An Efficient Method for Predicting the 5-year Survivability of Breast Cancer

Turan Jahanbazi
Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran

Mohammad H. Nadimi
Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran

ABSTRACT
Breast cancer is one of the most sever type of cancers and is the most common cause of death among the female cancer patients. In order to ease the process of decision making and financial arrangements, it is essential to be aware of survivability of patients. In recent years, effective data-mining techniques have been employed to predict the 5-year survivability of cancer patients, showing reasonable accuracy. The efficiency of these models can be improved by making them accessible on smartphones. In order to achieve this, it is essential to reduce the maximum required memory occupied by the prediction models, since a smartphone has a limited available memory. This issue, which is still an open area of research, is the concern of the present study. A hybrid method is enhanced by combining synthetic minority over-sampling technique (SMOTE), information gain attribute evaluation (InfoGainAttributeEval), AdaBoost.M1 algorithm and a decision tree. The more effective attributes are selected using InfoGainAttributeEval and the less effective nodes are removed by decision tree pre-pruning during the tree building. The hybrid method is further simplified by employing the post-pruning technique on the decision tree after its creation. The proposed method was subjected to a 5-year cancer survivability dataset, showing considerable reduction in the maximum required memory while maintaining the accuracy of prediction.

Keywords
Breast cancer, Decision tree, Synthetic minority over-sampling technique, Information gain attribute evaluation, maximum required memory, smartphones, hybrid method.

1. INTRODUCTION
Cancer is a serious problem in health care and one of the major causes of death. Hence, the treatment result needs a careful assessment. Globally, a number of 7.4 million (13% of all deaths) people die from cancer annually, and breast cancer is one of the five life-threatening cancers. For instance, DeSantis et al. [1] reported that about 232340 women in the United States were diagnosed with breast cancer, and 39620 patients died of breast cancer in 2013. Considering the serious impact of this disease, not only patients but also their families suffer. So, being aware of patients' survivability to ease the decision-making process regarding medical treatment and financial preparation is necessary [2]. Various predictive models of machine learning and data mining were employed to make predictions on survivability. Furthermore, with the advancement of technology, there is demand for the use of predictive models on smartphones, but given that the phones have limited memory, and also the volume of data is increasing rapidly, their memory management is very important. Thus, the maximum required memory on smartphones will be decreased by providing a new hybrid approach in this study.

Recently, the breast cancer data sets have been imbalanced, such that the number of survivors is higher than patients who died. Since the standard classifiers are not applicable for imbalanced data, in order to deal with the problem of imbalanced data in this study, SMOTE was employed. This technique was proposed by Chawla et al. [3] which is a famous re-sampling method in data pre-processing and has been applied in several articles, such as Pelayo and Dick [4], Zhao et al. [5], Gu et al. [6]. Using SMOTE technique, the number of samples in minority class can be increased by creating new synthetic samples instead of repeating them, so that the over-fitting problem in learning algorithm is avoided.

According to the limitations of classification, attribute selection technique was proposed in order to overcome the disadvantages of classifier. At this stage, the utilization of effective techniques that can possibly select the important and appropriate attributes is very helpful. By selecting the attribute, classifier workload is decreased and this enhances the classifier's accuracy [7]. In this study, InfoGainAttributeEval was used to select appropriate attributes. Novakovic [7] significantly enhanced the accuracy of decision tree employing this attribute selection technique.

Following the data preprocessing, AdaBoost.M1 technique was employed to improve the prediction performance. This technique is one of the most powerful learning ideas introduced in the last twenty years, which was developed by Yoav Freund and Robert Schapire [8]. AdaBoost.M1 technique has low error rate and has good performance in low noise data sets. It is also utilized as an alternative to Boosting algorithm to combine a set of weak classifiers in form of a model with higher prediction results [9]. Combinations of AdaBoost.M1 with other classifiers have been used in many articles, including Thongkam et al. [9].

After pre-processing of data, decision trees proposed to build the model are adopted. A simplification method of decision tree will be utilized by pre-pruning and post-pruning techniques to handle issues like noise and fragmentation of the tree. Consequently, tree growth is well controlled and classification performance is enhanced [10]. Zhang et al. [10] presented a new method for building decision tree through the use of pruning techniques.

In this study, a new hybrid algorithm is presented as an extension of research [2] in order to decrease the maximum memory needed while maintaining acceptable accuracy. Wang et al. [2] introduced an algorithm to improve the effect of classification for 5-year survivability of patients with breast cancer from voluminous data sets with imbalance property which significantly enhances the effect of classification for imbalanced massive data sets. The current paper refers to the algorithm of Wang et al. [2] as the basic algorithm. This was adjusted and adapted for the purpose of the present study. The

proposed algorithm includes SMOTE, InfoGainAttributeEval, with AdaBoost.M1 and decision tree classifier. In the proposed algorithm considering the memory limitations, decision tree pruning is employed during tree construction and pre-pruning technique was applied for each class in the subset at any stage after selecting the most important attribute and classification of training data set, and then the label of leaf or non-leaf is determined in certain conditions for nodes. Post-pruning technique will be applied on the tree structure after creating the decision tree.

