ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2013.0081

CLINICAL DATABASE ANALYSIS USING DMDT BASED PREDICTIVE MODELLING

Srilakshmi Indrasenan[1] and Sathiyabhama Balasubramaniam[2]
Department of Computer Science and Engineering, Sona College of Technology, India
E-mail: [1]srilakshmi.indrasenan@gmail.com and [2]sathiyabhama@sonatech.ac.in

Abstract
In recent years, predictive data mining techniques play a vital role in the field of medical informatics. These techniques help the medical practitioners in predicting various classes which is useful in prediction treatment. One of such major difficulty is prediction of survival rate in breast cancer patients. Breast cancer is a common disease these days and fighting against it is a tough battle for both the surgeons and the patients. To predict the survivability rate in breast cancer patients which helps the medical practitioner to select the type of treatment a predictive data mining technique called Diversified Multiple Decision Tree (DMDT) classification is used. Additionally, to avoid difficulties from the outlier and skewed data, it is also proposed to perform the improvement of training space by outlier filtering and over sampling. As a result, this novel approach gives the survivability rate of the cancer patients based on which the medical practitioners can choose the type of treatment.

Keywords
Predictive Data Mining, Breast Cancer, Survivability Rate, Outlier Filtering, Diversified Multiple Decision Tree

1. INTRODUCTION

Data mining is a field of computer science in which extracting of data or knowledge is done [14]. The extraction is done from a huge amount of data. This can also be referred as knowledge discovery. Data mining consists of various sequences of steps. It starts with the process of gathering data from any dataset and is followed by cleaning, pre processing, transformation, mining, evaluation and representation.

Clinical databases generally possess large amount of information about the patients and their diagnosis reports. It reduces the workload of the medical practitioners since this method is active and semi-automated [15] in contrary to the passive and manual method of hand written patient records. There are numerous tools to evaluate and analyse the patients' data after it is stored in the databases. These tools help in detecting the disease, its progress and various other features. However, the research is always confined in analysis of the clinical databases [1] and not on prediction of the survivability rates of the patients involved.

1.1 CANCER DATABASE ANALYSIS USING PREDICTIVE DATA MINING TECHNIQUES

Predictive data mining is a field of data mining that automatically creates a classification model from a given set of examples once when the model is built it can be used to predict the classes of other examples automatically. The outcome is generally because of certain probability or on basis of detection theory. Many models and classifiers exist for predictive data mining namely k-Nearest Neighbours (k-NN), Naïve Bayes, Logistic Regression, and Majority Classifiers. The predictive data mining techniques has varied applications ranging from archaeology, medical sciences and bioinformatics. The predictive modelling is mainly preferred for bio medical research to target best diagnosis and appropriate treatment, developed detailed patients' and disease profile and also to diagnose and prevent diseases appropriately. The predictive data mining technique allows the physician/surgeon [8] to,

- Predict the disease through symptoms quickly and accurately.
- Predict the survivability rate of the patients.
- Find the intensity of the diseases more precisely.

Thus predictive data mining helps the medical practitioners to decide about the type of treatment based on the predictions.

1.2 BREAST CANCER

Breast cancer is a major concern of any country in the world today. Breast cancer is the most frequently diagnosed form of cancer and is the second leading cause of death in women [17]. Even though in the east few years' research and its treatments were increased, still it is a major cause of cancer in the medical field.

Breast cancer is a form of a malignant tumour that develops when cells present in the tissues of breast multiplies by dividing without normal controls of call death and division. In women this is the most common form of cancer [17]. The treatment for breast cancer may be broadly classified as local and systematic. While surgery and radiation comes under local category, chemotherapy and hormone therapy are considered systematic [12].

1.3 MULTIPLE DECISION TREES

Generally, predictive data mining is done through single decision tree classifier namely the naive bayes, C4.5. However, these decision tree classifiers utilize only one of the attribute in the dataset. The attribute is known as the vital data. Based on the vital data [9] only the prediction is done. However, cancer is a disease which develops day by day and minute by minute, therefore many of the attributes in the dataset although being trivial will contribute [13] in the prediction of the intensity of the disease and the survivability rate.

