Scientific Programming 20 (2012) 29–42
DOI 10.3233/SPR-2012-0335
IOS Press

Lung cancer survival prediction using ensemble data mining on SEER data [1]

Ankit Agrawal *, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi and Alok Choudhary
Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA
E-mails: {ankitag, smi539, ran310, choudhar}@eecs.northwestern.edu, lpolepeddi@u.northwestern.edu

Abstract. We analyze the lung cancer data available from the SEER program with the aim of developing accurate survival prediction models for lung cancer. Carefully designed preprocessing steps resulted in removal/modification/splitting of several attributes, and 2 of the 11 derived attributes were found to have significant predictive power. Several supervised classification methods were used on the preprocessed data along with various data mining optimizations and validations. In our experiments, ensemble voting of five decision tree based classifiers and meta-classifiers was found to result in the best prediction performance in terms of accuracy and area under the ROC curve. We have developed an on-line lung cancer outcome calculator for estimating the risk of mortality after 6 months, 9 months, 1 year, 2 year and 5 years of diagnosis, for which a smaller non-redundant subset of 13 attributes was carefully selected using attribute selection techniques, while trying to retain the predictive power of the original set of attributes. Further, ensemble voting models were also created for predicting conditional survival outcome for lung cancer (estimating risk of mortality after 5 years of diagnosis, given that the patient has already survived for a period of time), and included in the calculator. The on-line lung cancer outcome calculator developed as a result of this study is available at http://info.eecs.northwestern.edu:8080/LungCancerOutcomeCalculator/.

Keywords: Ensemble data mining, lung cancer, predictive modeling, outcome calculator

1. Introduction

Respiratory (lung) cancer is the second most common cancer [26], and the leading cause of cancer-related deaths among men and women in the USA [8]. Survival rate for lung cancer is estimated to be 15% after 5 years of diagnosis [30].

The Surveillance, Epidemiology, and End Results (SEER) Program [25] of the National Cancer Institute is an authoritative repository of cancer statistics in the United States [27]. It is a population-based cancer registry which covers about 26% of the US population across several geographic regions and is the largest publicly available domestic cancer dataset. The data includes patient demographics, cancer type and site, stage, first course of treatment, and follow-up vital status. The SEER program collects cancer data for all invasive and in situ cancers, except basal and squamous cell carcinomas of the skin and in situ carcinomas of the uterine cervix [30]. The 'SEER limited-use data' is available from the SEER website on submitting a SEER limited-use data agreement form. Ries et al. [31] presents an overview study of the cancer data at all sites combined and on selected, frequently occurring cancers from the SEER data. The SEER data attributes can be broadly classified as demographic attributes (e.g., age, gender, location), diagnosis attributes (e.g., primary site, histology, grade, tumor size), treatment attributes (e.g., surgical procedure, radiation therapy), and outcome attributes (e.g., survival time, cause of death), which makes the SEER data ideal for performing outcome analysis studies.

There have been numerous statistical studies using the SEER data like demographic and epidemiological studies of rare cancers [40], assessing susceptibility to secondary cancers that emerge after a primary diagnosis [32], performing survival analysis [30], studying the impact of a certain type of treatment on overall survival [12], studying conditional survival (measuring prognosis of patients who have already survived a period of time after diagnosis) [11,36,37], amongst many

[1] A workshop version of this paper appeared in *Proceedings of the 10th International Workshop on Data Mining in Bioinformatics (BIOKDD 2011)* [3].

* Corresponding author: Ankit Agrawal, Department of Electrical Engineering and Computer Science, Northwestern University, 2145 Sheridan Rd, Evanston, IL 60201, USA. E-mail: ankitag@eecs.northwestern.edu.

others. There also have been scattered applications of data mining using SEER data for breast cancer survival prediction [5,13,15,24] and a few studying lung cancer survival [10,17].

