Vol III Issue VIII May 2014

ISSN No : 2249-894X

Monthly Multidisciplinary Research Journal

# Review Of Research Journal

**Welcome to Review Of Research**

**RNI MAHMUL/2011/38595** **ISSN No.2249-894X**

Review Of Research Journal is a multidisciplinary research journal, published monthly in English, Hindi & Marathi Language. All research papers submitted to the journal will be double - blind peer reviewed referred by members of the editorial Board readers will include investigator in universities, research institutes government and industry with research interest in the general subjects.

**Address:-Ashok Yakkaldevi 258/34, Raviwar Peth, Solapur - 413 005 Maharashtra, India**
**Cell : 9595 359 435, Ph No: 02172372010 Email: ayisrj@yahoo.in Website: www.ror.isrj.net**

Review Of Research
Vol. 3 | Issue. 8 | May. 2014
Impact Factor : 2.1002 (UIF)
ISSN:-2249-894X

Available online at www.ror.isrj.net

ORIGINAL ARTICLE

# PERFORMANCE ANALYSIS OF DATA MINING ALGORITHMS FOR DIAGNOSIS AND PREDICTION OF HEART AND BREAST CANCER DISEASE

**Vikas Chaurasia and Saurabh Pal**

Lecturer, KHBS College of Pharmacy, Jaunpur, UP, India.
Head, Dept. of MCA,VBS Purvanchal University, Jaunpur, UP, India.

**Abstract:**

*Heart disease or cardiovascular diseases are the number one cause of death and they are projected to remain so. An estimated 17 million people died from cardiovascular disease in 2005, representing 30% of all global deaths. Of these deaths, 7.2 million were due to heart attacks and 5.7 million due to stroke. About 80% of these deaths occurred in low- and middle income countries. If current trends are allowed to continue, by 2030 an estimated 23.6 million people will die from cardiovascular disease (mainly from heart attacks and strokes).*

*Breast cancer is the second most common cancer in women. The World Health Organization's International estimated that more than 1,50,000 women worldwide die of breast cancer in year. In India, breast cancer accounts for 23% of all the female cancer death followed by cervical cancer which accounts to 17.5% in India.*

*The main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for heart disease patients and breast cancer survivability. We used five popular data mining algorithms (Naïve Bayes, RBF Network, Simple Logistic, J48 and Decision Tree) to develop the prediction models using a large dataset (270 Heart disease and 683 breast cancer cases). We also used 10-fold cross-validation methods to measure the unbiased estimate of the five prediction models for performance comparison purposes. The results (based on average accuracy of Heart and Breast Cancer data set) indicated that the Naïve Bayes is the best predictor with 87.01% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), RBF Network came out to be the second with 86.9% accuracy, Simple Logistic came out to be third with 85.65% accuracy, J48 came out fourth with 84.85% accuracy and the Decision table models came out to be the worst of the five with 83.34% accuracy.*

**KEYWORDS:**

Cardiovascular disease, Breast cancer, Naïve Bayes, RBF Network, Simple Logistic, J48, Decision Tree.

## 1.INTRODUCTION

The number and the size of databases recording medical data are increasing rapidly. Medical data, produced from measurements, examinations, prescriptions, etc., are stored in different databases on a continuous basis. This enormous amount of data exceeds the ability of traditional methods to analyze and search for interesting patterns and information that is hidden in them. Therefore new techniques and tools for discovering useful information in these data depositories are becoming more demanding [1]. Analyzing these data with new analytical methods in order to find interesting patterns and hidden knowledge is the first step in extending the traditional function of these data sources. The main purpose of this research work involves methodology that starts with understanding the domain, locating proper data sources, preparing the raw data, applying advanced analysis techniques, and extracting and validating the resulting knowledge for heart disease and breast cancer survivals.

### 1.1 Heart disease

Heart disease is a type of cardiovascular disease. In addition to heart disease, the term cardiovascular disease encompasses a variety of heart conditions, such as high blood pressure and stroke. Coronary heart disease (CHD) is caused by a narrowing of the coronary arteries, which results in a decreased supply of blood and oxygen to the heart. CHD includes myocardial infarction, commonly referred to as a heart attack, and angina pectoris, or chest pain. A heart attack is caused by the sudden blockage of a coronary artery, usually by a blood clot. And chest pain occurs when the heart muscle does not receive enough blood. Another type of heart disease is a heart rhythm disorder, which includes rapid heart, heart murmurs, and other unspecified disorders. Congestive heart failure (CHF), which is often the end-stage of heart disease, is another disease of the heart.

The major causes of cardiovascular disease are tobacco use, physical inactivity, and an unhealthy diet. Over 80% of cardiovascular disease deaths take place in low-and middle-income countries and occur almost equally in men and women.

### RISK FACTORS

Tobacco use, an unhealthy diet, and physical inactivity increase the risk of heart attacks and strokes.
High blood pressure has no symptoms, but can cause a sudden stroke or heart attack.
Diabetes increases the risk of heart attacks and stroke.
Being overweight increases the risk of heart attacks and strokes.
Low socioeconomic status increases the chances of exposure to risk factors and increases the vulnerability to develop CVD.