2. RELATED WORK

Abdelghani Bellaachia and Erhan Guven conducted an analysis for the prediction of rate of survivability in breast cancer patients using predictive data mining techniques [9]. The data which has been used is from the SEER Public-Use Data. The dataset contained 16 fields and almost 150000 data has been considered. They investigated three most commonly used data

mining techniques namely the Naïve Bayes, the C4.5 decision tree and the back-propagated neural network algorithms. Various experiments were done using these algorithms. The results were compared to existing systems. However, it was found that C4.5 algorithm had a high performance in comparison to the other two techniques.

Three popular data mining algorithms [12] namely the artificial neural networks (ANN), decision trees and logical regression are analysed for prediction of breast cancer survivability rate prediction among various patients. Out of these decision trees were found out to be the best in precision and accuracy.

For erythemato-squamous disease an action rule based decision system was proposed [1]. Adaptive Neuro Fuzzy Inference System (ANFIS) is used to identify the disease correctly from the given set of symptoms. The ANFIS is the highest possible approach in knowledge discovery [3].

Different descriptive and predictive data mining techniques were deployed for diagnosing heart disease [7]. Almost fifteen attributes are needed for predicting the heart disease. Out of the various methods and approaches deployed decision tree classification was found out to be the most efficient and precise approach in predicting the disease.

Godswill Chukwugozie Nsofor compares five different predictive data mining techniques and five different data sets have been used [6]. There are a few non-linear types predictive data mining problems and a blend of linear and non-linear algorithm are best for them. He discusses about 4 linear and one non linear technique.

Riccardo Bellazzi and Blaz Zupan [8] concentrates on the recent research trends in predictive data mining and publishes the current issues and the guidelines for handling them. It also explains the need for data mining technique in medical field.

Jaree Thongkam, et al. focuses on the main problem in breast cancer survivability predictions. The main obstacle that becomes a hindrance in calculating survivability is the presence of outlier and skewed data present in the datasets. Therefore [11] proposes two approaches for removing the outlier and skewed data namely the C- Support Vector Classification (C-SVC) algorithm for filtering the data and an over sampling approach to increase the number of instances in the minority class.

Marko NF, et al. [16] proposes a nomogram based approach for prediction of breast cancer in women and also to display the survivability ratio in a better manner.

The system used in [1] [3] [6-9] [11 -12] provides the various clinical database analysis and its merits and demerits. It is also clear that predicting breast cancer survivability is highly concentrated but the solution is not clear. The system design proposed will deal exclusively on predicting the survivability rates of breast cancer patients. The system will find status of the patients' disease intensity accurately since it uses more than one attributes to predict the survival rate.

3. PROPOSED SYSTEM

To predict the cancer survivability, the data sets if breast cancer patients are gathered. The medical practitioner will collect the patients' personal information like name, age, address and occupation in addition to the cancer details diagnosed and the attributes related to it. In this proposed system the data gathered is given into pre processing by outlier filtering and over sampling.

The outlier and skewed data are removed by outlier filtering using C-Support Vector Classification (C-SVC) algorithm and any imbalance is balanced through the non-heuristic process of over sampling. The Fig.3 explains how this process is done. Apart from filtering and sampling the data also undergoes 10-fold cross validation.

The pre processed data is used to predict the survivability rates based on DMDT algorithm shown in Fig.2. The algorithm uses the concept of multiple decision tree and it construct trees based on more than one attribute.

3.1 SYSTEM ARCHITECTURES

The following diagram specifies the system architecture of the proposed model. The medical practitioner notes the patient general information. The data is pre processed and the predictive data mining is done through DMDT algorithm.