Applying data mining techniques to cancer data is useful to rank and link cancer attributes to the survival outcome. Further, accurate outcome prediction can be extremely useful for doctors and patients to not only estimate survivability, but also aid in decision making to determine the best course of treatment for a patient, based on patient-specific attributes, rather than relying on personal experiences, anecdotes, or population-wide risk assessments. Here we use supervised classification methods to predict survival of lung cancer patients, at the end of 6 months, 9 months, 1 year, 2 years and 5 years of diagnosis. Experiments with several classifiers were conducted to find that many meta classifiers used with decision trees can give impressive results, which can be further improved by combining the resulting prediction probabilities from several classifiers using an ensemble voting scheme. We have developed an on-line lung cancer outcome calculator to estimate the patient-specific risk for mortality due to lung cancer at the end of 6 months, 9 months, 1 year, 2 years and 5 years. Further, to estimate the risk of mortality after 5 years of diagnosis who have already survived a period of time, we also constructed ensemble voting models for predicting conditional survival for lung cancer, and have included them in our calculator.

The rest of the paper is organized as follows: Section 2 summarizes the recent research relevant to the problem, followed by a description of the major classification schemes used in this study in Section 3. The survival prediction system is presented in Section 4, and experiments and results are presented in Section 5. The lung cancer outcome calculator is described in Section 6, and the results of conditional survival prediction are presented in Section 7. Finally, the conclusion and future work are presented in Section 8.

2. Related work

With SEER data being available in the public domain, there is a mature literature on the statistics of SEER data [11,12,30,32,36,37,40], many of them using the SEERStat software provided by SEER itself.

In addition, there also have been a few data mining applications, which has become a very significant component of cancer research and survivability analysis. A number of techniques based on data mining have been proposed for the survivability analysis of various cancers. Zhou and Jiang [41] uses decision trees and artificial neural networks for survivability analysis of breast cancer, diabetes and hepatitis. Lundin et al. [24] uses artificial neural networks on SEER data to predict breast cancer survival. Delen et al. [13] empirically compared three techniques: neural networks, decision trees and logistic regression for the task of predicting 60 months breast cancer survival. They applied these techniques on 2000 version of SEER data. They found that decision trees performed the best with 93.6% accuracy, followed by neural networks. Bellaachia and Guven [5] found that the pre-classification process used by Delen et al. [13] was not accurate in determining the records of the 'not survived' class. The authors of [5] corrected this and investigated Naive Bayes, the back-propagated neural networks, and the C4.5 decision tree algorithm using the data mining tool WEKA. Decision Trees and Neural networks performed the best with 86.7% and 86.5% accuracy, respectively. According to the authors, the difference in results reported by Delen et al. [13] and those obtained by them is due to the facts that they used a newer database (2000 vs. 2002), a different class-distribution (109,659 and 93,273 vs. 35,148 and 116,738) and different toolkits (industrial grade tools vs. WEKA). They also reported the relative importance of various attributes for survival prediction, with extension of tumor, stage of cancer and lymph node involvement (EOD) coming out as the most important attributes.

The authors in [15] studied 5-year survival of follow-up patients in SEER data in 2002 who were diagnosed as breast cancer from 1992–1997. They compared seven methods (artificial neural network, naive Bayes, Bayes network, decision trees with naive Bayes, decision trees (ID3), decision trees (J48) and logistic regression model). The conclusion was that logistic regression (accuracy 85.8%) and decision trees (accuracy 85.6%) were the best ones with high accuracies and high sensitivities. Bellaachia and Guven [5] and Endo et al. [15] also showed that there is a significant imbalance between survived and not-survived classes for the five year survival problem: 80% survived, 20% not-survived. This imbalance in data can potentially affect the accuracy of the developed model. Ya-Qin et al. [39] addressed this problem and used under-sampling to balance the two classes. The conclusion was that the performance of the models is best when the class distribution is approximately equal, i.e., about 50% each (both in training and testing data). Thongkam et al. [34] used adaboost algorithms to pre-

dict breast cancer survivability on data gathered from breast cancer databases of Srinagarined hospital in Thailand.