A healthy lifestyle can reduce the risk of heart disease by as much as 80 percent [2]. People who are not overweight, do not smoke, consume about one alcoholic drink a day, exercise vigorously for 30 minutes a day or more, and eat a low-fat, high-fiber diet have the lowest risk for heart disease. Heart disease is largely preventable by virtue of a healthy lifestyle.

### 1.2 Breast cancer

The organs and tissues of the body are made up of tiny building blocks called cells. Cancer is a disease of these cells. Although cells in each part of the body may look and work differently, most repair and reproduce themselves in the same way. Normally, cells divide in an orderly and controlled way. But if for some reason the process gets out of control, the cells carry on dividing and develop into a lump called a tumour. Breast tumours are usually caused by an overgrowth of the cells lining the breast ducts. They can be either benign or malignant. In a benign tumour, the cells grow abnormally and form a lump. But they don't spread to other parts of the body and so are not cancer. The most common type of benign breast tumour is called a fibroadenoma. This may need to be surgically removed to confirm the diagnosis. No other treatment is necessary. In a malignant tumour, the cancer cells have the ability to spread beyond the breast if they are left untreated. For example, if a malignant tumour in the breast isn't treated, it may grow into the muscles that lie under the breast. It can also grow into the skin covering the breast. Sometimes cells break away from the original (primary) cancer and spread to other organs in the body. They can spread through the bloodstream or lymphatic system. When these cells reach a new area they may go on dividing and form a new tumour. The new tumour is often called a secondary or metastasis. Breast cancer occurs when cells within the breast ducts and lobules become cancerous. If caught at an early stage, breast cancer can often be cured. If the cancer has spread to other areas of the body it can't usually be cured, but it can normally be effectively controlled for a long time.

## RISK FACTORS

being a woman
getting older
having an inherited mutation in the BRCA1 or BRCA2 breast cancer gene
lobular carcinoma in situ (LCIS)
a personal history of breast or ovarian cancer
a family history of breast, ovarian or prostate cancer
having high breast density on a mammogram
having a previous biopsy showing atypical hyperplasia
starting menopause after age 55
never having children
having your first child after age 35
radiation exposure, frequent X-rays in youth
high bone density
being overweight after menopause or gaining weight as an adult
postmenopausal hormone use (current or recent use) of estrogen or estrogen plus progestin

**The best way to find breast cancer early is to get screened.**

A mammogram is an X-ray of the breast. It is the best screening tool used today to find breast cancer early. A mammogram can find cancer in its earliest stages, even before a lump can be felt. All women age 40 and older should have a mammogram every year. If you are younger than age 40 with either a family history of breast cancer or other concerns, talk with health care provider about when to start getting mammograms or other screening tests, like MRI, and how often to have them.

A clinical breast exam is done by a health care provider who checks breasts and underarm areas for any lumps or changes. Many women have a clinical breast exam when they get their Pap test. Women should have a clinical breast exam at least every 3 years between the ages of 20 and 39 and every year starting at age 40.

### 1.3 Knowledge discovery in databases and data mining

Data mining can be a solution by generating rules from those enormous datasets which can be used in echo readings. Data mining is a crucial step in discovery of knowledge from large datasets. In recent years, Data mining has found its significant hold in every field including health care. Mining process is more than the data analysis which includes classification, clustering, and association rule discovery. It also spans other disciplines like Data Warehousing, Statistics, Machine learning and Artificial Intelligence (Larose 2005). With the rising volumes of electronic patient records, data mining has become general to excerpt hidden patterns inpatient data for healthier understanding of relationships within the data. Data mining in medical domain is single from that in other domains due to the special characteristics of medical datasets. Medical datasets are often privacy-sensitive, huge and heterogeneous with data collected from dissimilar sources. The collected data may also need to be branded mathematically. In clinical domain, data mining methods have been applied to big clinical repositories containing clinical and administrative data collected from electronic sources to identify new disease associations. The techniques applied include pattern detection to identify commonly happening associations in the dataset, predictive analysis to predict upcoming outcome for a patient based on the existing patient records, and association mining to excerpt interesting rules from the recognized associations. Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into broad categories [3, 4].

The remainder of this paper is organized as follows. Section 2 provides the reader with the background information on heart disease patients and breast cancer research, previously published relevant literature. In Section 3, we explained in detail methodology that is used in this paper. In Section 4 and 5, explained the heart and breast cancer dataset in detail. In Section 6, the evaluation methods are described. The paper concludes with Section 7 and 8 where we summarize the research findings, and outline the limitations and further research directions.

## 2BACKGROUND

Several studies have been reported that have focused on cardiovascular disease diagnosis and breast cancer survivals. These studies have applied different approaches to the given problem and achieved high classification accuracies. Here are some examples:

Robert Detrano's experimental results showed correct classification accuracy of approximately 77% with logistic regression derived discriminant function [5].