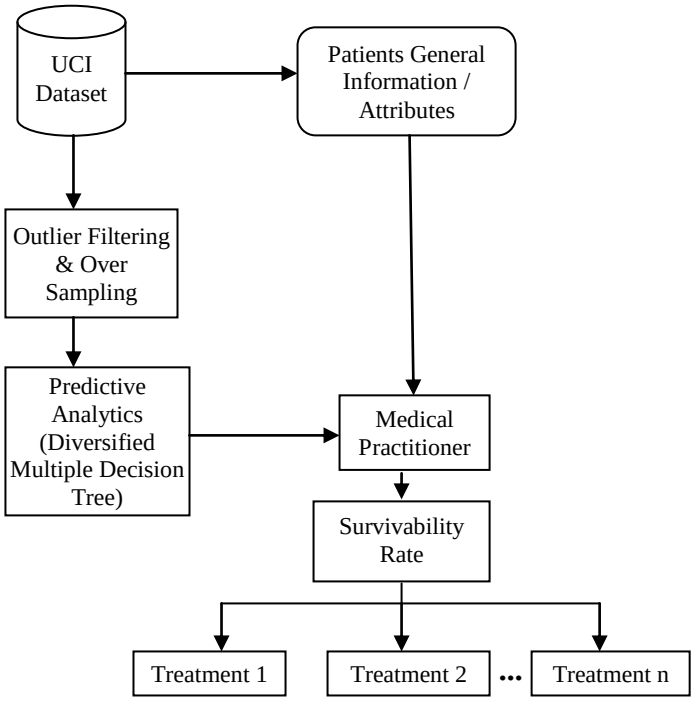

Fig.1. System Architecture

3.2 DIVERSIFIED MULTIPLE DECISION TREE (DMDT) ALGORITHM

The Diversified Multiple Decision Tree algorithm is as shown below. The algorithm consists of two parts namely the training phase and classification phase. In the training phase the microarray data is given as input and set of disjoint trees are produced as output. The trained trees (output of training phase) is given to classification phase and the output is class x. This class is the one which is used to predict the survivability rate of breast cancer patients.

```
Training (D,T,n)
INPUT: Dataset D and number of trees n.
OUTPUT: A set of disjointed trees T
Let T=ø;
For i = 0 to n-1 do
        call C4.5 to build tree T_i on D;
        remove genes used in T_i from D;
        T=T U T_i.
End for
Output T;

Classification (T, x, n)
INPUT: A set of trained trees T, a test sample x,
and the number of trees n.
OUTPUT: A class label of x
Let vote(i) = 0 where i = 1 to c = the number of
classes.
For j=1 to n do
        Let c be the class outputted by T_j;
        vote(c) = vote(c) + accuracy(T_j);
End for
Output c that maximizes vote(c);
```

Fig.2. Diversified Multiple Decision Tree (DMDT) Algorithm

Generally for any type of medical patient records only the persons' most vital attribute for the disease is utilized for prediction of the disease. However, this is very misleading especially for cancer dataset, because in course of time the tumour cells develop and spread hence making other cell attributes of the affected part also an important factor in predicting the survivability. If more than one attribute set is to be considered then a single decision tree classifier will be inefficient. Hence a diversified multiple decision tree classifier is used.

3.3 IMPROVING TRAINING SPACE

In this step, the data set collected are filtered using C-Support Vector Classification Algorithm (C-SVC) algorithm so that any sort of outlier data will be removed. The C-SVC algorithm checks the dataset with the k-nearest neighbours and removes any outlier data which is present.

Over sampling is one of the non-heuristic approaches which are used to balance the imbalanced data by increasing the size of minority classes.

Outlier data are those which may contain certain instances which are wrong and may affect the models' performance. Therefore, outlier filtering by C-SVC algorithm is used to filter these types of data. The over sampling approach is also done along with outlier filtering because the number of instances in few classes (minority classes) may be less than that of other classes (majority classes) [11]. Fig.2 illustrates C-SVC algorithm [11]. Fig.3 is the block diagram for the process adapted from Jaree Thongkam et al. [11].

There is the need to adapt both outlier filtering and over sampling because the hybrid approach significantly improves over the datasets than being done individually or any other procedure like under sampling. Table.1 illustrates that the combined approach has a significant improvement in the performance than the separate approaches of outlier filtering and over sampling.

