Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/01/jls.htm
2015 Vol.5 (S1), pp. 4330-4339/Karim
Research Article

A COMPARITIVE SURVEY ON DATA MINING TECHNIQUES FOR BREAST CANCER DIAGNOSIS AND PREDICTION

***Hamid Karim Khani Zand**
Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
**Author for Correspondence*

ABSTRACT
Breast cancer is one of the deadliest diseases, is the most common of all cancers and is the leading cause of cancer deaths in women worldwide. The classification of Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. In this paper we present a comparative survey on data mining techniques in the diagnosis and prediction of breast cancer and also an analysis of the prediction of survivability rate of breast cancer patients. The data used is the SEER Public-Use Data.

Keywords: *Data Mining, Technique, Breast Cancer, Diagnosis, Prediction, Method, SEER*

INTRODUCTION
Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. Breast cancer is one of the most common cancers among women. Breast cancer is one of major causes of death in women when compared to all other cancers. Cancer is a type of diseases which causes the cells of the body to change its characteristics and cause abnormal growth of cells. Most types of the cancer cells eventually become a mass called tumor. The occurrence of the breast cancer is increasing globally. It's a major health problem and represents a significant worry for many women (Chaurasia and Pal, 2014). Early detection of breast cancer is essential in reducing life losses. Earlier treatment, however, requires the ability to detect breast cancer in early stages. Early diagnosis requires accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones. Automatic diagnosis of breast cancer is an important, real-world medical problem. Therefore, finding an accurate and effective diagnosis method is very important. Recently machine learning methods have been widely used in prediction, particularly in medical diagnosis. Medical diagnosis is one of major problems in medical application (Liou and Chang, 2015).
The classification of Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. One major class of problems in medical science involves the diagnosis of disease, based upon different tests performed upon the patient. Because of this reason the use of classifier systems in medical diagnosis is gradually increasing (Eshlaghy *et al.,* 2013).
Predicting outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. Use of computers with automated tools, large volumes of the medical data are being collected and made available to the medical research groups. As a result, data mining techniques has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict outcome of a disease using the historical datasets (Saleema *et al.,* 2014).

Literature Review
Bellaachia and Guven (2006) used the SEER data to compare three prediction models for detecting breast cancer. They have reported that C4.5 algorithm gave the best performance of 86.7% accuracy.
Delen *et al.,* (2005) in their work preprocessed the SEER data for to remove redundancies and missing information. They have compared predictive accuracy of the SEER data on three prediction models indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample.
Choi *et al.,* (2009) implemented common machine learning algorithms to predict survival rate of breast cancer patient. This study is based upon data of SEER program with high rate of positive examples (18.5

%). Logistic regression had the highest accuracy, artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity Kotsiantis and Pintelas (2004) did a work on Boosting, Bagging and Combination of Bagging and Boosting as a single ensemble using different base learners such as Naïve Bayes, C4.5, One R and Decision Stump. These were experimented on several benchmark datasets of UCI Machine Learning Repository.

Breast Cancer

Breast cancer is a malignant tumor which develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. It is the most common cancer among women. Although scientists do not know the exact causes of most breast cancer, they know some of the risk factors that increase the likelihood of a woman developing breast cancer. These factors contain such attributes as age, family history and genetic risk.

Treatments for breast cancer are separated into two main types, systematic and local. Surgery and radiation are examples of local treatments whereas chemotherapy and hormone therapy are examples of systematic therapies. Usually for the best results, the two types of treatment are used together.

Although breast cancer is the second leading cause of cancer death in women, but the survival rate is high. With early diagnosis, 97% of women survive for 5years or more (Jerez-Aragonés *et al.*, 2003).

Data Mining Classification Methods

The data mining consists of various methods. Different methods serve various purposes, each method offering its own advantages and disadvantages. Classification and clustering are the two most common techniques of data mining which are used in field of medical science. However, most data mining methods commonly used are of classification category as the applied prediction techniques assign patients to either a ”benign” group that is non- cancerous or a ”malignant” group that is cancerous and generate rules for the same. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems. In data mining, classification is one of the most important tasks. It maps data in to predefined targets. It’s a supervised learning as targets are predefined. The aim of classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification tasks can be classified into the following groups.