This paper is structured as follows: previous works are introduced in section 2. The proposed method and details of proposed algorithm are presented in section 3. Section 4 shows an example. Section 5 presents the results and discussions. Section 6 includes conclusion.

2. RELATED WORKS

Research on breast cancer has resulted in improved treatment methods in the form of less-invasive predictive medicine. Thus, mortality rate for this cancer has declined in recent years [11]. Using methods that would decrease the maximum memory requirements can also help to utilize predictive systems on smartphones and as a result increase their application performance. The following describes some related works.

Delen et al. in 2005 compared C5.0 decision tree techniques, artificial neural networks and logistic regression to analyze survivability of breast cancer [12]. In this study, it was concluded that the decision tree of C5.0 is the best predictor with an accuracy of 93.6% and artificial neural networks is best next predictor with an accuracy of 91.2% and logistic regression model is the worst predictor with an accuracy of 89.2%.

Bellaachia et al. in 2006 provided an analysis of predicting the survivability of patients with breast cancer employing data mining techniques [13]. C4.5 is a well-known classification technique in decision tree induction which is utilized along with two other techniques of Naïve Bayes and Back-Propagated Neural Network. C4.5 decision tree was determined as the best model with an accuracy of 86.7%.

Liu et al. in 2009 provided predictive models for 5-year survivability of patients with breast cancer employing decision trees based on imbalanced data [14]. In this study, the under-sampling technique and bagging algorithm were employed to deal with the imbalance problem, so that the prediction performance will be enhanced. Finally, the combination of under-sampling technique and decision tree were selected.

Wang et al. in 2013 provided a series of new algorithms to improve the effect of classification for the 5-year survivability of patients with breast cancer from a large database with imbalanced specificity [2]. The results demonstrate that hybrid algorithm of SMOTE + PSO + C5.0 is the best algorithm with an accuracy of 94.26% for the classification of a 5-year survivability of patients with breast cancer between all combinations of algorithms.

Thongkam et al. in 2008 provided the combination of AdaBoost and RF algorithms in order to create a model that predicts breast cancer survivability [9]. The proposed method enhances the accuracy by approximately 88.60% compared with only a classifier and other combined classifiers for predicting survivability of breast cancer.

Mair et al. in 2014 introduced LT-map algorithm and given its hierarchic and nonmetric nature, the LT-map provides compatibility with memory limitations [15]. LT-map method is highly scalable and is dynamically consistent with available memory. Producing LT-map to display a particular route can lead to: (1) reduced memory consumption. (2) In case of lack of memory, tree leaves can easily be pruned so that local and short-term information are eliminated.

According to researches carried out for predicting breast cancer survivability, decision tree is one of the powerful classification algorithms that becomes increasingly more common by growth of data mining and models created using decision trees with high accuracy. By selecting the appropriate attributes and also by simplification techniques of decision tree, a decision tree can be created that needs less maximum memory. In other researches, it has been demonstrated that boosting algorithms are employed for imbalanced data sets and also enhance the prediction performance of decision tree. Some of the articles related to the prediction of survivability are summarized in Table 2.

3. PROPOSED METHOD

In the proposed method, a massive medical data was used with imbalanced property. Using 10-fold cross validation, training and test sets isolated from each other. In each layer, using the technique of SMOTE, the number of samples of training set is resized. Thereafter, the attributes were compared by InfoGainAttributeEval and a subset of appropriate attributes are selected. AdaBoost.M1 algorithm was employed for better performance of the classifier and using the proposed classifier, the decision trees will be created, as in making the decision tree with pre-pruning technique, its low-impact nodes will be detected and eliminated. After making the decision tree, post-pruning technique of the decision tree can be employed for further simplification. Details of the proposed algorithm are presented in Fig. 4.

3.1 10-fold cross validation

To evaluate the performance of the proposed model, a complete data set is divided into two subsets: training and test sets (Fig. 4). The entire data set in this method is divided into 10 equal parts. Nine parts were used as a training set and the model is constructed on the basis of them and the evaluation operation will be performed with the remaining part. This process will be repeated 10 times so that each of the 10 parts will be used only once for evaluation and each time accuracy is calculated for the model built. In this method, the final accuracy evaluation of the classifier will be equal to the average of 10 accuracies calculated.

3.2 Resize data set

Synthetic minority over-sampling technique is carried out on training data set and samples number of minority class is increased by creating new synthetic samples in the original data set. New synthetic samples are created with two specific parameters: over-sampling rate (%) and the number of nearest neighbors (K). In this study, SMOTE technique was applied in order to increase the number of minority class samples and as a result, balancing the data set (Fig. 4).