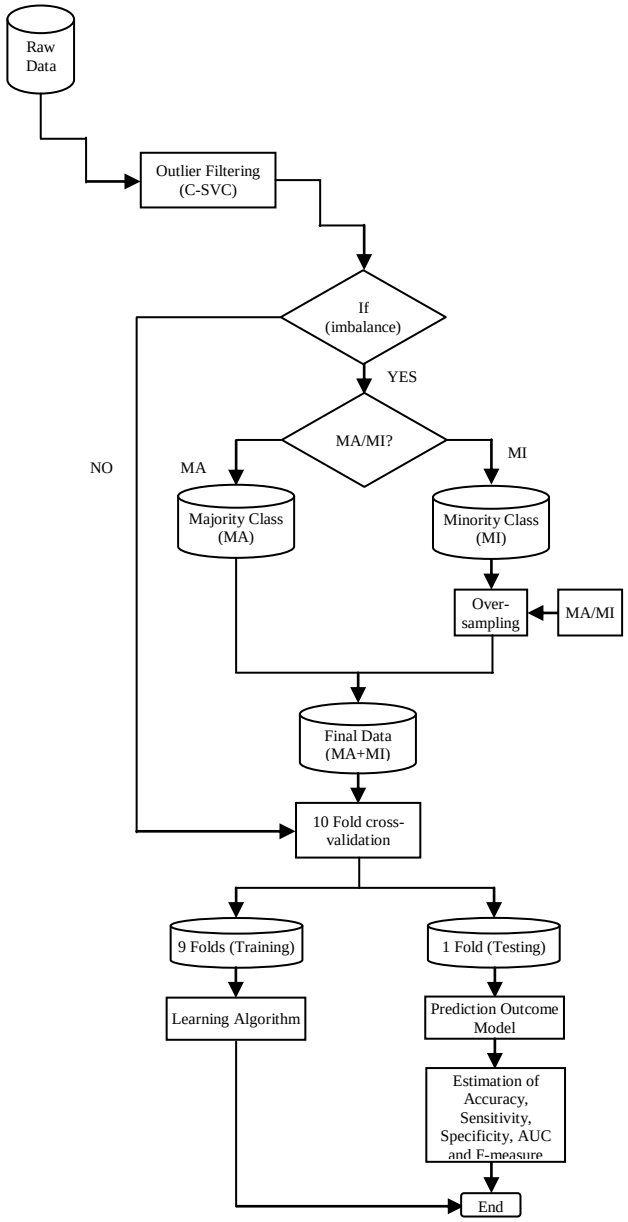

Fig.3. Outlier Filtering and Over Sampling

3.4 STATISTICAL HYPOTHESIS TESTING

In the proposed system the decision variable is considered to be the attributes of the cells that cause tumor (cancer) to spread. By keeping that as the decisive variable and other attributes as inputs, chi-square test calculates a value called *p* by comparing the value of the statistic to a chi-squared distribution. The number of degrees of freedom is equal to the number of cells ***n***, minus the reduction in degrees of freedom, *p*.

Definition: Pearson's chi-squared test (χ^2) is the best-known of statistical procedures [18], whose results are evaluated by reference to the chi-squared distribution.

Lemma: A test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution. (Goodness of Fit test for χ^2).

Proof:

In this case N observations are divided among ***n*** cells of the dataset table. In the proposed model **N** is the number of patient datasets observed and n is the attributes considered. A simple application is to test the hypothesis that, in the general population, values would occur in each cell with equal frequency. The "theoretical frequency" for any cell (under the null hypothesis of a discrete uniform distribution) is thus calculated as,

$$E_i = \frac{N}{n}, \quad (1)$$

and the reduction in the degrees of freedom is $p = 1$, notionally because the observed frequencies $\boldsymbol{O_i}$ are constrained to sum to $\boldsymbol{N}$.

The value of the test-statistic [18] is

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

where,

X^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution

O_i = an observed frequency

E_i = an expected (theoretical) frequency, asserted by the null hypothesis

n = the number of cells in the dataset table

4. MATERIALS AND METHODS

The data which has been used for our study is taken from the UCI [19] data set. The data set under consideration has 586 records which consist of all the 16 available fields from the UCI database for cancer patients. The dataset is of breast cancer patients. The data set is chosen with utmost care because the breast cancer details are sensitive and vulnerable.

Weka toolkit is used to perform the classification, association rules and visualization for the given data. The toolkit is developed using Java. Java swing is utilized for GUI front end and Oracle 10g is used to store and manage the dataset.