B. Zheng Yao applied a new model called R-C4.5 which is based on C4.5 and improved the efficiency of attribution selection and partitioning models. An experiment showed that the rules
created by R-C4.5s can give health care experts clear and useful explanations [6].
C. Resul Das introduced a methodology that uses SAS base software 9.13 for diagnosing heart disease. A neural networks ensemble method is at the center of this system [7].
D. Colombet et al. evaluated implementation and performance of CART and artificial neural networks comparatively with a LR model, in order to predict the risk of cardiovascular disease in a real database [8].
E. Engin Avci and Ibrahim Turkoglu study an intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases [9].
F. Imran Kurt , Mevlut Ture , A. Turhan Kurum compare performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease [10].
G. The John Gennari's CLASSIT conceptual clustering system achieved a 78.9% accuracy on the Cleveland database [11].
V. Chauraisa and S. Pal given a diagnosis system based on different data mining techniques on early prediction of heart diseases [12-13].
Jinyan LiHuiqing Liu's [14] experimented on ovarian tumor data to diagnose cancer using C4.5 with and without bagging.
Dong-Sheng Cao's [15] proposed a new decision tree based ensemble method combined with feature selection method backward elimination strategy with bagging to find the structure activity relationships in the area of chemometrics related to pharmaceutical industry.
Liu Ya-Qin's [16] experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability.
Tan AC's [17] used C4.5 decision tree, bagged decision tree on seven publicly available cancerous micro array data, and compared the prediction performance of these methods.
V. Chauraisa and S. Pal [18-19] experimented on breast cancer data to diagnose cancer using RepTree, RBF Network and Simple Logistic and SMO, IBK and BF Tree methods.

## 3.METHODOLOGY

This paper uses five data mining algorithms each on heart dataset and breast cancer dataset, Naïve Bayes, RBF Network, Simple Logistic, J48 and Decision Tree. These classification algorithms are selected because they are very often used for research purposes and have potential to yield good results. Moreover, they use different approaches for generating the classification models, which increases the chances for finding a prediction model with high classification accuracy.

### 3.1 Naïve Bayes

The Naïve Bayes is a simple probabilistic classifier [20]. It is based on an assumption about mutual independency of attributes (independent variable, independent feature model). Usually this assumption is far from being true and this is the reason for the naivety of the method. The probabilities applied in the Naïve Bayes algorithm are calculated according to the Bayes' Rule
[21]: the probability of hypothesis *H* can be calculated on the basis of the hypothesis *H* and evidence about the hypothesis *E* according to the following formula:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \qquad (3.1.1)$$

Depending on precision of the probability model, the Naïve Bayes may give a model with high effectiveness for a supervised learning problem [18]. Frequently the Naïve Bayes uses a method of maximum likelihood (particularly in practical applications). In practice the Naïve Bayes method works effectively in various real-world situations.

The structure of a Naïve Bayes model forms a Bayesian network of nodes with one node for each attribute. The nodes are interconnected with directed edges and form a directed acyclic graph. One of the main advantages of the method is small amount of data that is required for estimation of a mean and a variance. The reason for this is the independence of the variables which is assumed. This implies only the need to determine the variances of the variables for each class – not the entire covariance matrix. The Naïve Bayes for each class value estimates whether a given instance belongs to it. The method can only represent simple distributions on the contrary to other classification methods, for instance decision trees.

The Bayesian networks were introduced in 1980s. In 1990s their first applications in medicine was shown. The Bayesian formalism is a way of representation of uncertainties what is essential during diagnosis, prediction of patients' prognosis and treatment selection [22]. With the use of this network it is possible to present the interactions among variables. The Bayesian networks are often understood as cause-and-effect relationships.

### 3.2 RBF Network

The RBF network bears a feed forward structure composed of a single hidden layer of J locally tuned units that are completely interconnected with an output layer of L linear units. n - Dimensional real valued input vector X is fed to all the hidden units simultaneously. The absence of hidden-layer weights is the primary distinction between the RBF network and MLP. The hidden-unit outputs are not determined with the aid the weighted-sum mechanism/sigmoid activation, instead the hidden-unit output $Z_j$ is obtained by closeness of the input $X$ to an n - dimensional parameter vector $\mu_j$ associated with the jth hidden unit [23]. The response characteristics of the $j^{th}$ hidden unit ( j = 1,2,......, J ) is assumed as,

$$Z_j = K\left(\frac{\|X - \mu_j\|}{\sigma_j^2}\right) \qquad (3.2.1)$$

Where K is a strictly positive radially symmetric function (kernel) with a unique maximum at its 'centre' $\mu_j$ and which drops off rapidly to zero away from the centre. The parameter σj is the width of the receptive field in the input space from unit j. This implies that $Z_j$ has an appreciable value only when the distance $\|X - \mu_j\|$ is smaller than the width σj. Given an input vector $X$, the output of the RBF network is the L -dimensional activity vector Y , whose l'h component (l = 1,2,...., L) is given by,

$$Y_l(X) = \sum_{j=1}^{J} w_{lj} Z_j(X) \qquad (3.2.2)$$

For l = 1, mapping of eq. (3.2.1) is similar to a polynomial threshold gate. However, in the RBF network, a choice is made to use radially symmetric kernels as 'hidden units'.

