

## II. Literature Review

Many researchers have developed heart disease prediction system. Most of this research has used heart symptoms attributes instead of risk factors attributes, and some attempts have added one or two attributes in order to improve the accuracy of the prediction system. In what follows, we discuss the previous studies that have contributed to heart disease prediction, focusing on research that has added heart symptoms attributes and their impact on system accuracy.

Authors in [7] authors designed a heart disease prediction system using a group of data classification techniques (Decision Tree, Naïve Bayes). They used data from the "Cleveland Clinic Foundation Heart Disease Dataset" composed of 13 attributes and 303 instances. The system was designed and implemented using Weka and IBM SPSS Modeler. Results showed that the accuracy of the mining techniques was as follows: Decision Trees 79%, Naive Bayes 83.7%.

Authors in [8] developed prediction systems for heart disease based on three data mining classification techniques (Decision Trees, Naive Bayes, and Neural Networks). They added two new attributes (obesity and smoking). The data was analyzed and interpreted using Weka data mining software. The results showed that the accuracy of the three mining techniques using 13 attributes was as follows: Decision Trees 96.66%, Naive Bayes 94.44%, and Neural Networks 99.25%, while 15 attributes gives: Decision Trees 99.62%, Naive Bayes 90.74%, and Neural Networks 100%. The results showed that the Neural Networks predicts heart disease with the highest accuracy.

Authors in [9] used medical datasets to build an automatic heart diseases prediction system based on different Neural Networks technique. The historical data consists of 14 attributes and 414 instances; 13 attributes and a class attribute the input layer contains 13 neurons to represent 13 attributes. The data was analyzed and interpreted using Weka data mining software. The results show that the accuracy was about 82.90%.

Authors in [10] improved a Heart Disease Prediction System (HDPS) based on Artificial Neural Networks (ANN). The dataset was composed of 303 samples and the number of attributes was 13. System designed using a C- programming language to execute heart disease classification and prediction ANN. The accuracy of the prediction techniques was nearly 80%.

Authors in [11] were developed heart disease prediction systems based on different classification techniques (Decision Trees, Naive Bayes, and Neural Networks). The data consists of 3000 instances, where 70% of the data is used for training the model and the remaining 30% for testing it. The system was developed using 15 attributes. Only 13 attributes were used for prediction, then two new attributes (obesity and smoking) were added to test the accuracy of heart disease diagnosis. The data was analyzed and processed using Weka 3.6.6 data mining tools. The result of analysis using 13 attributes is as follows: Artificial Neural Network 85.53%, Decision Tree 89%, Naive Bayes 86.53%. When using 15 attributes, the results were: Artificial Neural Network 100%, Decision Tree 99.62 %, Naive Bayes 96.5%. The results showed that Neural Networks predict Heart disease with the highest accuracy.

Authors in [12] developed a Heart disease prediction with a comparison of three basic classification algorithms (Neural Networks, Naive Bayes, and Decision Trees) for implementation in health care. They collected data of 305 instances and 14 attributes. Comparing the results of the three used algorithms, Neural Networks perform better than the other algorithms: Neural Networks 82%, Decision Tree 80% and Naïve Bayes 81%.

Authors in [13] applied four basic classification algorithms: Bayes Net, Decision Tree (J48), Naive Bayes and Genetic Algorithm, to build an heart attack prediction system from a patient dataset obtained from medical practitioners. The data used for the analysis were collected by medical practitioners in South Africa and consisted of 11 attributes. Data analysis, discovery and prediction patterns developed by WEKA data mining software results showed that the decision tree (J48) predicts heart diseases with a higher accuracy than other techniques: J48 99.0%, Naïve Bayes 97.22%, Bayes Net 98.14%, and genetic algorithm 99.07%.

Authors in [14] used four different classification data mining techniques (J48 Decision Tree, K Nearest Neighbors, Naïve Bayes and SMO) in order to improve heart disease prediction accuracy. The dataset set was collected from a hospital in Iran which composed of 209 samples with 8 attributes used for Prediction. The data were analyzed and tested using the WEKA data mining software. The results of the J48 decision tree, Naive Bayes, K Nearest Neighbors and SMO were compared, and the results showed that the best classification accuracy is 83.73% achieved by J48 decision tree with medical data set, while K Nearest Neighbors yielded 82.775%, Navy Bayes 81.818%, and SMO 82.775%.

After discussing the previous studies in the field of predicting heart disease, we noted the following:

The employed frameworks can be used only for dedicated classification algorithms, and is not adaptable with other types of classification algorithms.

Most studies have used the same number of attributes and some added one or two to improve the efficiency of the system, while other approaches tried to use new classification techniques that rely on the old data and same attributes. The

number of attributes used was insufficient to improve the accuracy and efficiency of the prediction system.