3.3 Features selection

Diverse attribute ranking and attribute selection techniques have been proposed in the literature of machine learning. The objective of these techniques is to eliminate irrelevant or redundant attributes from the set of attributes. The methods of information gain, gain ration, symmetrical uncertainly, relief-F, one-R and chi-Squared will be utilized to evaluate the

attributes [7]. In this study, the InfoGainAttributeEval is considered for selecting appropriate attributes (Fig. 4).

3.4 AdaBoost.M1

AdaBoost is a well-known ensemble method and, significantly increases the prediction accuracy of the basic learner [8]. This technique is a learning algorithm utilized to generate various classifiers so as to utilize them in building the best classifier [16]. AdaBoost.M1 algorithm was employed in this study to enhance the classifier accuracy (Fig. 4).

3.5 Proposed classifier

At this stage, an induction method will be suggested that is different from the other induction algorithms of decision tree (Fig. 4). Below are the steps to create the proposed algorithm, subsequently each of the steps described above are explained [10].

3.5.1 Definitions

The training data set, the feature set and the class set are identified with D, A, C, respectively.

3.5.2 Choosing the most important feature

A) For each class C_i in C, the probability Pi is calculated. n_i is samples belonging to C_i and ||D|| is the number of samples in D ((1)).

$$P_i = \frac{n_i}{\|D\|} \quad (1)$$

B) With respect to Category C including m outputs, entropy C is defined in the following formula. H(C) measures the information that is needed to classify an attribute in C ((2)).

$$H(C) = \sum_{i=1}^{m} -P_i * \log_2 P_i \quad (2)$$

C) The attribute $A^i \in A$ with values of $\{A_1^i, A_2^i, ... A_v^i\}$ is considered as the root node in the tree. D is divided into subsets of $\{D_1^i, D_2^i, ... D_v^i\}$, D_j^i includes parts whose attribute value for A^i is A_j^i. Entropy of the sub-tree of D_j^i is $H(D_j^i)$. $\|D_j^i\|$ is the number of samples in the subset D_j^i. Information required for this sub-tree is introduced with A^i as its root in (3). Weight of J-th term is the proportionality of samples in D belonging to D_j^i.

$$E(A^i) = \sum_{j=1}^{v} \frac{\|D_j^i\|}{\|D\|} H(D_j^i) \quad (3)$$

D) Finally, the information gain is given by D divided by Ai in (4):

$$Gain(A^i) = H(C) - E(A^i) \quad (4)$$

E) Calculation of Gain(Ai) for all $A^i \in A$

F) The attribute with the highest information gain is shown with $\bar{A}^i$ and is selected as the root node of the tree.

3.5.3 Grouping of training data set

Training set of D for any $\bar{A}^i$ is classified into the subsets of $\{D_1^i, D_2^i, ... D_v^i\}$. Grouping criteria is combination of samples of the training data set which are in a group of D_j^i with same attribute value of $\bar{A}_j^i$.

3.5.4 Proposed pre-pruning technique

A) Calculation of the probability P_i' for each class in a subset D_j^i (j = 1, 2, ..., v). n_i is the number of samples of class C_i in D and n_j is the number of samples of class C_i in D_j^i ((5)).

$$P_i' = \frac{n_j}{n_i} \quad (5)$$

B) Samples from the subset D_j^i that the probability of the class P_i' is less than the threshold will be eliminated from the subset. Threshold is a parameter which prunes the noise samples from the training data set and can ensure that there are adequate samples in the training data set. Consequently, the prepared decision tree provides high accuracy for test data. The pseudo code of the proposed pre-pruning technique is shown in Table 3.

3.5.5 Determining leaf nodes

For value A_j^i of the selected attribute $\bar{A}^i$, leaf node is formed under the following conditions:

A) If the subset D_j^i is empty, then the output A_j^i is specified with "unknown".

B) If the subset D_j^i includes class C_i (homogeneous subset), then the output A_j^i will be C_i.

C) If the subset D_j^i is heterogeneous with classes with equal probability of P_i' and there is no attribute in A that can be utilized to divide more, then the output A_j^i is specified with "unknown".

D) If the output of leaf node is "unknown" for the attribute value of A_j^i, the leaf node label will be replaced with the label of class with the highest probability of P_i' in the subset related to the attribute of $\bar{A}^i$.

3.5.6 Determining non-leaf nodes

Subset $D_j^i \in \{D_1^i, D_2^i, ... D_v^i\}$ which does not have any of the conditions of the leaf node in the previous step is considered as non-leaf node. Further expansion of the tree nodes in the subarea of this node are carried out through the following steps:

A) Removal of the attribute of $\bar{A}^i$ from the attribute set of A

B) Repetition of steps 2-6 for each non-leaf node subset

3.5.7 Post-pruning technique

Once the tree is formed, post-pruning technique is for further simplification in order to deal with the following:

A) If the outputs of class labels are similar for more than one

attribute value of $\overline{A}^{i}$, such as C_i, then all the leaf nodes of the attribute values will be merged as a leaf node labeled with the class C_i.

B) If outputs for all possible attribute values $\overline{A}^{i}$ are equal, such as class C_i, then $\overline{A}^{i}$ with a leaf node is replaced with the class C_i as the output.