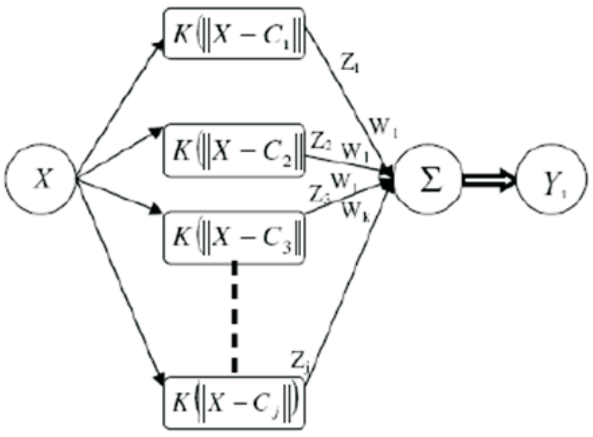


**Figure 1: Radial Basis Function Network**

RBF networks are beneficial for continuous or piecewise continuous real-valued mapping approximations f: $R^n \rightarrow R^L$ , where in n is satisfactorily small. Classification problems, an exceptional case, also come under these approximations. According to equations (3.2.1) and (3.2.2) the RBF network can be looked upon as approximating a function of interest *f (X)* by superposition of non-orthogonal, bell shaped basis functions. Three parameters namely the number of basis functions used, their location and their width [23] determine the degree of accuracy of the RBF networks.

### 3.3 Simple Logistic

Logistic regression [24] is a statistical method used to describe the relation between predictor variables denoted by $x = (x_1, x_2, \ldots, x_p)$ and a response variable, which is a categorical variable with two values (here, "survival" or "non-survival"). The conditional probability of non-survival patient can be written as $P(Y = 1|x) = \pi(x)$. Thus, the LR model for p predictor variables can be written as

$$\pi(\boldsymbol{x}) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}}{1 + e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}} \qquad (3.3.1)$$

Where $0 \leq \pi(x) \leq 1$
A useful transformation of LR is the logit transformation, defined as

$$g(\boldsymbol{x}) = ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (3.3.2)$$

The parameters $\beta = \beta_0, \beta_1, \beta_2, \ldots, \beta p$ are obtained by maximum likelihood method. This method finds the estimators of parameters that maximize the likelihood function:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad (3.3.3)$$

The log likelihood of (5) is defined as

$$L(\beta) = \ln l(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)[1-\pi(x_i)]\} \qquad (3.3.4)$$

To find the maximum likelihood estimators, L(β) is differentiated with respect to each parameter, and then the resulting terms are set as equal to zero. Other methods such as Newton's method can be utilized. The odds ratio (OR) is widely used to interpret the model. It associates with one unit change in xj represented with $e^{(\beta j)}$.

The highly correlated variable or multi collinearity problem leads to unstable parameter estimation. Ridge regression method can decrease the impact of multi collinearity in ordinary least squares regression, and is applied to logistic regression to find the ridge estimator. Ridge estimator βr is defined as [25]

$$\boldsymbol{\beta}_r = \left(\boldsymbol{X}' \boldsymbol{V}\boldsymbol{X} + k\boldsymbol{I}\right)^{-1} \boldsymbol{X}' \boldsymbol{V}\boldsymbol{X}\beta_{mle} \qquad (3.3.5)$$

Where $\beta_{mle}$ is the maximum likelihood estimator of β, V is the diagonal matrix of the maximum likelihood estimators of success probabilities, I is the identity matrix, and k is the ridge constant.

### 3.4 J48

J48 is a Java implementation of C4.5 which is a classic decision tree algorithm in machine learning. It is used to build a tree structure for classifying a data set related to a class attribute consisting of nodes and leaves. It employs Gain Ratio for selecting the best attribute from instances before applying the top-down greedy strategy to build a tree. In this way, models generated from C4.5 are easy to interpret from a tree structure and only needs a short computation time. Therefore, much research has utilized C4.5 to build the prediction models. For instance, Yao, Liu, Lei and Yin successfully exploited C4.5 to build prediction models. However, it has limitation in over fitting and time consuming in computation.

### 3.5 Decision Tree

Decision trees are one of the most frequently used techniques of data analysis. Decision trees are, among other things, easy to visualize and understand and resistant to noise in data. Commonly, decision trees are used to classify records to a proper class. Moreover, they are applicable in both regression and associations tasks. While building a decision tree it is essential to choose the best attributes to go into each of the nodes. In order to do that several mathematical formulas apply, like entropy which measures amount of information an attribute carries.

Another important aspect of decision trees construction is the problem of overtraining. The overtraining is usually observed in cases when the learning phase was performed for too long or the training examples are rare. Then the learner may adjust to the specific random features of the training data which may negatively influence its predictive power. This happens when the performance increases on the training instances but decreases on the unseen instances. The overtraining may result from a tree being too deep. In order to overcome or avoid this problem the pruning has been introduced. It relies on removing superfluous parts of a decision tree. The decision trees are successfully applied in medicine for instance in prostate cancer classification.

## 4.HEART DISEASE DATASET

The data used in this study is the Statlog (Heart) Data Set. Heart disease data set available at http:// http://archive.ics.uci.edu/ml/datasets/Statlog + (Heart). The data set has 13 attributes. However, all of the published experiments only refer to 11 of them. Consequently, to allow comparison with the literature, all the predictor and response variables which were derived from the database are given in Table 1 for reference. The data set contains 270 rows.