Most of previous researches did not use suitable evaluation methods to test the efficiency of the system and determine the accuracy of the results.

Most of previous researches used heart symptoms as risk factors in building heart disease prediction system, which is a totally insufficient way to predict a disease that is already exist. Those factors were as follows:

### A. Chest Pain Type (Cp)

Chest pain is generally caused by one of the parts of the chest (heart, lung, or esophagus) or by the chest wall (skin, muscle, or bone) [15].

- Typical angina: It is serious, and may be a sign that a heart attack could happen soon.
- Atypical angina: Often does not cause pain; it feels a vague discomfort in the chest; patients experience shortness of breath, feel tired or nauseous, have indigestion, or pain in the back or neck.
- Non-anginal pain: It occurs when tiny vessels in the heart become narrow and stop functioning properly.
- Asymptomatic: It often occurs while the patient is resting, and it cannot be predicted. It may cause severe pain, and is usually the result of a spasm in a coronary artery.

### B. Resting Electrocardiographic Results

An electrocardiogram (ECG) is a measure that tests the electrical activity of human heart to check if it works normally. An ECG records the heart rhythm. This attribute is divided into three values [16]:

- Value 0: Normal

$$0.12<PR<0.20 \text{ second}$$

$$0.08<QRS<0.10 \text{ second}$$

$$0.5<ST\text{-}T<0.55 \text{ mV}$$

- Value 1: Where ST-T wave abnormal (T inversion or ST dispersion)
- Value 2: Where ST-T wave having hypertrophy.

Vale 1, 2 indicates that the heat works normally

### C. Exercise Induced Angina (Exang)

Angina means that the heart is not getting enough blood and oxygen, and it may result in chest pain, the two most common types of angina being stable angina and unstable angina [15]:

- Stable angina: Occurs during exercise, when the heart has to work harder than normal.
- Unstable angina: It is more serious, and may be a sign that a heart attack could happen soon. It should always be treated as an emergency. People with unstable angina are at increased risk for a heart attack.

ST depression induced by exercise relative to rest. ST segment depression is the classical response to stress during exercise stress testing. ST segment depression estimated by measuring the distance between the isoelectric line of the QRS complex and the beginning of T-wave. A positive value represents an ST elevation, and a negative value represents an ST depression.

### D. The Slope of the Peak Exercise ST Segment

ST depression can be either up sloping, down sloping, or horizontal. This attribute is subdivided into three values:

- Up sloping: Progressive ischemia
- Flat: Ischemia
- Down sloping: Myocardial ischemia

The three values indicate that heart is not working normally.

The selection of these attributes to build a system for the prediction of heart disease has led to a defect in the results. To correct this error, we replaced all these attributes with new heart risk factors attributes. In this research, we build **AHDBP** using three basic classification techniques: neutral network, decision tree, and naïve Bayes.

## III. Methodology

### A. Heart Disease Data

In 2015, World Health Organization published an article on the risk factors leading to a heart attack. These factors are numerous, and are divided into several groups based on the degree of their impact on the heart. To use all these factors, they must be available in the database, and often this does not happen, so we used the attributes available in the database of The Cleveland Heart Disease. Some factors were not available in the database but they related to other factors by means mathematical or medical relationships, which were used to obtain values of these factors, the data set consists of 387 instance and 18 attributes. Attribute values are a mixture of nominal and numeric. These attributes are illustrated in **Table 1.** Waikato Environment for Knowledge Analysis (**WEKA**) data mining software used due to its proficiency in discovering, analysis and predicting patterns. The values of some attributes are derived from other attributes related to them through mathematical or medical relationships, which are as follows:

- Body Mass Index (BMI): Weight that is higher than what considered as a healthy weight for a given height is described as overweight or obese. Body Mass Index, or BMI, used as a screening tool for excessive weight or obesity, it calculated from relation:

$$BMI = \frac{\text{weight in kilograms}}{\text{square of height in meters}} \quad (1)$$

A high BMI can be an indicator of high body fatness [16]. At an individual level, BMI can be used as a screening tool, but it is not a diagnostic of the body fatness or the health of an individual.