4. EXAMPLE

In this section, an example will be given to clarify the proposed method and how it works. The brief list of attributes is shown in

Table 8. The data set includes 10 attributes and 20 samples. Each sample belongs to one of two classes of survival and non-survival (Table 1).

Table 1.The classes of attributes

Classes	*Survival*	*Non-survival*
Number of samples per class	12	8

SSG2000 attribute with the highest information gain is selected by calculating attribute information gain. Thereafter, the data set for all attribute values of SSG2000 were grouped into subsets. Then, the probability of classes in each subset was calculated and compared with the threshold (0.45). If any of the derived probabilities is lower than the threshold, samples related to the class will be eliminated from the subset. According to the description mentioned in the proposed algorithm in this example, samples associated with subsets L, IS and D were removed and leaf nodes related to them were labeled as "unknown". Moreover, samples related to SSG2000 attribute were updated and leaf nodes labels with "unknown" values were substituted with the class label that has the highest probability in the parent node. In Table 3, the probability of classes on the first level of decision tree is compared with the threshold.

Subset R does not have any of the leaf node conditions and is considered as a non-leaf node and after removing SSG2000 attribute, the construction steps of tree in the subset R is repeated (Fig 1). In the end, a decision tree is generated as shown in Fig 2 Post-pruning technique can be employed for further simplification of the decision tree. Fig 3 illustrates the post-pruning of the decision tree made in the previous step. In the first tree level, subsets of L, IS and D except R have leaves with the same label and all the leaf nodes are combined with survival class as a common leaf node labeled. In the second tree level, all leaf nodes have the same label; as a result, NO_SURG attribute was substituted with a single leaf node labeled with non-survival class (Fig 3). At this stage, a proposed model was provided based on the proposed method in order to decrease the maximum memory requirements. Using the proposed method, it is expected that the number of nodes in the tree produced which represents the maximum required memory will be decreased by selecting more effective attributes and abandonment of less effective nodes by determining the appropriate threshold during the manufacturing process of the decision tree.

5. RESULTS AND DISCUSSIONS

5.1 Testing environment

The proposed method was developed in Java environment (Weka software). Many experiments were performed on data sets of 5-year survivability of breast cancer with C5.0, J48, BFTree, REPTree and SimpleCart decision tree algorithms using Weka and Clementine software.

5.2 Evaluation parameters

In this study, the parameter of accuracy was used to evaluate the performance of the prediction of model as well as the parameter of required maximum memory. Accuracy shows the percentage of correctly classified records among the total number of records. Accuracy formula is shown in (6). The maximum required memory shows the number of nodes resulting from the construction of decision tree.

$$Accuracy = (TP + TN) / (TP + FP + TN + FN) \quad (6)$$

Where TP is the number of positive samples which were correctly detected and FN is the number of positive samples which were found to be falsely negative and FP as the number of negative samples which were found to be falsely positive and, TN as the number of negative samples which were correctly diagnosed.

5.3 Data set

To carry out this study, the SEER cancer incidence database from 1973-2012 was applied. SEER program is a part of the monitoring research program at the National Cancer Institute and authoritative source of information on cancer incidence and survivability in the United States [17]. After data cleaning, the dataset includes 146461 records.

5.4 Experiments

Using a random function, 12000 data records were selected between 146461 obtained records and will be assessed given that the number of values utilized in the attributes in the original data set and selected data set are the same. In the proposed method, training and testing data will be separated using 10-fold cross validation and in each layer, the training set will be resized utilizing SMOTE in over-sampling rate of 700%, so that the number of samples in two classes is roughly balanced, and k=5 (5-nearest neighbor). Using InfoGainAttributeEval among 20 attribute of basic research [2], 13 more effective attributes were selected to predict 5-year survivability of breast cancer. Ranking of the attributes is shown in Fig. 5.

Data set is classified in two class "survival" with 89.68% and the class of "non-survival" with 10.32%. Binary target variable is considered as 1 (survival) and 0 (non-survival). Distribution of cancer survivability class is demonstrated in the common model in Table 2.

Table 2. The distribution of cancer survivability class

Classes	*Number of records*	*Percentage*
Survival=1	10761	89.68
Non-survival=0	1239	10.32
Total	12000	100

Using AdaBoost.M1 algorithm and selection of proposed decision tree, the model is built. In the process of making decision tree, the pruning value of the tree is determined by selecting the appropriate threshold. The best possible results were obtained in terms of average accuracy and number of nodes with the threshold value of 0.111. Results were evaluated in Table 6 and

Table 7 in terms of accuracy and maximum amount of memory required.