**Table 1: SELECTED STATLOG (HEART) DATA SET**

| Name | Type | Description |
|---|---|---|
| Age | Continuous | Age in years |
| Sex | Discrete | 1 = male<br>0 = female |
| Cp | Discrete | Chest pain type:<br>1 = typical angina<br>2 = atypical angina<br>3 = non-anginal pa<br>4 =asymptomatic |
| Trestbps | Continuous | Resting blood pressure (in mm Hg) |
| Chol | Continuous | Serum cholesterol in mg/dl |
| Fbs | Discrete | Fasting blood sugar > 120 mg/dl:<br>1 = true<br>0 = false |
| Restecg | Discrete | Resting electrocardiographic results:<br>0 = normal<br>1 = having ST-T wave abnormality<br>2 =showing probable or define left ventricular hypertrophy by Estes'criteria |
| Thalach | Continuous | Maximum heart rate achieved |
| Exang | Discrete | Exercise induced angina:<br>1 = yes<br>0 = no |
| Slope | Discrete | The slope of the peak exercise segment :<br>1 = up sloping<br>2 = flat<br>3= down sloping |
| Diagnosis | Discrete | Diagnosis classes:<br>0 = healthy<br>1= possible heart disease |

## 5.BREAST-CANCER-WISCONSIN DATASET

The data used in this study are provided by the UC Irvine machine learning repository located in breast-cancer-Wisconsin sub-directory, filenames root: breast-cancer-Wisconsin having 699 instances, 2 classes (malignant and benign), and 9 integer-valued attributes. We removed the 16 instances with missing values from the dataset to construct a new dataset with 683 instances (see Table 2). Class distribution: Benign: 458 (65.5%) Malignant: 241 (34.5%).

Table 2: BREAST CANCER DATA SET

| Attribute | Domain |
|---|---|
| 1. Sample code number | id number |
| 2. Clump Thickness | 1 - 10 |
| 3. Uniformity of Cell Size | 1 - 10 |
| 4. Uniformity of Cell Shape | 1 - 10 |
| 5. Marginal Adhesion | 1 - 10 |
| 6. Single Epithelial Cell Size | 1 - 10 |
| 7. Bare Nuclei | 1 - 10 |
| 8. Bland Chromatin | 1 - 10 |
| 9. Normal Nucleoli | 1 - 10 |
| 10. Mitoses | 1 - 10 |
| 11. Class | 2 for benign,<br>4 for malignant |

## 6.EVALUATION METHOD

This topic presents WEKA (Waikato Environment for Knowledge Analysis) version 3.6.9, the tool which is chosen in experiment to analyze medical data sets and evaluate the performance of data mining techniques applied to these sets. The selected data mining methods are presented with detailed description of parameters they use for analyses. Furthermore the measures of model performance are presented which are the bases for the comparison of methods' effectiveness and accuracy. Finally the visualization of each algorithm's performance is shown for medical data sets. This is based on own experience with WEKA environment supported with information included in.

## 7.EXPERIMENT'S RESULTS AND DISCUSSION

The heart diseases database consists of eleven conditional attributes. We analyze heart data set visually using different attributes and figure out the distribution of values.

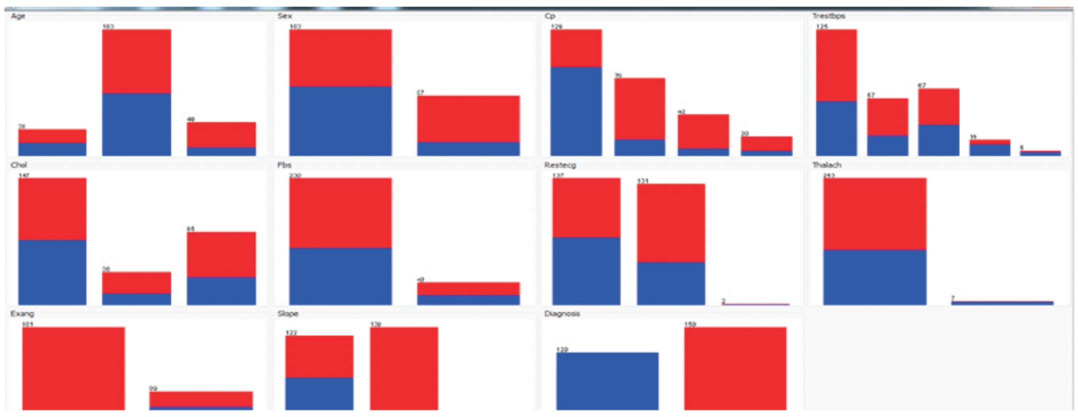


Figure 2: Distribution of the attributes of the heart diseases data

The breast cancer database consists of nine conditional attributes. The decisional attribute takes the values 0 or 1. As presented in the Figure 3 the distributions of almost all values of attributes are even. In case of almost all of the attributes the number of instances in which the attributes take the lowest values is the greatest. All conditional attributes are multi-valued.