TABLE I. LIST OF ATTRIBUTES USED IN SYSTEM DESIGN

| Attribute | features | Probability | Description |
|---|---|---|---|
| Age | 25-29 | 0.6375 | Young |
| | 30-39 | 1.53 | Youth adult |
| | 40-59 | 2.29 | Adult |
| | >60 | 5.73 | Old age |
| Sex | 1 | 1.53 | Male |
| | 0 | 2.29 | Female |
| BP(Blood Pressure) | 120-129 | 0 | Normal |
| | 130-139 | 5.73 | Stage 1 hypertension |
| | 140-179 | 9.72 | Stage 2 hypertension |
| | >180 | 14.58 | Hypertension crisis |
| Fbs (history of diabetes) | 1 | 3.18 | History |
| | 0 | 0 | No history |
| Thalach (Maximum Heart Rate achieved) | > 75% | 3.18 | Up normal |
| | 50% - 75% | 0 | Normal |
| | < 50% | 2.29 | medium |
| HRR ( Heart Resting Rate) | 60-80 | 0 | Normal |
| | 80-99 | 5.73 | Medium risk |
| | >100 | 9.72 | Up normal |
| Smoke (Tobacco use) | 1 | depend | Smoking |
| | 0 | depend | No Smoking |
| Cig (Number of Cigarettes per Day) | 0 | 0 | No Smoking |
| | 1-4 | 1.27 | Light smoker |
| | 5-19 | 5.73 | Medium smoker |
| | >20 | 14.58 | Huge smoker |
| Year (Smoking years) | 0 | 0 | No smoking |
| | 1-19 | 1.27 | Medium |
| | 20-29 | 5.73 | High stage 1 |
| | 30-39 | 7.65 | High stage 2 |
| | 45-49 | 9.72 | Very high |
| | >50 | 14.58 | Critical |
| Famhist (Family history of coronary) | 1 | 3.18 | History |
| | 0 | 0 | No history |
| TC (Total Cholesterol) | 0-199 | 0 | Desirable |
| | 200-239 | 3.82 | Borderline high |
| | >240 | 9.72 | High |
| TG (Triglycerides (fasting)) | 0-150 | 0 | Desirable |
| | 150-199 | 3.82 | High |
| | 200-249 | 7.65 | Very high |
| | >250 | 9.18 | Critical high |
| HDL (High Density lipids) | 0-40 | 9.18 | Poor |
| | 40-59 | 0.76 | Better |
| | >60 | 0 | Best |
| LDL (Low Density Lipids) | 100-129 | 0 | Optimal |
| | 130-159 | 5.73 | Borden high |
| | 160-189 | 9.18 | High |
| | >190 | 14.58 | Very high |
| BMI (Body Mass Index(obesity)) | <18.5 | 1.14 | Under weight |
| | 19-24.9 | 0 | normal weight |
| | 25 - 29.9 | 5.73 | Over weight |
| | >30 | 7.63 | Obese |
| Ca (Cardiovascular Disease History) | 1 | 3.18 | History |
| | 0 | 0 | No history |
| Va (Vascular Disease History) | 1 | 3.18 | History |
| | 0 | 0 | No history |
| LVH (Left ventricular hypertrophy) | 1 | 3.18 | History |
| | 0 | 0 | No history |
| NUM (Predictable attribute) | 1 | patient | Heart patient |
| | 0 | Not patient | Not patient |

- Total cholesterol (TC): Cholesterol is a fatty and waxy substance called a lipid, which is essential for maintaining the outer membranes of cells, but it becomes unhealthy in excessive amounts. A high level of "bad" cholesterol indicates that the heart arteries are lined with fatty deposits, possibly leading to heart attack or stroke.
- Low-density lipoprotein (LDL): A combined reading of LDLs and VLDLs (very low-density lipoproteins). LDLs form a plaque buildup in the arteries, narrowing them. They are referred to as "bad" cholesterol [17].
- High-density lipoprotein (HDL): Transport cholesterol in the bloodstream back to the liver and reduce the amount of cholesterol; called "good" cholesterol.
- Triglycerides (TG): They contribute to the narrowing and hardening of the arteries.

Total cholesterol (TC) is a combination of HDL, LDL and TG according the following equation:

$$TC=LDL+HDL+\frac{TG}{5} \quad (2)$$

- Vascular disease history (Va): A class of diseases of the blood vessels (the arteries and veins of the circulatory system of the body). It is a subgroup of cardiovascular disease. Disorders in this vast network of blood vessels can cause a range of health problems, which can be severe or prove fatal. There are several types of vascular disease, (which is a subgroup of cardiovascular disease), and the signs and symptoms depend on which type [18]. Since it is related with cardiovascular disease, we used Ca (Cardiovascular Disease History) that is available in the database to obtain (Va) values.
- Left ventricular hypertrophy history (LVH): Left ventricular hypertrophy is found in hypertensive patients, and it increases the risk of stroke and death. Recent research indicated it is a modifiable risk factor of heart disease. LVH is diagnosed on electrocardiogram (ECG) when the myocardium is hypertrophied: there is a larger amount of myocardium for electrical activation to pass through; thus the amplitude of the QRS complex, representing ventricular depolarization, is increased [19]. Since LVH is related with ECG, we used ECG attribute that is available in the database to obtain values. For any patient having up normality in ECG, it considered as LVH history.