Naive Bayes (NB)

The Naive Bayes is a quick method for creation of statistical predictive models. NB is based on the Bayesian theorem. This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class. During training, the probability of each class is computed by counting how many times it occurs in the training dataset. This is called the “prior probability” P(C=c). In addition to prior probability, the algorithm also computes probability for the instance x given c with the assumption that the attributes are independent. This probability becomes the product of the probabilities of each single attribute. Then the probabilities can be estimated from the frequencies of the instances in the training set.

Decision Trees (C4.5)

Decision tree is a tree where each non-terminal node represents a test or decision on the considered data item.

Choice of a certain branch depends upon the outcome of test. To classify a particular data item, we start at root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees also can be interpreted as a special form of rule set, characterized by their hierarchical organization of rules.

Neural Networks

Neural networks (NN) are those systems modeled based on the human brain working. As the human brain includes millions of neurons that are interconnected by synapses, neural network is a set of connected input/output units in which each connection has a weight associated with it. Network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input.

Techniques for Breast Cancer Diagnosis

Clinical diagnosis of breast cancer helps in predicting the malignant cases. A lump felt during examination roughly give clues as to the size of tumour and its texture. Various common methods used for breast cancer diagnosis are Mammography, Positron Emission Tomography, Biopsy and Magnetic Resonance Imaging. The results obtained from these methods are used to recognize the patterns which are aiming to help the doctors for classifying the malignant and benign cases. There are different data mining techniques, statistical methods and machine learning algorithms that are applied for this purpose. This section includes the review of various technical and review articles on data mining techniques applied in breast cancer diagnosis.

Sarvestani *et al.,* (2010) provided a comparison among the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map (SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which are used to classify WBC and NHBCD data. The performance of such neural network structures was investigated for breast cancer diagnosis problem. PNN and RBF were proved as the best classifiers in the training set. But PNN gave the best classification accuracy when the test set is considered. This work also showed that statistical neural networks can be effectively used for breast cancer diagnosis as by applying several neural network structures a diagnostic system was constructed that performed quite well.

Abdelaal *et al.,* (2010) investigated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases. Here, SVM techniques show the promising results for increasing diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve comparable to values for tree boost and tree forest.

Wei-Pin and Der-Ming (2008) explored that the genetic algorithm model yielded better results than other data mining models for the analysis of the data of breast cancer patients in terms of the overall accuracy of the patient classification, expression and complexity of the classification rule. The artificial neural network, logistic regression, decision tree, and genetic algorithm were used for the comparative studies and the accuracy and positive predictive value of each algorithm were used as the evaluation indicators. WBC database was incorporated for the data analysis followed by the 10-fold cross-validation. The results showed that the genetic algorithm described in the study was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible.

Gandhi *et al.,* (2010) in their paper constructed classification rules using the Particle Swarm Optimization Algorithm for breast cancer datasets. In that study to cope with heavy computational efforts, the problem of the feature subset selection as a pre-processing step was used which learns fuzzy rules bases using GA implementing the Pittsburgh approach. It was used to produce smaller fuzzy rule bases system with higher accuracy. The resulted datasets after feature selection were used for classification using particle swarm optimization algorithm. The rules developed were with rate of accuracy defining the underlying attributes effectively.

Padmavati (2011) performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with the logistic regression. The logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB software. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models had a better predictive power over logistic regression. When comparing MLP and RBF neural network models, it was found that RBF had good predictive capabilities and also time taken by RBF was less than MLP.