The dataset obtained is been pre processed and the training space is improved using outlier filtering and over sampling approach. The approach uses a C-SVC algorithm [11] that removes any outlier i.e. ungrouped data. The steps involved in the improvement of training space are as follows,

- Removing outlier values from both dead and alive classes of the data set.
- The data sets are categorized into major and minor classes.
- Size of the minority classes are improved by over sampling.
- The majority and minority classes are combined in order to form a balanced data set.

DMDT algorithm can be used to predict the survivability accurately because the ratio between majority and minority classes are almost 1 after applying the above specified steps.

Table.1. Improving Training Space

Data Sets	Ratio of MA/MI			
	Original	Outlier Filtering	Over Sampling	Combined
1 – 10	321	1256	1.00	1.00
11 – 20	186	503	0.99	1.00
21 - 30	135	201	1.00	1.00
31 - 40	246	815	0.99	1.00

Table.1 shows the ratio between the majority and minority classes. The optimal ratio should be 1 in order to get the survivability rates of the patients accurately. However it is clearly visible from Table.1 that the ratios are so high in case of original data as well as outlier filtered data. However in over sampled data the ratio is almost near 1. To obtain a perfect equivalent number of majority and minority classes a combined approach which yields 1 as the ratio between MA and MI across all the data sets is used.

On the improved training space the DMDT algorithm is applied which is again done with the help of Weka ensemble toolkit. In Weka toolkit the J48 algorithm is selected for this process. J48 algorithm is well known and is accurate for decision based classification. The process of multiple classifications in the proposed model is done iteratively for the different trivial and vital attributes such as tumour cell condition, patient history, smoking habits, previous treatments. The DMDT algorithm works as shown in Fig.2. The output of DMDT on the improved training space gives a class label of the test sample x.

The class label is classified from a set of trained trees. Therefore output of the DMDT phase after the improved trained space phase provides a class label x which predicts the survivability rate of the given breast cancer patient. This rate of survivability is essential to select the correct type of treatment for curing the disease.

5. RESULTS AND DISCUSSION

The important benefit of the proposed system is the improved training space which on applying DMDT algorithm gives a class label x that predicts the survivability rate of the given breast cancer patients. The outcome of the system predicts the survivability rate of the patients which is very essential in decision of the type of treatment to be given. When the survivability is clearly known the patients' can be saved from the brutal disease with ease.

The results are given to the medical practitioner who analyses and categorizes patients based on their survivability rate and other factors and start giving appropriate treatments.

The performance of this system can be computed by following measures;

- Accuracy
- Sensitivity
- Specificity
- AUC (Area Under the ROC (Receiving Operating Characteristics) Curve)
- F-Measure

The measures can be given by the following formula;

$$\text{Accuracy (AC)} = \frac{(TP+TN)}{(TP+TN+FN+FP)} \quad (3)$$

$$\text{Sensitivity (SE)} = \frac{TP}{(TP+FN)} \quad (4)$$

$$\text{Specificity (SP)} = \frac{TN}{(TN+FP)} \quad (5)$$

The AUC can be represented by the following graph;

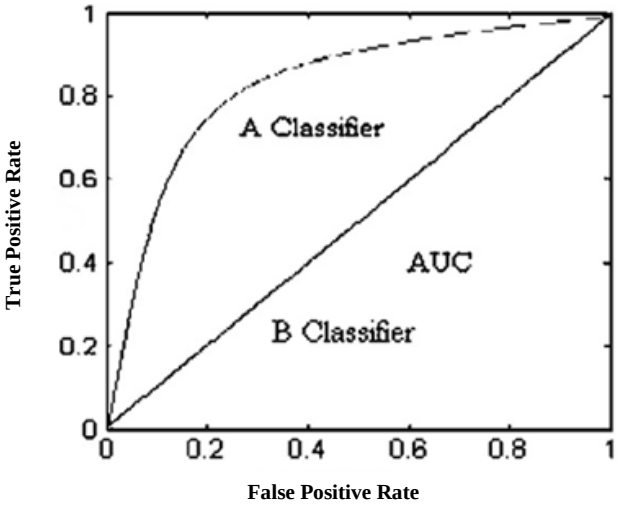

Fig.4. Area under the ROC Curve (AUC)

Area under ROC Curve is the region under the ROC curve. ROC is the graph across the True Positive Rate and False Positive Rate.