## IV. Heart Disease Prediction Model

In this research, framework of AHDBP is built to run with different algorithms of the Decision Tree, Naive Bayes, and Neural Networks classification techniques, by using data from 370 patients. **Fig. 1** shows the proposed system which consists of different layers: (1) Data preprocessing phase to prepare the data before analysis, (2) Data mining phase (3) Pruning phase (4) Evaluation phase.

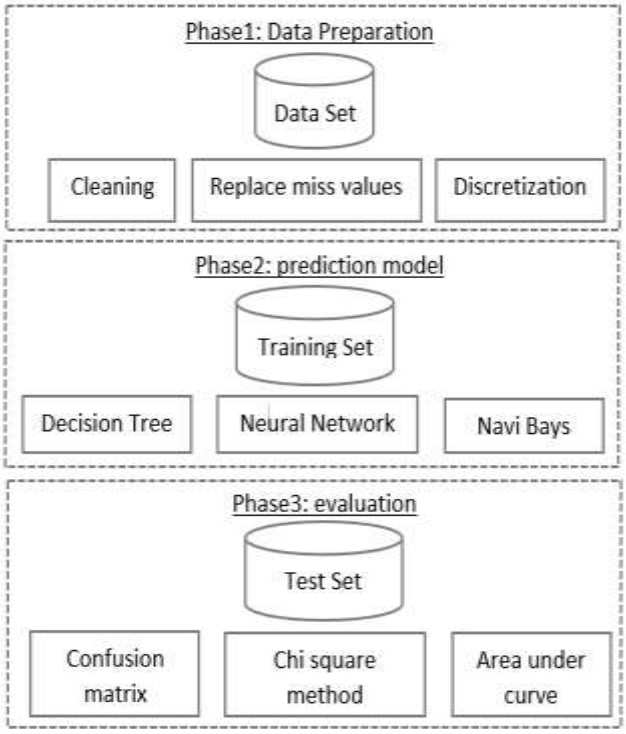



Fig. 1. The Framework of Proposed Heart Disease Prediction System.

### A. Data Preprocessing Phase.

In this phase, a group of operations is applied to clean empty records, to replace missed data, and to perform data discretization.

- Data cleaning; replacing missing process: The purpose of this process is to remove the instances and attributes that contain empty values only. While attributes and instances feature partially missing data, "replace missing" filter is applied to the data to replace empty values with certain ones through applying statistical methods. The data before applying this step contained 24 attributes and 394 instances, and after applying this process only 18 attributes and 370 instances are produced. Then the heart disease data divided into two groups: the first group used for system design involves 200 instances, while the second group consists of 170 instances and is used for training.
- Data discretization process: The number of attributes used is 18, and each attribute has a set of features according to its type and its effect on the heart, for example the smoking attribute includes two features (smoke – non-smoke). The same scenario was applied to the rest of the attributes until the number of features became 51 as illustrated in Table 1. Based on the value and degree of effect of these features, they are divided into a number of groups; each group is assigned to a number with no overlaps in between according to a statistical calculation, as shown in **Fig. 2.**