Lee *et al.,* (2001) in their study proposed a new classification method based on the hierarchical granulation structure using the rough set theory. Hierarchical granulation structure was adopted to find the classification rules effectively. Classification rules had minimal attributes and the knowledge reduction was accomplished by using the upper and lower approximations of rough sets. One simulation was performed on WBC dataset to show the effectiveness of the proposed method. Simulation result showed

that the proposed classification method generated minimal classification rules and made the analysis of information system easy.

Hassanien and Ali (2004) in their paper presented a rough set method for generating classification rules from a set of observed 360 samples of the WBC data. The attributes were selected, normalized and then the rough set dependency rules were generated directly from the real value attribute vector. Then the rough set reduction technique was applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. They showed that the total number of generated rules was reduced from 472 to 30 rules after applying the proposed simplification algorithm. They also made a comparison between the obtained results of rough sets with the well known ID3 decision tree and concluded rough sets showed higher accuracy and generated more compact rules.

Sawarkar *et al.,* (2006) applied SVM and ANN on the WBC data. The results of SVM and ANN prediction models were found comparatively more accurate than the human being. The 97% high accuracy of these prediction models can be used to take decision to avoid biopsy.

Jamarani *et al.,* (2005) presented an approach for early breast cancer diagnosis by applying combination of ANN and multiwavelet based sub band image decomposition. The proposed approach was tested using the MIAS mammographic databases and images collected from local hospitals. Best performance was achieved by BiGHM2 multiwavelet with areas ranging around 0.96 under ROC curve. Proposed approach could assist the radiologists in mammogram analysis and diagnostic decision making.

Techniques for Breast Cancer Prognosis

Once a patient is diagnosed with breast cancer, the malignant lump should be excised. During this procedure physicians must determine the prognosis of the disease. It is the prediction of the expected flow of the disease. Prognosis is important because the type and intensity of the medications are based on it. The prognosis problem is also called as "analysis of survival or lifetime data". It poses more difficult problem than that of diagnosis since the data is censored. That is, there are only few cases where we have an observed recurrence of the disease. In this case, we can classify patient as recur and we know the time to recur (TTR). On the other hand, we do not observe recurrence in most patients. Due these, there is no real point at which we can consider the patient a non recurrent case. So, data is considered censored since we do not know the time of recurrence. For such patients, all known is only the time of their last check-up. We call this the disease-free survival time (DFS). Prognosis helps in establishing a treatment plan by predicting the outcome of a disease (Pantel, 1998). There are three predictive foci of cancer prognosis: 1) prediction of cancer susceptibility (risk assessment), 2) prediction of cancer recurrence and 3) prediction of cancer survivability. The most widely accepted prognostic factor for breast cancer is the American Joint Commission on Cancer (AJCC) staging system based on the TNM system (T, tumor; N, node; M, metastasis) (Choi *et al.*, 2009) and survival is considered as any incidence of breast cancer where the person is still living from the date of diagnosis. The objective of prognostic predictions is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time. Therefore, breast cancer prognostic problems are mainly in the scope of the widely discussed classification problems. This section consists of the review of different technical and review articles on data mining techniques applied in breast cancer prognosis.

C4.5 is a well known decision tree induction learning technique which has been used by Bellaachia and Guven (2006) along with two other techniques for example Naïve Bayes and Back-Propagated Neural Network. They presented an analysis of prediction of the survivability rate of breast cancer patients using above data mining techniques and used the new version of the SEER Breast Cancer Data. The preprocessed dataset consists of 151,886 records that have all the available 16 fields from the SEER database. They have adopted different approach in the pre-classification process by including three fields: STR (Sarvival Time Recode), VSR (Vital Status Recode), and COD (Cause of Death) and used the Weka toolkit to experiment with these three data mining algorithms. Several experiments were conducted using these algorithms. Achieved prediction performances are comparable to existing techniques. They found out that model generated by C4.5 algorithm for given data has a much better performance than the other two techniques.

Burked *et al.,* (1999) has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku to evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. Values of ROC curve for 5 years was evaluated as 0.909, for 10 years as 0.086 and for15 years as 0.883, these values were used as a measure of accuracy of the prediction model. They also compared 82/300false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy.