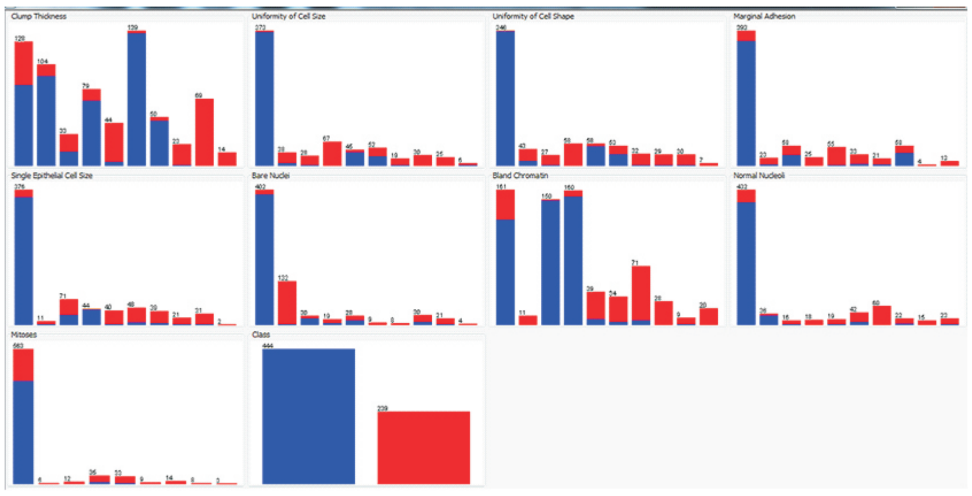


**Figure 3: Distribution of the attributes of the breast cancer data**

The results of the comparison of the algorithms are presented in the Table 3 and Table 4. The table shows the ranking of the algorithm in case of each of the performance measures and databases. The unquestionable leader in majority of cases is the Naïve Bayes. The worst results this algorithm gained for the heart disease databases. Nevertheless, overall performance was always better in comparison to other algorithms. When it comes to the RBF Network, it wins the second place in terms of the performance. For most of the databases and metrics the results gained by this algorithm were slightly worse than for the Naïve Bayes in most of the cases. The worst results this algorithm delivered for the heart data. Finally, the worst results were yielded by the decision Tree. Its results were the worst in terms of both errors and AUC in comparison to all of the algorithms. The reason for this may the nature of medical data. Its complexity and heterogeneity of values of attributes can hinder data mining.

**Table 3: Performance of the classifiers with respect to a testing configuration for the heart and breast cancer database**

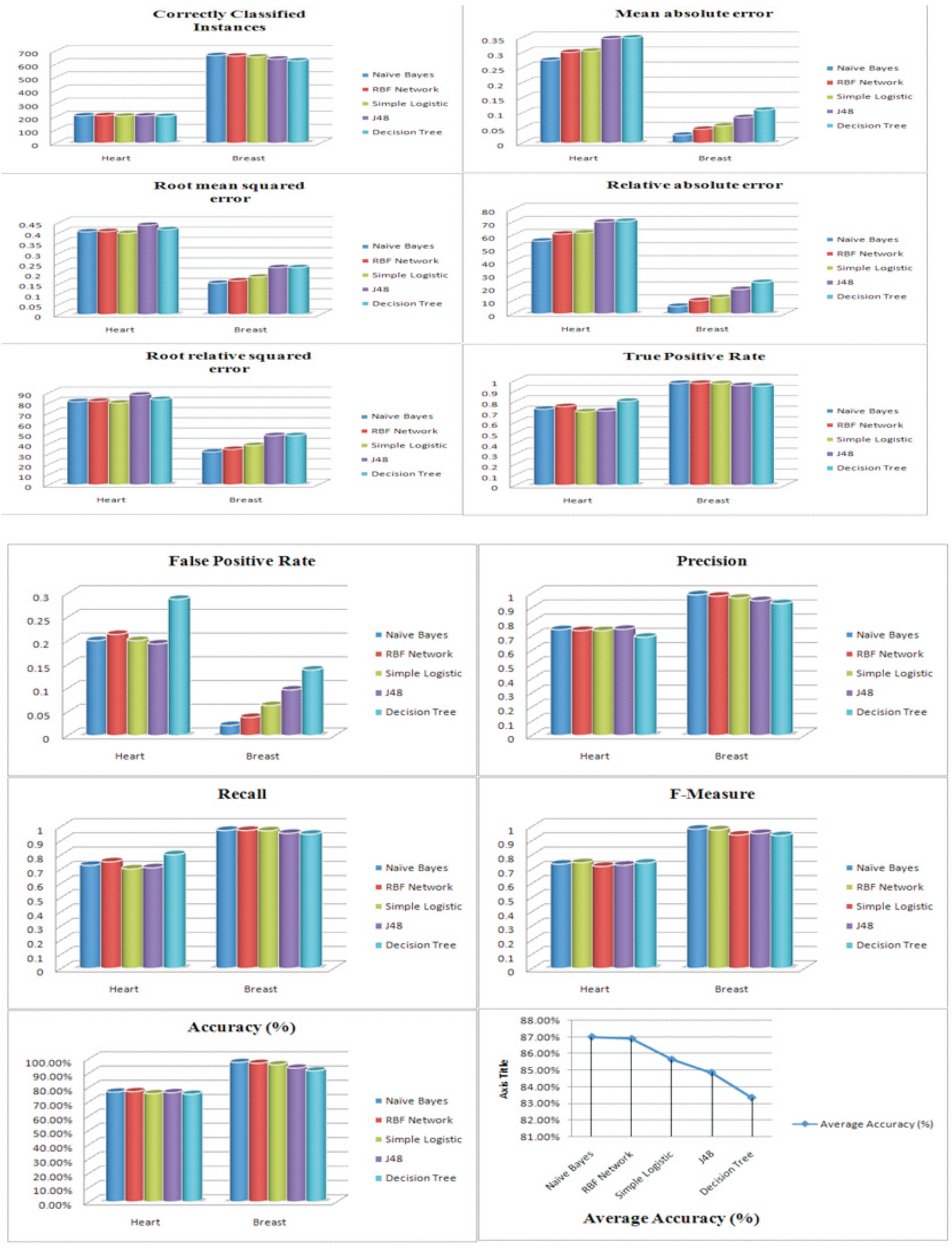
| Testing Method | Naïve Bayes | RBF Network | Simple Logistic | J48 | Decision Tree | Diseases |
|---|---|---|---|---|---|---|
| Correctly Classified Instances | 207 | 208 | 204 | 206 | 203 | **Heart** |
| | 665 | 661 | 654 | 638 | 625 | **Breast** |
| Incorrectly Classified Instances | 63 | 62 | 66 | 64 | 67 | **Heart** |
| | 18 | 22 | 29 | 45 | 58 | **Breast** |
| Kappa statistic | 0.5263 | 0.5358 | 0.5025 | 0.5176 | 0.5053 | **Heart** |
| | 0.9425 | 0.9295 | 0.9066 | 0.855 | 0.8119 | **Breast** |
| Mean absolute error | 0.2738 | 0.3004 | 0.3039 | 0.3455 | 0.3481 | **Heart** |
| | 0.0257 | 0.0455 | 0.0561 | 0.0843 | 0.1087 | **Breast** |
| Root mean squared error | 0.4037 | 0.4052 | 0.3966 | 0.4353 | 0.4141 | **Heart** |
| | 0.1524 | 0.1647 | 0.1827 | 0.2288 | 0.2289 | **Breast** |
| Relative absolute error | 55.4313 | 60.8267 | 61.5275 | 69.9475 | 70.4917 | **Heart** |
| | 5.6471 | 10.0075 | 12.325 | 18.5134 | 23.8742 | **Breast** |
| Root relative squared error | 81.2354 | 81.539 | 79.8089 | 87.5937 | 83.3399 | **Heart** |
| | 31.9596 | 34.5262 | 38.2991 | 47.9727 | 48.0005 | **Breast** |
| Accuracy (%) | 76.66% | 77.03% | 75.55% | 76.29% | 75.18% | **Heart** |
| | 97.36% | 96.77% | 95.75% | 93.41% | 91.50% | **Breast** |
| Average Accuracy (%) | 87.01% | 86.90% | 85.65% | 84.85% | 83.34% | **Combined Accuracy(Heart + Breast)** |