F-Measure is specified by the following formulae;

F-Measure = 2PR/ (P+R); where P – Precision and R- Recall.

Based on the above discussed measures of performance the DMDT algorithm is compared with the normal decision tree classification algorithm and the results are tabulated.

Table.2. Test Case Generation

	Diseases present	**Diseases absent**
Test positive	50(directly) 30(TP)	5(FP)
Test negative	10(FN)	50/30/5(TN)
Total	90	90

Table.3. Performance of C4.5 Decision Tree and DMDT

No. of Data	**C4.5**		**DMDT**	
	Sensitivity	**Specificity**	**Sensitivity**	**Specificity**
10	99	8	100	3
20	97	8	98	4
40	94	10	97	3
80	97	8	99	4
100	98	7	98	4

The Table.3 clearly explains that DMDT is better than normal decision tree classification because the sensitivity is high and the specificity is low. Although, the differences between the two decision trees' performance is minimal, the difference matters when it comes to accuracy since the DMDT algorithm improves the performance and accuracy more than the best known[12] decision tree algorithm C4.5. DMDT uses C-SVC algorithm for outlier filtering and multiple decision tree classifier. The graph for the performance comparison between various algorithms is shown below,

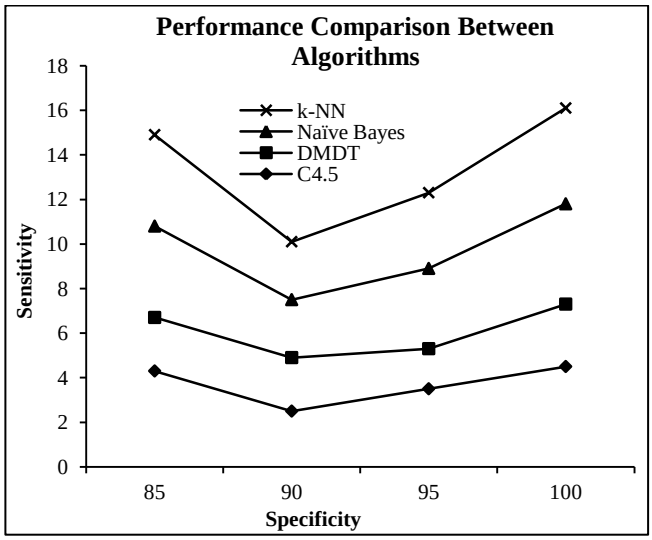

Fig.5. Performance Graph

In X-axis and Y-axis, specificity and sensitivity values have been taken into account respectively. When specificity increases, the sensitivity decreases. The inverse linear relationship is clearly inferred from the ROC curve. Hence one can clearly say that as true positive rate rises, the false negative rate falls automatically. In conclusion, a good system is the one which has high sensitivity and less specificity.

For an efficient system the F-Measure should be high. While it is proved that the DMDT algorithm is best using basic measures of performance it is proved that over sampling and outlier filtering is the best option for improving the training space through F-Measure as shown in Table.4. The improved training space directly has an influence on the performance of the model. Since the outliers are filtered as well the over sampled.

Table.4. F-Measure for the Improved Training Space

Data Sets	**F-Measure**			
	Original	**Outlier Filtering**	**Over Sampling**	**Combined**
1-10	35.78	73.89	66.09	92.00
11-20	46.33	67.90	64.01	92.22
21-30	55.80	79.05	71.06	87.36
31-40	66.66	91.70	77.89	96.05

It can be inferred from the Table.4 that the combined approach of outlier filtering and over sampling has a high F-Measure which means the system has high accuracy and precision.

6. CONCLUSION AND FUTURE WORK

This system generates the survivability rate of the breast cancer patients through predictive data mining. Breast cancer is a tough fight to battle for both the doctors as well as the patients however with the type of treatment known precisely the patients and the medical practitioners can effectively fight against the breast cancer. To choose the treatment the survivability rates are very essential. This proposed system finds gets data from UCI data set and it improves the training space using outlier filtering and over sampling and predicts the survivability rate by DMDT algorithm. DMDT algorithm utilizes more than one data attribute for the predictive analytics. Therefore it provides precise and reliable survivability rate to the medical practitioners. This system can be further improved by hybridizing this DMDT based predictive data mining technique with a combination of k-NN (k-Nearest Neighbors) and Majority Classifiers to improve the precision even further because k-NN algorithm helps in classification and pattern recognition which is very vital for cancer dataset.