Nick (1998) applied ANN classification to Wisconsin Prognostic Breast Cancer and SEER datasets for the analysis of survival. He developed novel encoding as good and poor prognosis of censored data in an ANN architecture to provide a framework for prognostic prediction.

Chi *et al.,* (2007) used the Street's ANN model for Breast Cancer Prognosis on WPBC data and Love data. In their study they used recurrence at five years as a cut point to define the level of risk. Applied models successfully predicted recurrence probability and separated patients with good (>5 yrs) and bad (<5 yrs) prognoses.

Delen *et al.,* (2005) compared ANN, logistic regression and decision tree techniques for breast cancer survival analysis. They used the SEER data's twenty variables in the prediction models. The decision tree with 93.6% accuracy and ANN with 91.2% were found more superior to logistic regression with 89.2% accuracy. Choi *et al.,* (2009) compared the performance of an Artificial Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. Hybrid Network combined both ANN and Bayesian Network. Nine variables of SEER data which were clinically accepted were used as inputs for the networks. The accuracy of ANN (88.8%) and Hybrid Network (87.2%) were very similar and they both outperformed the Bayesian Network. They found proposed Hybrid model can also be useful to take decisions.

Khan *et al.,* (2008) investigated a hybrid scheme based on fuzzy decision trees on SEER data; they performed experiments using different combinations of number of decision tree rules, types of the fuzzy membership functions and inference techniques. They compared performance of each for cancer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification.

Burke *et al.,* (1997) compared the TNM staging system's predictive accuracy with that of ANN for5years survival of the patients. They made the comparison over three different datasets these are SEER data, PCE data and PCE colorectal dataset. They found ANNs more accurate than the TNM staging system in all cases.

MATERIALS AND METHODS

In this paper, we have investigated three data mining techniques: Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. In this paper, we used these algorithms to predict the survivability rate of SEER breast cancer data set. We selected these three classification techniques to find the most suitable one for predicting cancer survivability rate.

The Naïve Bayes technique depends on the famous Bayesian approach following a clear, simple and fast classifier. It has been called 'Naïve' due to the fact that it assumes mutually independent attributes. In practice, it is almost never true but is achievable by preprocessing the data to remove the dependent categories. This method has been used in many areas to represent, utilize, and learn the probabilistic knowledge and significant results have been achieved in machine learning.

The second technique uses artificial neural networks. In this study, a multi-layer network with back-propagation (also known as a multi-layer perceptron) is used.

The third technique is the C4.5 decision-tree generating algorithm. C4.5 is based on the ID3algorithm.It has been shown that the last two techniques have better performance (Zhou and Jiang, 2003; Delen *et al.,* 2005). Therefore we have included them in our analysis.

We have used the Weka toolkit to experiment with these three data mining algorithms. The Weka is an ensemble of tools for data classification, clustering, regression, visualization, and association rules. The toolkit is developed in Java and is open source software issued under the GNU General Public License.

Preprocessing the input data set for a knowledge discovery goal using a data mining approach usually consumes the biggest portion of the effort devoted in the entire work. We have developed a set of tools to extract and cleanup the raw SEER data.

A simple analysis shows that the SEER data has missing information in the fields of Extent of Disease (EOD) and Site Specific Surgery (SSS) fields for almost half of the records. Most of the missing information is in the records, which are gathered prior to 1988. Since we wanted to use all available fields in the SEER database, we removed these records from the test data set. These records have Coding System for EOD coded as '4'. The SSS field usage has changed after 1998. Instead of theregular field, the information is split in five other fields. A mapping scheme from new SSS to old SSS is developed to fill the missing SSS fields. After this step, the records with missing information are removed from the data set.

The EOD field is composed of five fields including the EOD code. These fields (size of tumor, number of nodes, number of positive nodes, and number of primaries) contain missing information coded such as '999', '99' or '9' representing the 'unknown' information. Please note that, the statistics in Table 1 do not contain fields with 'unknown' values. The table also shows the fields used in our analysis.