According to the results presented in Table 6, the average

number of nodes in the proposed and basic methods with the selected attributes were obtained as 472.9 and 872.2, respectively. Reduction of maximum required memory while maintaining accuracy is 46%. The average number of nodes in the proposed method with selected attributes and basic method with tree J48 and selected attributes were obtained as 472.9 and 735.2, respectively. Reduction of the maximum required memory while maintaining accuracy was 36%. After activating the pre-pruning option in the settings of BFTree, the average number of nodes in the proposed method and basic method with BFTree and selected attributes were obtained as 472.9 and 20.6, respectively while average accuracy is also reduced. The average number of nodes in the proposed method with the selected attributes and the basic method with REPTree and selected attributes are 472.9 and 617.7, respectively and reduction of the maximum required memory while maintaining accuracy was 23%. The average number of nodes in the proposed method with selected attributes as well as basic method with tree SimpleCart and selected attributes were obtained as 472.9 and 651.2, respectively. Reduction of the maximum required memory while maintaining accuracy was 27%. By comparing the results, it was found that in the proposed method with selected attributes while maintaining accuracy, less maximum memory is required and it will be compared with the results in

Table 7. The average number of nodes in the proposed method with the selected and basic attributes were obtained as 472.9 and 715.3, respectively and reduction of the maximum required memory while maintaining accuracy was 34%. The average number of nodes in the proposed method with the selected attributes and the basic method with the basic attributes were 472.9 and 956.6, respectively and reduction of the maximum required memory while maintaining accuracy was 51%.

According to the results provided in Table 6 and Table 7, the proposed method with selected attributes achieved better results when compared with other methods in reducing the maximum required memory.

After making the decision tree, tree post-pruning technique was utilized. Average reduction in memory in the proposed method just with pre-pruning technique and proposed method with both pre-pruning and post-pruning techniques are presented in Fig 6.

6. CONCLUSION

Effective models have been introduced to predict 5-year survivability for breast cancer. Nowadays there is a demand to run such models on smartphones. Considering the limited available memory of smartphones, the present research aimed at reducing the maximum required memory of the prediction models. The proposed hybrid method combines synthetic minority over-sampling technique (SMOTE), information gain attribute evaluation (InfoGainAttributeEval), AdaBoost.M1 algorithm and a decision tree. The information gain attribute evaluation allowed selecting the more effective attributes (referred to as selected attributes). Less effective nodes were removed by employing decision tree pre-pruning during tree building. The hybrid method is further simplified by employing the post-pruning technique on the decision tree after its creation. The method was subjected to a 5-year cancer survivability dataset and the number of nodes was compared with those of several benchmark algorithms (C5.0, J48, BFTree, REPTree, SimpleCart) to evaluate the maximum required memories. In all comparisons a reduction in the maximum memory was obtained by the proposed method, while the accuracy of prediction was maintained. Compared to the main benchmark algorithm (from a reference study) the maximum required memory was reduced to 51%. The proposed method could be subjected to other data sets of special cancers such as lung and colon cancer in future studies to evaluate its performance further.