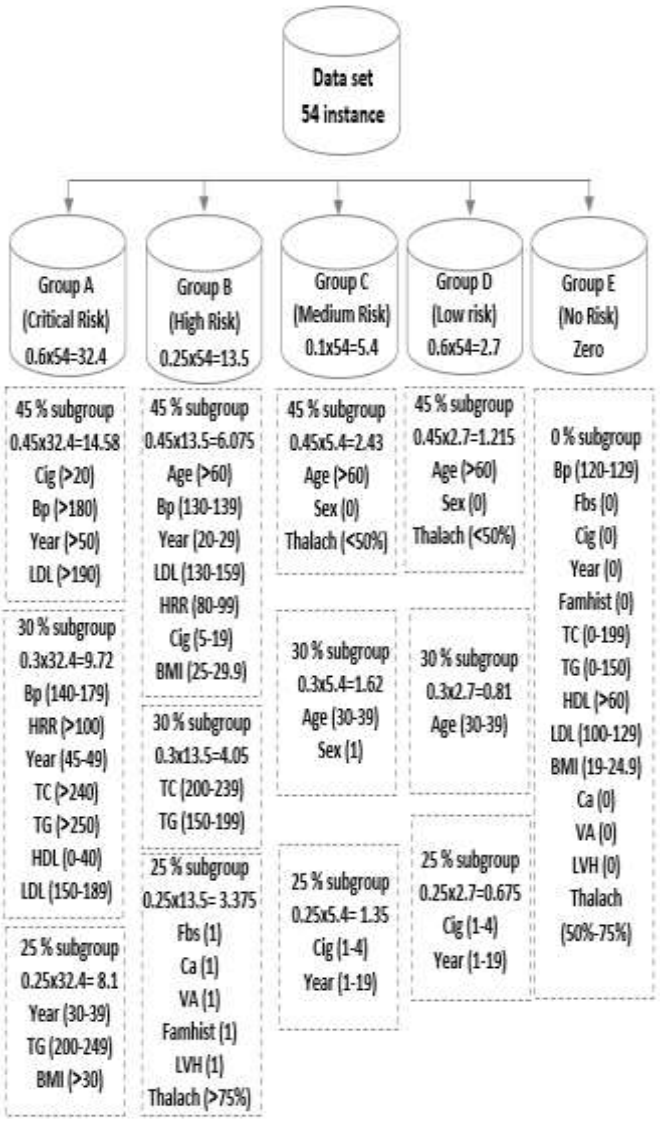



Fig. 2. Data Discretization Process.

### *B. Binning Process*

Data binning or bucketing is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values that fall in a given small Interval, a bin replaced by a value representative of that interval, often the central value. Authors in [20] used the number of bin of Equal Width discretization to<k = max(1,2log(l))> , where $l$ is the number of distinct values of the attribute. However, in most cases, the number of bins is always set to 10 for the Equal Frequency and Equal Width methods. In the literature, the problem of choosing the optimal number of bins has not been considered in supervised learning.

### *C. Prediction Model Phase*

Three basic data mining techniques are applied to build our prediction system based on the best accuracy produced from each model.

- Decision Tree Model: Weka J48 implements the decision-tree learning algorithm applied to the training data set. Some basic constructive steps on which this algorithm is based is presented in the following sections.

A decision tree grows recursively by partitioning each set into successively subsets. Let $D_t$ be the set of records which are associated with the node t, and let $y = \{y_1, y_2, y_3, \ldots . y_c\}$ be the class label.

Step 1: If all records in $D_t$ belong to the same class $y_t$, then t is a leaf node labeled as $y_t$.

Step 2: If the records contained in $D_t$ belong to more than one class, then the records are portioned into smaller subsets, then a child node is created for each output of the test condition and the records in the $D_t$ will be distributed to the children based on the outcomes. The algorithm is then recursively applied to each child node.

Step 3: If the child node created in step 2 is empty, the node is declared as leaf node with the same class label. If the records of $D_t$ have identical attributes, then the split process cannot be applied any more, and in this case the node is declared as leaf node with the same class label.

In order to divide the attributes into smaller subsets, an attribute test condition must be selected for each tree growing process. To perform this step, the employed algorithm must provide a method for specifying the test condition for different types of attributes. The tree growing process must be stopped. This happens when all records that belong to the same class or all records have identical attribute value.

Decision Trees involve many different types, and the selection of each is based on the mathematical model used for the selection of attributes splitting. The best types used in previous studies are as follows: Information Gain, and Gain Ratio. Counting the best split is defined in term of the class classification of the record before and after splitting. At a given node t, let P(i | t) contain a fraction of records that belongs to the same class i. The best split is often based on the degree of impurity of the child nodes: the smaller the degree of impurity, the more class distribution.

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} P(i|t) \log_2 P(i|t) \quad (3)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [P(i|t)]^2 \quad (4)$$

$$\text{Classification error}(t) = 1 - \max_i f_0[P(i|t)] \quad (5)$$

Where c is the number of the class.

The minimum values for the measure attained when all the records belong to the same class.

This method is used to reduce the effect of the resulting bias from the use of Information Gain [18]. The Gain Ratio adjusts the Information Gain for each attribute to allow for the breadth and uniformity of the attribute values.

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}} \quad (6)$$

- Naive Bayes: The Weka Naive Bays learning algorithm was applied to the data sets. This classifier is based on the Bayes theorem that assumes that all attributes in the same class are conditionally independent from each other. The following steps are applied:

Step 1: Collect exemplars for each class.

The Bayes theorem express a class of independent attributes as follows:

$$X=\{x_1, x_2... x_n\} \quad (7)$$

Where X is the evidence.

Step 2: Estimate class a priori probabilities.

Each specific class combines a group of evidence X. The class prior may be calculated as follows:

$$\text{prior}= \frac{\text{number of samples in the class}}{\text{total number of samples}} \quad (8)$$

And probability P (H|X) of given evidence is calculated as follows:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (9)$$

Where I is the hypothesis means.

Step 3: Estimate class means.

Data are segmented by the class, and then we compute the mean and variance of in each class. Let $\mu_c$ be the mean of the values in associated with the class, and let ${\sigma_c}^2$ be the variance of the values in associated with the class. Then the probability of some value given a class can be computed as:

$$P(X)= \frac{1}{\sqrt{2\pi{\sigma_c}^2}} e^{\frac{(x-\mu_c)^2}{2{\sigma_c}^2}} \quad (10)$$

Step 4: Construct a classifier from the probability model

The Naive Bayes classifier uses this model with a decision rule. Using one common rule to pick the most probable hypothesis, this is known as the maximum a posteriori decision rule.