Table 1: Survivability Attributes

Nominal variable name	**No. of distinct values**
Race	19
Behavior code	2
Grade	5
Histologic type	48
Marital status	6
Primary site code	9
Extension of tumor	23
Site specific surgery code	19
Lymph node involvement	10
Radiation	9
Stage of cancer	5

Numeric variable name	**Mean**	**Std. Dev.**	**Range**
Age	58	13	10-110
Tumor size	20	16	0-200
Number of nodes	15	6.8	0-95
No of positive nodes	1.5	3.7	0-50
Number of primaries	1.25	0.5	1-8

As stated in the previous section, we have adopted a different approach in the pre-classification process. Unlike (Delen *et al.,* 2005), we have included three fields:

STR, VSR, and COD. The STR field ranges from 0to 180 months in the SEER database. The pre classification process is outlined as follows.

// Setting survivability dependent variable for 60months threshold
if STR ≥ 60 months and VSR is alive **then**
the record is pre-classified as "survived"
else if STR < 60 months and COD is breast cancer, **then**
the record is pre-classified as "not survived"
else
Ignore the record
end if

In the above approach, ignored records correspond to those patients that have an STR less than 60 months and are still alive, or those patients that have an STR less than 60 months but the cause of their death is not breast cancer.
Table 2 and Table 3 show the classes of our pre classification process and the approach used in (Delen *et al.*, 2005), respectively.

Table 2: Proposed Survivability Class Instances

Class	No of instances	Percentage
0: not survived	35,148	23.2
1: survived	116,738	76.8
Total	151,886	100

Table 3: Survivability Class Instances based on the Previous Work (Delen *et al.*, 2005)

Class	No of instances	Percentage
0: not survived	162,381	58.3
1: survived	116,282	41.7
Total	278,663	100

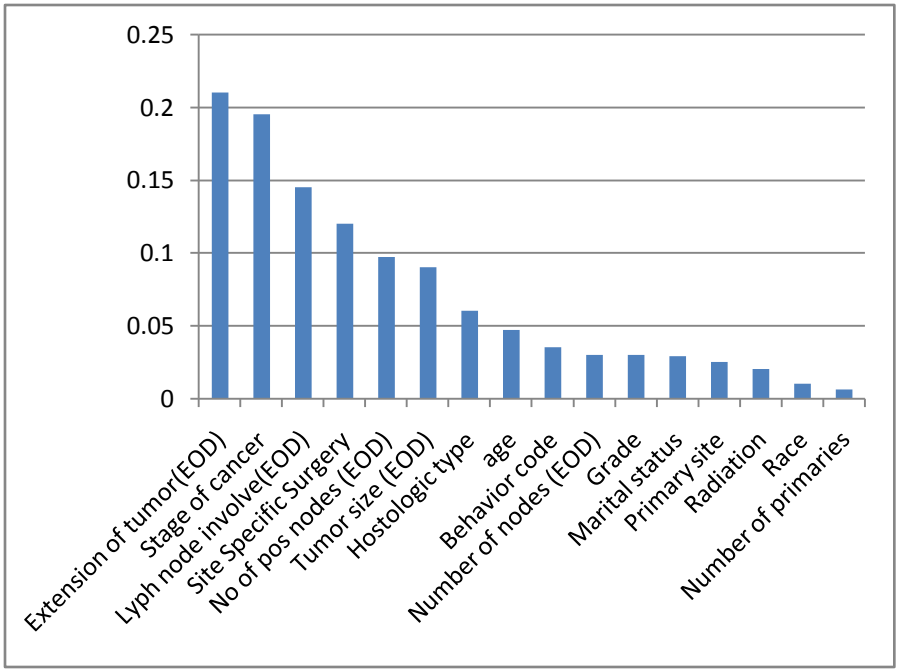

Figure 1: Ranked Survivability Attributes

After the preprocessing step, a common analysis would be determining the effect of the attributes on prediction, or attribute selection. We used the information gain measure to rank the attributes due to the fact that it is a common method and theC4.5 decision tree technique utilizes this measure. Information gain (IG) is measured as the amount of the entropy (H) difference when an attribute contributes the additional information about the class.
The following is the information gain and entropy before and after observing the attribute Xi for the class C:

$$\mathbf{H}(C) = -\sum p(c)\boldsymbol{log}p(c)\ , c \in C \quad (1)$$

$$\mathbf{H}(C|X_i) = -\sum p(x)\sum p(c|x)\boldsymbol{log}p(c|x)\ , x \in X_i, c \in C \quad (2)$$

$$\mathbf{IG_i} = \mathbf{H}(C) - \mathbf{H}(C|X_i) \quad (3)$$

Figure 1 shows the ranked survivability attributes of our data as calculated by the Weka toolkit. It clearly shows that Extension of Tumor has a higher rank than the Tumor Size.

We will use the performance metrics of accuracy, precision and recall comparing the three techniques. In order to have a fair measure of the performance of the classifier; we used a cross validation with 10 folds. Cross-validation, in its most elementary form, consists of dividing the data into k subgroups. Each subgroup is predicted via the classification rule constructed from the remaining (k-1) subgroups, and the estimated error rate is the average error rate from these k subgroups. In this way, the error rate is estimated in an unbiased way.

The final classifier rule is calculated from the entire data set. After running the classifier 10 times with 10folds, we obtain the metrics of precision, accuracy A_i, recall and the Cross Validation Accuracy(CVA) to represent a classifier performance:

$$\mathrm{CVA} = (1/10)\sum \mathrm{A}_i \quad \mathrm{i} = 1, 2, \ldots, 10 \qquad (4)$$

A_i = # records correctly classified / total # records

The Weka toolkit can calculate all these performance metrics after running a specified k-fold cross-validation.

RESULTS AND DISCUSSION

In this study, the accuracy of three data mining techniques is compared. The aim is to have high accuracy, besides high precision and recall metrics. Although such metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics (Witten and Frank, 2005).

Experimental results of our approach as presented in Table 4.

Table 4: Combined Results (our study)

Classification Technique	Accuracy (%)	Class	Precision	Recall
Naïve Bayes	84.5	0	0.70	0.57
		1	0.88	0.93
Artificial Neural Net	86.5	0	0.83	0.52
		1	0.87	0.97
C4.5	86.7	0	0.80	0.56
		1	0.88	0.96

Table 5: Results for C4.5 (dataset as in Table 3)

Classification Technique	Accuracy (%)	Class	Precision	Recall
C4.5	81.3	0	0.86	0.81
		1	0.76	0.81

As can be seen in Table 4, neural net and decision tree have comparable performances. Table 5 shows the experimental results using the pre-classification approach used in [9] and the same dataset used in our approach. The results clearly show that the classification rate (81%) is much lower than the classification rate of our approach (~87%). It may be worth noting that the computation times of the algorithms Naïve Bayes, neural net andC4.5 (on an AMD Athlon 64 4000+ machine) were in the ranges of 1 minute, 12 hours and 1 hour, respectively.

Conclusion

This paper provides a study of different technical and review papers on breast cancer diagnosis and prognosis problems and also has outlined and resolved the issues, algorithms, and techniques for the problem of breast cancer survivability prediction in SEER database.

Data mining techniques offer great promise to uncover patterns hidden in the data that can help the clinicians in decision making. From above study it is observed that the accuracy for the prognosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy.

REFERENCES

Abdelaal MMA, Farouq MW, Sena HA and Salem A (2010). Using data mining for assessing diagnosis of breast cancer. *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (Imcsit)* (IEEE) 11-17.

Bellaachia A and Guven E (2006). Predicting breast cancer survivability using data mining techniques. *Age* **58** 10-110.

Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, Marks JR, Winchester DP and Bostwick DG (1997). Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* **79** 857-862.