7. APPENDIXES

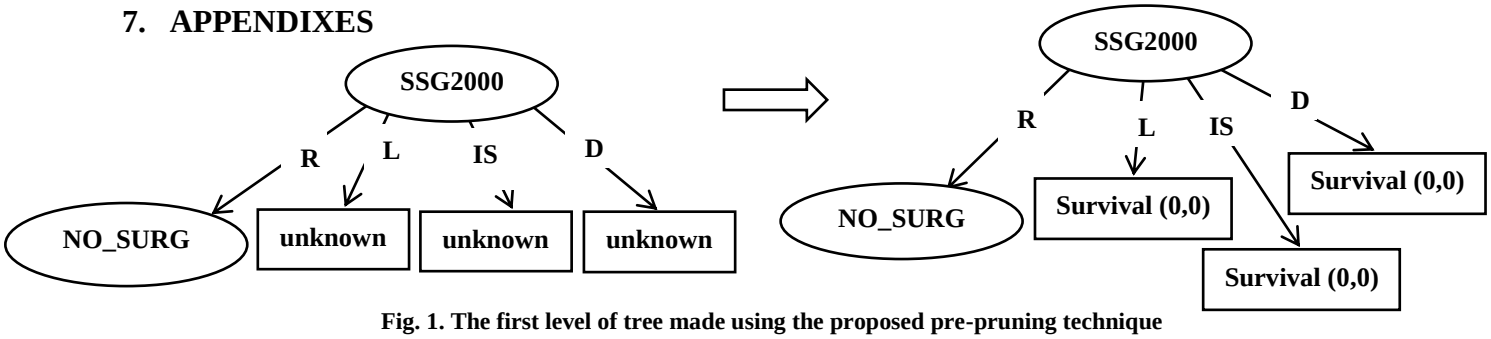

Fig. 1. The first level of tree made using the proposed pre-pruning technique

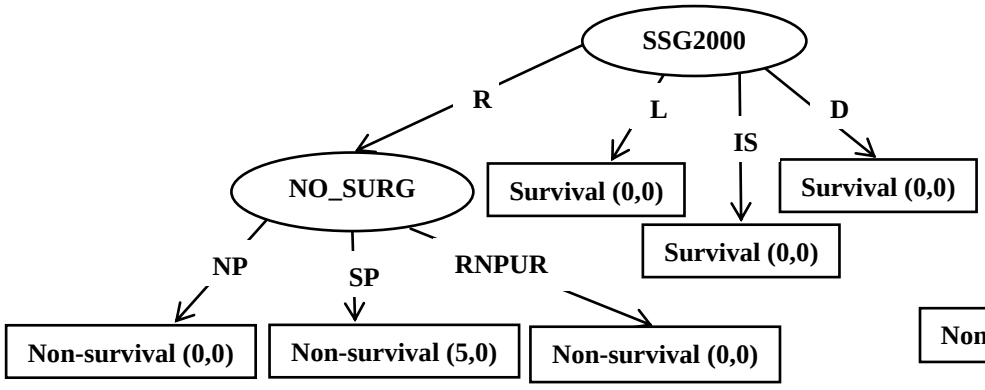

Fig. 2. The tree made using the proposed pre-pruning technique

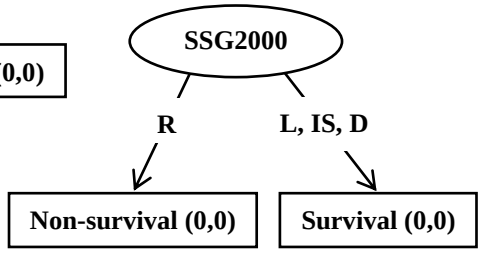

Fig. 3. The tree made using the post-pruning technique

Table 3.The Comparison of the Class Probabilities

Subset R: 7/12 < 0.45 =probability of survival class 5/8 < 0.45 =probability of non-survival class	Subset L: 3/12 < 0.45 =probability of survival class 0/8 < 0.45 =probability of non-survival class
Subset IS: 2/12 < 0.45 =probability of survival class 0/8 < 0.45 =probability of non-survival class	Subset D: 0/12 < 0.45 =probability of survival class 3/8 < 0.45 =probability of non-survival class

Table 4. Previous Researches related to Predict the Survivability of Breast Cancer

Sources	*Class distribution*	*Classifier methods*	*Accuracy performances*
[12]	Survival: 46% Non-survival: 54%	C5.0 DT ANN LR	93.62% 91.21% 89.20%
[13]	Survival: 76.80% Non-survival: 23.20%	C4.5 DT ANN Naïve BN	86.70% 86.50% 84.50%
[14]	Survival: 86.52% Non-survival: 13.48%	C5.0 DT Under-sampling + C5.0 DT Bagging algorithm + C5.0 DT	88.05% (AUC = 0.067) 74.22%(AUC = 0.748) 76.59% (AUC = 0.768)
[2]	Survival:90.68% Non-survival: 9.32%	LR PSO + LR SMOTE + PSO + LR C5.0 PSO + C5.0 SMOTE + PSO + C5.0 1-nn PSO + 1-nn SMOTE + PSO + 1-nn	91.50% 91.30% 74.27% 90.27% 89.14% 94.26% 86.46% 86.66% 87.35%