- Neural Network Model

Neural networks use many types based on different types of rules. In our current research we use the feed forward network type, where input information comes in one direction starting from the input layer and passing through the hidden layers, and finally ending up with the output layer. A feed forward neural network is an artificial neural network where connections between the units do not form a cycle. Several hidden layers can be placed between the input and output layers.

Input Layer: The activity of the input units represents the raw information that is fed into the network.

Hidden Layer: The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units.

Output Layer: The behavior of the output units depends on the activity of the hidden units.

## V. Pruning Evaluation Phase

### A. Pruning

Pruning is process applied in order to reduce errors resulting from classification errors, which occurred because of specialization in the training set. The application of reduced error pruning provides more compact decision rules and reduces the number of extracted rules [19]. After the whole production processes of the decision tree, which classify all the training set instances, the pruning process was applied to make the tree more generic.

### B. Evaluation

The accuracy of the system was tested and validated using a test set by the holdout cross validation and chi square methods.

In the holdout cross validation, the data set was divided into the training set and the testing set [20]. The training set was only used to fit a function approximation to predict the output values for the testing set (never seen these output values before). The errors of the approximate function were gathered to give the mean absolute test set error that was used to evaluate the model. This method is usually preferable to the other methods and takes less time to compute.

Chi square method used to determine whether there is a significant difference between the expected frequencies and the observed frequencies [21]. The observations are classified into mutually exclusive classes (null hypothesis) where it gives the probability that any observation falls into the corresponding class. The main objective of the test is to evaluate how likely the observations t would be assuming the null hypothesis is true. Assume that there are n observations in a random sample from a population classified into k mutually exclusive classes with respective observed numbers $x_i$ (for $i = 1,2,\dots,k$), and a null hypothesis gives the probability $p_i$ that an observation falls into the $i^{th}$ class. So we have the expected numbers $m_i = np_i$ for all $i$, where:

$$\times^2= \sum_{i=1}^{k} \frac{(x_i-m_i)^2}{m_i} \quad (11)$$

An algorithm has been built to perform the chi square test, and the following steps were applied:

Step 1: Null hypothesis is given

$H_0$: $patient$
$H_E$: $not\ patient$

Step 2: Degree of freedom calculated

$$\text{number}=(\text{coulms-1})\times(\text{rows-1})$$

Step 3: Observed number calculated form equation 11

Step 4: Critical value assigned according to the degree of freedom.

Step 5: Observed number compared to the critical vale to determine the hypothesis class. If the test statistic is improbably large according to that chi-squared distribution, then one rejects the null hypothesis of independence.

After applying the chi square test to the system using the test set, about 86.54% of the data gave valid and accurate results, while 13.45% needs more validation. This proves the reliability and efficiency of the proposed system.

## VI. Experimental Results

Most of the previous studies have used some heart disease symptoms as major risk factors to build a heart disease prediction, Also the developed frameworks can be used only for dedicated classification algorithms, and is not adaptable with other types of classification algorithms. In order to enhance the accuracy of the system, we employed AHDBP framework that run with different algorithms of the Decision Tree, Naive Bayes, and Neural Networks classification techniques.

Data for more than 370 heart disease patients with 18 risk factors attributes were used for analysis, also wrong attributes used in previous works are declared, identified and removed from the data.

The results were compared with the previous studies discussed previously, and they show the effect of new modifications on the accuracy and efficiency of the prediction system. Data for more than 370 heart disease patients with 18 attributes have undergone several stages of data analysis. Each stage has several results depending on the type of operation performed during the analysis stages.

At the beginning of the analysis, patients data were reviewed, and attributes that completely contain missing part were removed, while for attributes and instances that contains partial missing data, a "replace missing" filter was used to predict missing parts. This steps reduced the number of instances from 394 to 370, while the number of attributes fell from 24 to 18. Then the data were discretized using statistical methods.

After applying the model, a large scale of statistical information was obtained. These performance measures were used to evaluate the model as shown in **Table 2**. Then the resulting data where applied to different data mining algorithms. The accuracy results of the heart disease prediction module are shown in **Table 3**. Results show that the highest accuracy achieved using neural network was 94.91%, while the for the Naïve Bayes algorithm the highest accuracy was 91.54%, and for the pruned decision tree algorithm 90.34%. Neural networks can predict heart disease with more accuracy than naïve Bayes and the decision tree. **Table 4** compares the results obtained from current research with previous studies. **Fig. 3** shows a comparison between the Receiver Operating Characteristic (ROC) curves for the classification methods models and their sensitivity and specificity values at the optimal cutoff points.