REFERENCES

[1] T. Deepa, B. Sathiyabhama, J. Akilandeswari and N. P. Gopalan, "Knowledge Management Techniques for Analysis of Clinical Databases", *Lecture Notes in Computer Science, , Advanced Computing, Networking and Security*, Vol. 7135, pp. 198 - 206, 2012.

[2] X. Y. Djam and Y. H. Kimbi, "A Medical Diagnostic Support System for the Management of Hypertension (MEDDIAG)", *Journal of Science and Multidisciplinary Research*, Vol. 3, pp. 16 – 30, 2011.

[3] A. Abraham and B. Nath, "Hybrid intelligent systems: A review of a decade of research", *Technical Report Series, School of Computing and Information Technology, Faculty of Information Technology*, Vol. 5, pp. 1 – 55, 2000.

[4] R. Brause and F. Friedrich, "A Neuro-fuzzy Approach as Medical Diagnosis interface", *Proceedings of European Symposium on Artificial Neural Networks*, Vol. 1, pp. 201 – 206, 2000.

[5] X. Y. Djam and Y. H. Kimbi, "Fuzzy Expert System for the Management of Hypertension", *The Pacific Journal of Science and Technology*, Vol. 12, No. 1, pp. 390 – 402, 2011.

[6] Godswill Chukwugozie Nsofor, "A Comparative Analysis of Predictive Data-Mining Techniques", A Thesis Presented for the Master of Science Degree, The University of Tennessee, Knoxville, 2006.

[7] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", *International Journal of Computer Applications*, Vol. 17, No. 8, pp. 43 – 48, 2011.

[8] Riccardo Bellazzi and Blaz Zupan, "Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines", *International Journal of Medical Sciences*, Vol. 77, No. 2, pp. 81 – 97, 2008.

[9] Abdelghani Bellaachia and Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining*, 2006.

[10] B. Devendra Rao, B. Sathiyabhama, J. Akilandeswari and N. P. Gopalan, "Actionable Knowledge Discovery for Clinical Decision Support System", *Proceedings of the IEEE and CSI Sponsored National Conference on Computational Intelligence, Security and Systems*, Vol. 2, pp. 130 – 136, 2011.

[11] Jaree Thongkam, Guandong Xu, Yanchun Zhang and Fuchun Huang, "Toward breast cancer survivability prediction models through improving training space", *An International Journal on Expert Systems with Applications*, Vol. 36, No. 10, pp. 12200 – 12209, 2009.

[12] Dursun Delen, Glenn Walker and Amit Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", *Journal on Artificial Intelligence in Medicine*, Vol. 34, No. 2, pp. 113 – 127, 2004.

[13] Hong Hu, Jiuyong Li, Hua Wang, Grant Daggard and Mingren Shi, "A Maximally Diversified Multiple Decision Tree Algorithm for Microarray Data Classification", *Australian Computer Society, Inc, Conferences in Research and Practice in Information Technology (CRPIT)*, Vol. 73, 2006

[14] Zengyou He, Xiaofei Xu and Shengchun Deng, "Data mining for Actionable Knowledge: A Survey", *High Technology and Development Program of China*, 2003.

[15] Qiao Yang, "A multi-agent prototype system for helping medical diagnosis", *Memorial University of Newfoundland*, 2008.

[16] N. F. Marko, Xu Z, Gao T, M. W. Kattan and R. J. Weil, "Predicting survival in women with breast cancer and brain metastasis: A nomogram outperforms current survival prediction models", *Cancer Research, Cambridge University*, Vol. 118, No. 15, pp. 3749 – 3757, 2012.

[17] American Cancer Society, "Breast Cancer Facts & Figures 2005-2006", *Atlanta: American Cancer Society, Inc.*, 2005.

[18] Harald Cramer, "*Mathematical Methods of Statistics*", Princeton University Press, 1948.

[19] David Aha, "*UCI Machine Learning Repository*", Centre for Machine Learning and Intelligent System, Available at http://archive.ics.uci.edu/ml/datasets.html, 1987.