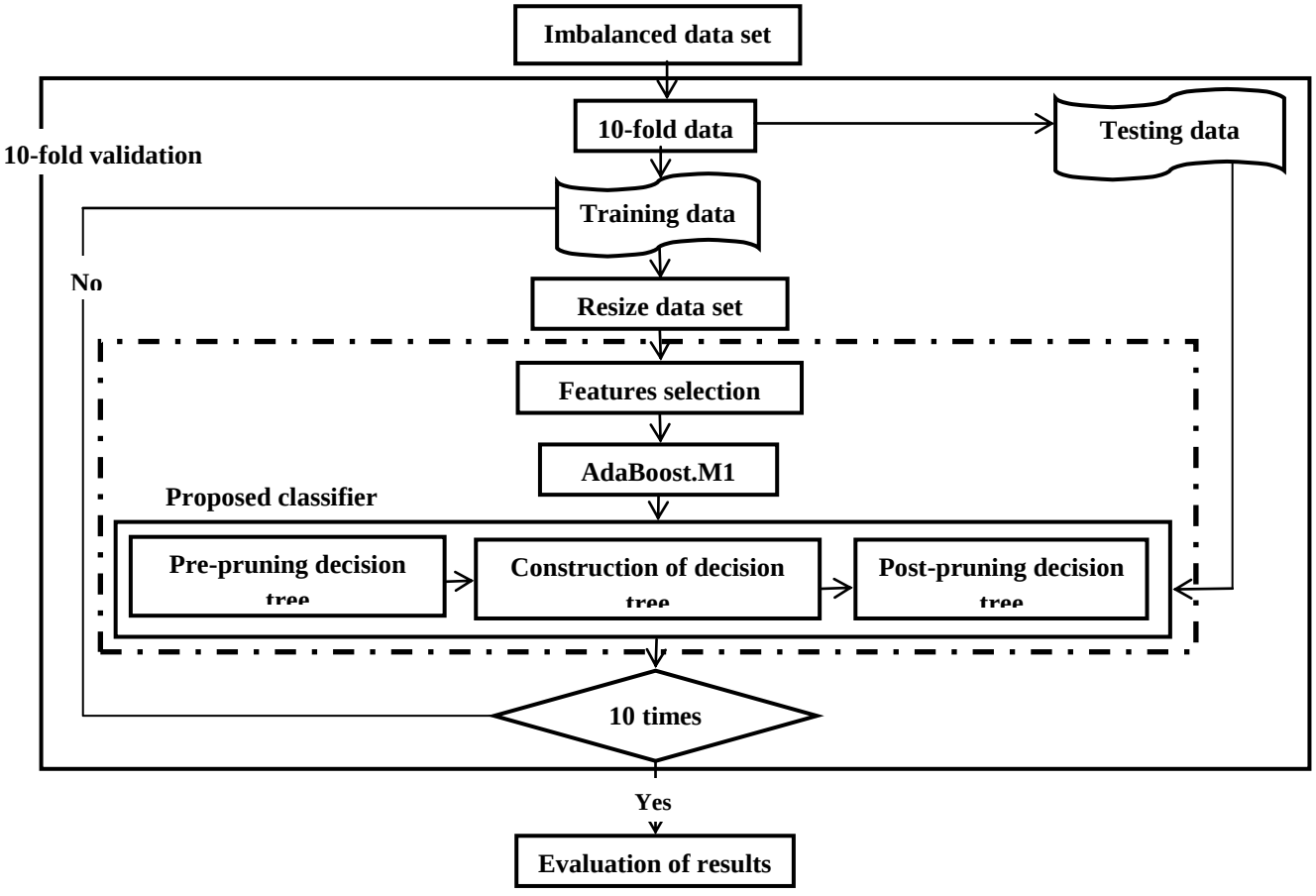

Fig. 4. The proposed model

Table 5. The pseudo code of the proposed pre-pruning_Decision_tree

```
Input: Dataset table
Output: The decision tree without the use of low impact nodes in the process of building tree
Method:
1.  Pre-pruning_Decision_Tree( Node current_node):
2.    survival_rate ← the number of survival class / number of survival class in parent node
3.    non-survival_rate ← the number of  non-survival class / number of non- survival class in parent node
4.    kparam ← The probability of classes defined by user
5.    For    instances in current_node     Do
6.          If  survival_rate < kparam THEN delete instances with survival class in current node
7.          If non survival_rate < kparam THEN delete instances with non-survival class current node
8.          childs ← Split remained instances in current_node to branches
9.          int i=0;
10.         d ←number of child nodes
11.      For  i←1 to d  Do
12.  Pre-pruning_Decision_Tree(childs[i]);
```

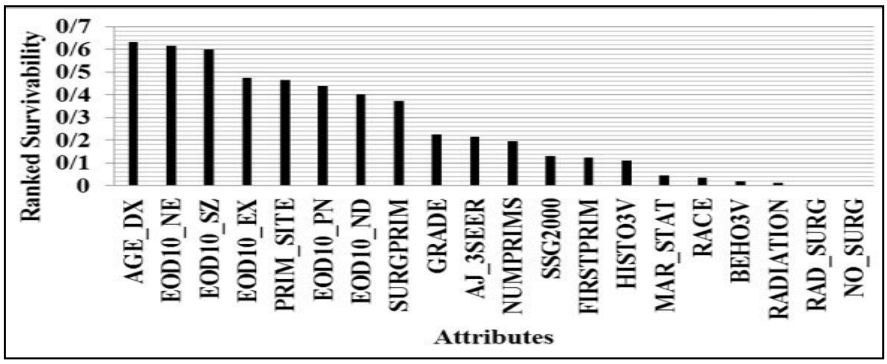

Fig. 5. The ranking of survivability attributes

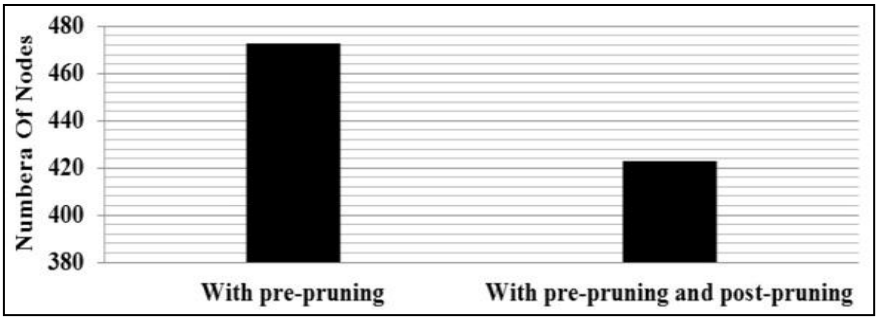

Fig. 6. The comparison of the node numbers

Table 6. The comparison of results with selected attributes

Methods name	Experiments	
	Average number of nodes	Average accuracy
Training data of basic method with selected attributes Testing data of basic method with selected attributes	872.2	91.37% 83.14%
Training data of proposed method with selected attributes (threshold =0.111) **Testing data of proposed method with selected attributes (threshold =0.111)**	**472.9**	**91.39%** **87.07%**
Training data of basic method with tree J48 and selected attributes (threshold =0.111) Testing data of basic method with tree J48 and selected attributes (threshold =0.111)	735.2	95.01% 88.46%
Training data of basic method with tree BFTree and selected attributes Testing data of basic method with tree BFTree and selected attributes	20.6	81.13% 83.42%
Training data of basic method with tree REPTree and selected attributes Testing data of basic method with tree REPTree and selected attributes	617.7	93.35% 87.55%