TABLE II. Performance of the AHDBP three Classification Techniques

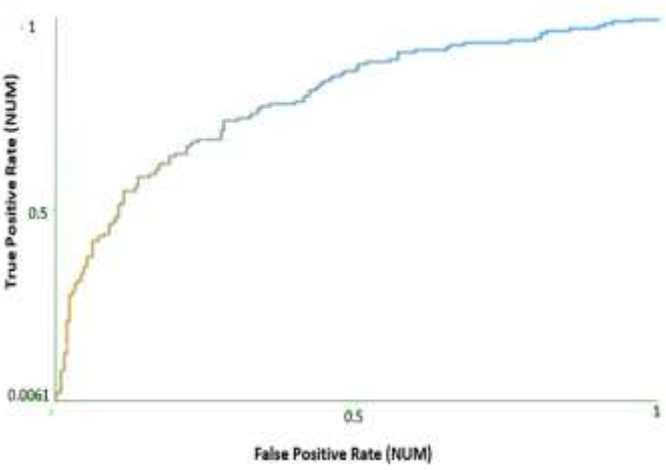
| Classification Technique | Class | TP | TF | SENS | ROC |
|---|---|---|---|---|---|
| Decision tree | Class 0 | 0.95 | 0.16 | 89.4% | 0.885 |
| | Class 1 | 0.84 | 0.05 | 83% | 0.932 |
| Naive Bayes | Class 0 | 0.95 | 0.16 | 78% | 0.885 |
| | Class 1 | 0.84 | 0.05 | 67.8% | 0.932 |
| Neural networks | Class 0 | 0.93 | 0.03 | 100% | 0.975 |
| | Class 1 | 0.96 | 0.07 | 98.5% | 0.919 |

TABLE III. Best Accuracy Results

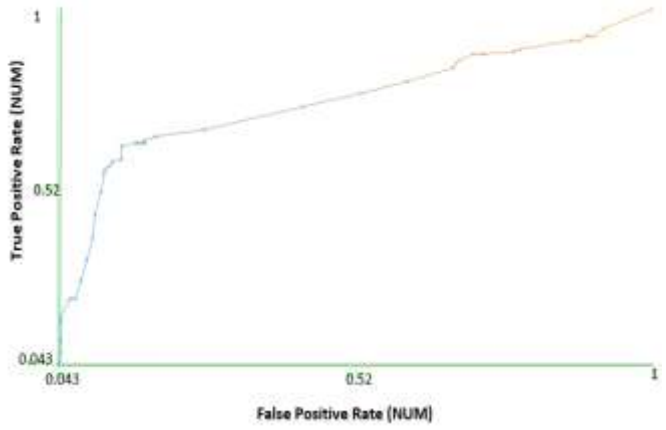
| Technique | Percentage |
|---|---|
| Neural Networks | 94.91% |
| Naive Bayes | 91.54% |
| Decision Tree | 90.34% |

TABLE IV. Comparison between Obtained Results and Previous Studies

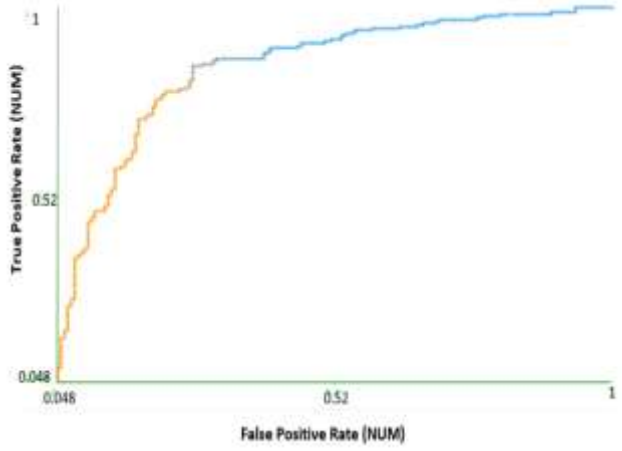
| Author | Attribute | Technique | % | AHDBP % | Accuracy |
|---|---|---|---|---|---|
| R. Assari | 13 | Decision tree | 79 | 90.34 | 11.34 |
| | | Naive Bayes | 83.7 | 91.54 | 7.84 |
| Nidhi Bhatla | 15 | Decision tree | 89 | 90.34 | 1.34 |
| | | Naive Bayes | 86 | 91.54 | 5.54 |
| | | Neural Networks | 85.3 | 94.91 | 9.61 |
| Aravinthan | 14 | Decision tree | 80.09 | 90.34 | 10.25 |
| | | Naive Bayes | 81.30 | 91.54 | 10.02 |
| | | Neural Networks | 82.56 | 94.91 | 12.35 |
| Boshra Bahrami | 8 | Decision tree | 83.7 | 90.34 | 6.67 |
| | | Naive Bayes | 81.8 | 91.54 | 9.74 |
| Shined S. B.1 | 13 | Naïve Bayes | 86.66 | 91.54 | 4.88 |




(A)




(B)




(C)

Fig. 3. ROC Curves, (A) Decision Tree, (B) Navi Bays, (C) Neural Networks.