Burked MLJLH, Pylkkänenc STL and Joensuua H (1999). Artificial neural networks applied to survival prediction in breast cancer. *Oncology* **57** 281-286.

Chaurasia V and Pal S (2014). Data mining techniques: to predict and resolve breast cancer survivability. *International Journal of Computer Science and Mobile Computing (Ijcsmc)* **3** 10-22.

Chi CL, Street WN and Wolberg WH (2007). Application of artificial neural network-based survival analysis on two breast cancer datasets. *Amia Annual Symposium Proceedings*, 2007, American medical informatics association 130.

Choi JP, Han TH and Park RW (2009). A hybrid bayesian network model for predicting breast cancer prognosis. *Journal of Korean Society of Medical Informatics* **15** 49-57.

Delen D, Walker G and Kadam A (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine* **34** 113-127.

Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR and Ahmad LG (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J health med inform* **4** 124.

Gandhi KR, Karnan M and Kannan S (2010). Classification rule construction using particle swarm optimization algorithm for breast cancer data sets. *International Conference on Signal Acquisition and Processing, 2010, Icsap'10* (IEEE) 233-237.

Hassanien AE and Ali JM (2004). Rough set approach for generation of classification rules of breast cancer data. *Informatica* **15** 23-38.

Jamarani S, Behnam H and Rezairad G (2005). Multiwavelet based neural network for breast cancer diagnosis. *Gvip* **5** 19-21.

Jerez-aragonés JM, Gómez-ruiz JA, Ramos-jiménez G, muñoz-pérez J and Alba-conejo E (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine* **27** 45-63.

Khan MU, Choi JP, Shin H and Kim M (2008). Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. Engineering in medicine and biology society, 2008, Embs 2008, *30th Annual International Conference of the IEEE* (IEEE) 5148-5151.

Kharya S (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *Arxiv Preprint Arxiv* 1205-1923.

Kotsiantis S and Pintelas P (2004). Combining bagging and boosting. *International Journal of Computational Intelligence* **1** 324-333.

Lee CH, Seo SH and Choi SC (2001). Rule discovery using hierarchical classification structure with rough sets. *Ifsa World Congress and 20th Nafips International Conference, 2001, Joint 9th* (IEEE) 447-452.

Liou DM and Chang WP (2015). Applying data mining for the analysis of breast cancer data. *Data Mining in Clinical Medicine* (Springer).
O'malley CD, Le GM, Glaser SL, Shema SJ and West DW (2003). Socioeconomic status and breast carcinoma survival in four racial/ethnic groups. *Cancer* **97** 1303-1311.
Padmavati J (2011). A comparative study on breast cancer prediction using rbf and mlp. *International Journal of Scientific and Engineering Research* **2** 1-5.
Pantel P (1998). Breast cancer diagnosis and prognosis. *University of manitoba*.
Saleema J, Shenoy PD, Venugopal K and Patnaik L (2014). Cancer prognosis prediction model using data mining techniques. *Data Mining and Knowledge Engineering* **6** 21-29.
Sarvestani AS, Safavi A, Parandeh N and Salehi M (2010). Predicting breast cancer survivability using data mining techniques. *2nd International Conference on Software Technology and Engineering (Icste)* (IEEE) v2-227-v2-231.
Sawarkar SD, Ghatol AA and Pande AP (2006). Neural network aided breast cancer detection and diagnosis using support vector machine. *Proceedings of the 7th Wseas International Conference on Neural Networks*, cavtat, croatia.
Society AC (2008). *Cancer Facts & Figures*, the society.
Street WN (1998). A neural network model for prognostic prediction. *ICML*. Citeseer 540-546.
Wei-pin C and Der-ming L (2008). Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data. *Journal of Telemedicine and Telecare* **9**.
Witten IH and Frank E (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, morgan kaufmann.
Zhou ZH and Jiang Y (2003). Medical diagnosis with c4. 5 rule preceded by artificial neural network ensemble. *Ieee Transactions on Information Technology in Biomedicine* **7** 37-42.