**Table 4: Performance of the classifiers with respect to a testing configuration for the heart and breast cancer database**

| Classifier | TP | | FP | | Precision | | Recall | | F-Measure | | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Naïve Bayes | Heart | Breast | Heart | Breast | Heart | Breast | Heart | Breast | Heart | Breast | Heart/Breast |
| | 0.725 | 0.971 | 0.2 | 0.021 | 0.744 | 0.989 | 0.725 | 0.971 | 0.734 | 0.98 | Presence of heart disease/Benign |
| | 0.8 | 0.979 | 0.275 | 0.029 | 0.784 | 0.947 | 0.8 | 0.979 | 0.792 | 0.963 | Absence of heart disease/Malignant |
| RBF Network | 0.75 | 0.971 | 0.213 | 0.038 | 0.738 | 0.98 | 0.75 | 0.971 | 0.744 | 0.975 | Presence of heart disease/Benign |
| | 0.787 | 0.962 | 0.25 | 0.029 | 0.797 | 0.947 | 0.787 | 0.962 | 0.792 | 0.954 | Absence of heart disease/Malignant |
| Simple Logistic | 0.7 | 0.968 | 0.2 | 0.063 | 0.737 | 0.966 | 0.7 | 0.968 | 0.718 | 0.967 | Presence of heart disease/Benign |
| | 0.8 | 0.937 | 0.3 | 0.032 | 0.769 | 0.941 | 0.8 | 0.937 | 0.784 | 0.939 | Absence of heart disease/Malignant |
| J48 | 0.708 | 0.95 | 0.193 | 0.096 | 0.746 | 0.948 | 0.708 | 0.95 | 0.726 | 0.949 | Presence of heart disease/Benign |
| | 0.807 | 0.904 | 0.292 | 0.05 | 0.776 | 0.908 | 0.807 | 0.904 | 0.791 | 0.906 | Absence of heart disease/Malignant |
| Decision Tree | 0.8 | 0.944 | 0.287 | 0.138 | 0.691 | 0.927 | 0.8 | 0.944 | 0.741 | 0.935 | Presence of heart disease/Benign |
| | 0.713 | 0.862 | 0.2 | 0.056 | 0.817 | 0.892 | 0.713 | 0.862 | 0.762 | 0.877 | Absence of heart disease/Malignant |

The results from the Table 3 and Table 4 have been also presented (for better visualization) in the figures in the Figure 4. These graphs confirm high performance of the Naïve Bayes in case of the breast database. However, overall best algorithm is the Naïve Bayes, with the RBF Network being the second.