Training data of basic method with tree SimpleCart and selected attributes Testing data of basic method with tree SimpleCart and selected attributes	651.2	94.93% 87.92%

Table 7. The comparison of results with basic attributes

Methods name	Experiments	
	Average number of nodes	**Average accuracy**
Training data of basic method with basic attributes Testing data of basic method with basic attributes	956.6	92.49% 83.76%
Training data of proposed method with basic attributes (threshold=0.111) Testing data of proposed method with basic attributes (threshold=0.111)	715.3	89.35% 84.23%

Table 8. The brief list of attributes

Race /ethnicityity: Race	Site-specific surgery code: SURGPRIM
Marital status at diagnosis: MAR_STAT	Stage of cancer: SSG2000
Primary Site code: PRIM_SITE	SEER modified AJCC stage 3rd ed: AJ_3SEER
Behavior code ICD-O-3: BEHO3V	First malignant primary indicator: FIRSTPRM
Grade: GRADE	Age at diagnosis: AGE_DX
Extension of disease: EOD10_EX	Tumor size: EOD10_SZ
Lymph node involvement: EOD10_ND	Number of positive nodes: EOD10_PN
Reason of no surgery: NO_SURG	Number of nodes: EOD10_NE
Radiation: RADIATN	Number of primaries: NUMPRIMS
Radiation sequence with surgery: RAD_SURG	Survival status: SURV_STAT
Histology: HISTO3V	

8. ACKNOWLEDGMENTS

The data used in the present research were provided by the National Cancer Institute in the framework of the Surveillance, Epidemiology, and End Results (SEER) Program.

9. REFERENCES

[1] C. DeSantis, J. Ma, L. Bryan, and A. Jemal, "Breast cancer statistics, 2013," CA: a cancer journal for clinicians, vol. 64, pp. 52-62, 2014.

[2] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," Applied Soft Computing, vol. 20, pp. 15-24, 2014.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[4] L. Pelayo and S. Dick, "Applying novel resampling strategies to software defect prediction," in NAFIPS 2007-2007 Annual Meeting of the North American Fuzzy Information Processing Society, 2007, pp. 69-72.

[5] X. M. Zhao, X. Li, L. Chen, and K. Aihara, "Protein classification with imbalanced data," Proteins: Structure, function, and bioinformatics, vol. 70, pp. 1125-1132, 2008.

[6] Q. Gu, Z. Cai, and L. Zhu, "Classification of imbalanced data sets by using the hybrid re-sampling algorithm based on isomap," in International Symposium on Intelligence Computation and Applications, 2009, pp. 287-296.

[7] J. Novakovic, "Using information gain attribute evaluation to classify sonar targets," in 17th Telecommunications forum TELFOR, 2009, pp. 24-26.

[8] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," Journal-Japanese Society For Artificial Intelligence, vol. 14, p. 1612, 1999.

[9] J. Thongkam, G. Xu, and Y. Zhang, "AdaBoost algorithm with random forests for predicting breast cancer survivability," in Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, 2008, pp. 3062-3069.

[10] Y. Liu, D. Zhang, and G. Lu, "Region-based image retrieval with high-level semantics using decision tree learning," Pattern Recognition, vol. 41, pp. 2554-2570, 2008.

[11] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, "Robust predictive model for evaluating breast cancer survivability," Engineering Applications of Artificial Intelligence, vol. 26, pp. 2194-2205, 2013.

[12] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial intelligence in medicine, vol. 34, pp. 113-127, 2005.

[13] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," Age, vol. 58, pp. 10-110, 2006.

[14] L. Ya-Qin, W. Cheng, and Z. Lu, "Decision tree based predictive models for breast cancer survivability on imbalanced data," in Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on, 2009, pp. 1-4.

[15] E. Mair, M. Augustine, B. Jäger, A. Stelzer, C. Brand, D. Burschka, et al., "A biologically inspired navigation concept based on the Landmark-Tree map for efficient long-distance robot navigation," Advanced Robotics, vol. 28, pp. 289-302, 2014.

[16] C. Edeki and S. Pandya, "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability," Mediterranean Journal of Social Sciences, 2012.

[17] SEER (2014) Surveillance, Epidemiology, and End Results (SEER) Program(www.seer.cancer.gov) Research Data (1973-2012). National Cancer Institute,DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2014 based on the November 2013 submission.