## VII. Conclusion

In this paper, a framework of an AHDBP was developed. The framework can be implemented by applying different algorithms of the decision tree, naive Bays and neural networks mining techniques. Data from 370 patients from different heart disease database resources has analyzed. The proposed system is built based on risk factors of patients instead of their symptoms, unlike in most of the previous researches. Working based on symptoms delivers totally inaccurate results.

The data went through several stages of analysis, the first of which is data preparation. Afterwards, the data were discretized using supervised and unsupervised discretization. Second, the preprocessed data applied to different data mining techniques (decision tree, nave Bayes and neural network). The experimental results showed that the neural networks can predict heart disease with more accuracy than naïve Bayes and the decision tree.

In the future, more data sets will be used to train other classifiers and to try more experiments. Other techniques will be applied, too, and more than one technique will be combined to reach as high accuracy as possible.

## References

[1] J. Han, M. Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.

[2] P. Sellappan, R. Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, 978-1-4244-1968-5/08/$25.00©2008 IEEE.

[3] D.M. Hlaudi, A. M. Mosima, “Prediction of Heart Disease using Classification Algorithms”, Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[4] A. Kumar, P.P. Pandey, K.L. Jaiswal, “A Heart Disease Prediction Model using Decision Tree” 2013 IUP. All Rights Reserved.

[5] W.J. Frawley, G. P. Shapiro, J. M. Christopher, “Knowledge Discovery in Databases: An Overview”, AI Magazine Volume 13 Number 3 (1992) (© AAAI).

[6] N. Guru, A. Dahiya, N. Rajpal, “Decision Support System for Heart Disease Diagnosis Using Neural Network”, Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

[7] R. Assari, P. Azimi, M.R. Taghva, “Heart Disease Diagnosis Using Data Mining Techniques” Int J Econ Manag Sci 6 (2017): 415. Doi: 10.4172/2162-6359.1000415.

[8] C.S. Dangare, S.S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888), Volume 47– No.10, June 2012.

[9] S.Vijayarani, S.Sudha, “A Study of Heart Disease Prediction in Data Mining”, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.5, October 2012.

[10] A.H Chen, S.Y Huang, P.S Hong, C.H Cheng, E.J Lin, “HDPS: Heart Disease Prediction System” Computing in Cardiology 2011; 38:557-560, ISSN 0276-6574.

[11] N. Bhatla, K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 8, October – 2012, ISSN: 2278-0181.

[12] K. Aravinthan, M.Vanitha, "A Novel Method for Prediction Of Heart Disease Using Naïve Bayes", International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 3, Special Issue 20, April 2016, ISSN 2394-3785.

[13] H.D. Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms." WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[14] B. Bahrami, M.H. Shirvani "Prediction and Diagnosis of Heart Disease by Data Mining Techniques 'Journal of Multidisciplinary Engineering Science and Technology, (2015), Vol 2, No (3159-0040).

[15] R.O. Bonow, N. Bohannon, W. Hazzard. Risk stratification in [127] Colditz GA, Stampfer MJ,Willett WC. A prospective study of coronary artery disease and special populations. Am J Med parental history of myocardial infarction and coronary heart disease 1996;101:4A17S–22S.

[16] J. Millan, Lipoprotein ratios: "Physiological significance and clinical usefulness in cardiovascular prevention", (2009). DOI: 10.2147/VHRM.S6269.

[17] Q. Sun, "Comparison of dual-energy x-ray absorptiometric and anthropometric measures of adiposity in relation to adiposity-related biologic factors. Am. J. Epidemiol.", 2010, 172(12), pp.1442–1454.

[18] A. Bikfalvi, "Encyclopedic Reference of Vascular Biology & Pathology". Springer. ISBN 9783642570636 (2013-12-19).

[19] S. Michael, S. Lauer, "Heart Rate Response in Stress Testing: Clinical Implications, the American College of Cardiology, Published by Elsevier Science Inc"., 1062-1458/01/$20.00. PII S1062-1458(01)00423-8.

[20] J.R. Dougherty, Kohavi, et al. (1995). "Supervised and unsupervised discretization of continuous features." In: Proceedings of the 12th international conference on machine learning. San Francisco: Morgan Kaufmann: p. 194–202.

[21] Cochran, William G. "The Chi-square Test of Goodness of Fit". The Annals of Mathematical Statistics. 23: 315–345. Figures and Tables.