Figure 4: Evaluation of performance of the data mining algorithms for the medical databases



The analysis delivered interesting results. The best classifier was the Naïve Bayes. Its overall performance turned out to be the highest in case of most of the databases. This may be caused by the nature of data being complex could have caused overtraining of the four other algorithms. The second place was won by the RBF Network. Its general performance was only slightly worse than Naïve Bayes'. On the third and fourth position were Simple logistic and J48. The worst algorithm, in turn, was the Decision Tree. Its poor performance was expressed by high rate of errors and low values of other metrics. It may have been caused by overtraining of the tree which was very complex in most of the cases.

## 8.CONCLUSIONS

Nowadays, huge amount of data is gathered in medical databases. Such databases may contain valuable information contained in nontrivial dependencies among symptoms and diagnoses. With the use of medical systems the process of uncovering such relationships in historical data is much easier to conduct. This knowledge can be used in diagnosis of future cases.

The main goal of the research paper was to identify the most common data mining algorithms, implemented in modern Medical Diagnosis, and evaluate their performance on several medical datasets. Five algorithms were chosen: Naïve Bayes, RBF Network, Simple Logistic, J48 and Decision Tree. For the evaluation two UCI databases were used: heart disease and breast cancer datasets. Several performance metrics were utilized: percent of correct classifications, True/False Positive rates, AUC, Precision, Recall, F-measure and a set of errors.

## 9.REFERENCES

1.Fayyad U, PiatetskyShapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996 Fal;17(3):37- 54.

2.Hu, F. (1999). Findings from the Nurses' Health Study presented at the 72nd Scientific Sessions of the American Heart Association, Atlanta, GA, November 8.

3.HianChyeKoh and Gerald Tan, ―Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Vol 19, No 2.

4.PrasannaDesikan, Kuo-Wei Hsu, JaideepSrivastava, ―Data Mining For Healthcare Management‖, 2011SIAM International Conference on Data Mining, April, 2011.

5.Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology,
Vol. 64, No. 3, 1987, pp. 304-310.

6.Yao, Z.; Lei, L.; Yin, J., "R-C4.5 Decision tree model and its applications to health care dataset". Proceedings of International Conference on Services Systems and Services Management 2005, pp. 1099-1103.

7.Das, R.; Abdulkadir, S. (2008). "Effective diagnosis of heart disease through neural networks ensembles". Elsevier, 2008.

8.Colombet, I.; Ruelland, A.; Chatellier, G.; Gueyffier, F. (2000). "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression". Proceedings of AMIA Symp 2000, p 156-160.

9.Avci, E.; Turkoglu, I., "An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases". Journal of Expert Systems with Application, Vol. 2, No. 1, 2009, pp. 2873-2878.

10.Kurt, I.; Ture, M.; Turhan, A., "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". Journal of Expert Systems with Application, Vol. 3, 2008, pp. 366-374.

11.Gennari, J., "Models of incremental concept formation". Journal of Artificial Intelligence, Vol. 1, 1989, pp. 11-61.

12.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. 4,2013, pp 56-66. ISSN: 2296-1739, Helvetic Editions LTD, Switzerland.

13.V. Chauraisa and S. Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques", Caribbean Journal of Science and Technology, Carib.j.SciTech, Vol.1, pp. 208-217, 2013. ISSN 0799-3757, West Indies.

14.Jinyan LiHuiqing Liu, See-Kiong Ng and Limsoon Wong," Discovery of significant rules for classifying cancer diagnosis data", Bioinformatics 19(Suppl. 2)Oxford University Press 2003.

15.Dong-Sheng Cao, Qing-Song Xu ,Yi-Zeng Liang, Xian Chen, "Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity", Chemometrics and Intelligent Laboratory Systems.

16.Liu Ya-Qin, Wang Cheng, Zhang Lu," Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data", 3rd International Conference on Bioinformatics and Biomedical Engineering , 2009.

17.Tan AC, Gilbert D. "Ensemble machine learning on gene expression data for cancer classification", Appl Bioinformatics. 2003;2(3 Suppl):S75-83.

18.V. Chauraisa and S. Pal, “Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability”, (IJCSMC) International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 1, January 2014, pg.10 – 22, ISSN 2320–088X, 2014, India.
19.V. Chauraisa and S. Pal, “A Novel Approach for Breast Cancer Detection using Data Mining Techniques”, (IJIRCCE) International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue. 1, January 2014, pg.2456 – 2465, ISSN(Online): 2320-9801, 2014, India.
20.Nong Y., The Handbook of Data Mining. Lawrence Earlbaum Associates, 2003
21.Tang Z., MacLennan J., Data Mining with SQL Server 2005. Indianapolis, Indiana, USA, Wiley Publishing Inc., 2005.
22.Lucas P.J.F., Boot H., Taal B.G., Decision-theoretic network approach to treatment management and prognosis. Knowledge-based Systems, 1998, vol. 11, 321–330.
23.P. Venkatesan and S. Anitha, "Application of a radial basis function neural network for diagnosis of diabetes mellitus", Current Science, vol. 91, no. 9, pp. 1195 - 1199, 10 November 2006.
24.Hosmer DW, Lemeshow S: Applied logistic regression (2nd Ed.). New York, USA: A Wiley-Interscience Publication, John Wiley & Sons Inc.; 2000.
25.Barker L, Brown C: Logistic regression when binary predictor variables are highly correlated. Stat Med 2001, 20:1431–1442.