Modeling survival for lung cancer is not as developed as for breast cancer. Ries and Eisner [29] perform a statistical analysis of the SEER data and computes survival percentage based on gender, race, geographic area, cancer stage, etc., Chen et al. [10] used SEER data containing records of lung cancer patients diagnosed from 1988 through 1998. They examined the following attributes: AJCC (American Joint Committee on Cancer) stage, grade, histological type and gender. For each of the first three attributes, they considered four popular values that are generally used in lung cancer studies. The attribute gender had two values: male and female. This gave them 128 ($4 \times 4 \times 4 \times 2$) possible combinations of values. They applied ensemble clustering on those combinations to get seven clusters and studied survival patterns of those clusters. Fradkin [17] used SEER data for patients diagnosed of cancer of lung or bronchus from the year 1988 through 2001. They studied 8 months survivability of lung cancer. The SEER data had 77 attributes to start with. They preprocessed the data by converting every field to binary, resulting in 98 fields. They removed cases when survival time was unknown or could not be determined. They also removed cases when cause of death was not lung cancer. They removed any attribute with more than 25% missing values. They also removed attributes with more than 95% same value (constant attributes). Any records that were missing more than 25% of the attribute values were also removed. After the preprocessing, 45 attributes were left. They divided the dataset into two parts. The records from 1988–1995 (120,318 records) were used as the training set. The records from 1996–2001 (97,240 records) were used as the test set. They compared penalized logistic regression and SVM for survival prediction of lung cancer, and found that logistic regression resulted in better prediction performance (in terms of ⟨sensitivity, specificity⟩ pair). They also noted that the task of building classification model using SVMs is slow (taking hours to train). Apart from classification modeling, association rule mining has also been applied on lung cancer data [1,2].

3. Classification schemes

We used several classification schemes resulting in identification of top 5 classification schemes, and subsequently used ensemble voting scheme to combine the prediction probabilities from the top 5 models (details presented in experiments and results section). This section presents a brief description of the classifiers and meta-classifiers used in the experiments reported in this paper.

(1) *Support vector machines*. SVMs are based on the Structural Risk Minimization (SRM) principle from statistical learning theory. A detailed description of SVMs and SRM is available in [35]. In their basic form, SVMs attempt to perform classification by constructing hyperplanes in a multidimensional space that separates the instances of different class labels. It supports both classification and regression tasks and can handle multiple continuous and nominal variables. Different types of kernels can be used in SVM models, like linear, polynomial, radial basis function (RBF), and sigmoid. Of these, the RBF kernel is the most recommended and popularly used, since it has finite response across the entire range of the real x-axis.

(2) *Artificial neural networks*. ANNs are networks of interconnected artificial neurons, and are commonly used for non-linear statistical data modeling to model complex relationships between inputs and outputs. The network includes a hidden layer of multiple artificial neurons connected to the inputs and outputs with different edge weights. The internal edge weights are 'learnt' during the training process using techniques like back propagation. Several good descriptions of neural networks are available [6,16].

(3) *J48 decision tree*. In a decision tree classifier, the internal nodes denote the different attributes whose values would be used to decide on the classification path, and the branches denote the split depending on the attribute values, while the leaf nodes denote the final value (classification) of the dependent variable. While constructing the decision tree, the J48 algorithm [28] identifies the attribute that must be used to split the tree further based on the notion of information gain/gini impurity.

(4) *Random forest*. The Random Forest [7] classifier consists of multiple decision trees. The final class of an instance in a Random Forest is assigned by outputting the class that is the mode of the outputs of individual trees, which can produce robust and accurate classification, and has the ability to handle a very large number of in-

put variables. It is relatively robust against overfitting and can handle datasets with highly imbalance class distributions.

(5) *LogitBoost*. Boosting is a technique that can dramatically improve the performance of several classification techniques by applying them repeatedly to re-weighted versions of the input data, and taking a weighted majority vote of the sequence of classifiers thereby produced. In [19], the authors explain the theoretical connection between boosting and additive models. The LogitBoost algorithm is an implementation of additive logistic regression which performs classification using a regression scheme as the base learner, and can handle multi-class problems.

(6) *Decision stump*. A decision stump [38] is a weak tree-based machine learning model consisting of a single-level decision tree with a categorical or numeric class label. Decision stumps are usually used in ensemble machine learning techniques.

(7) *Random subspace*. The Random Subspace classifier [22] constructs a decision tree based classifier consisting of multiple trees, which are constructed systematically by pseudo-randomly selecting subsets of features, trying to achieve a balance between overfitting and achieving maximum accuracy. It maintains highest accuracy on training data and improves on generalization accuracy as it grows in complexity.

(8) *Reduced error pruning tree*. Commonly known as REPTree [38], it is an implementation of a fast decision tree learner, which builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning.

(9) *Alternating decision tree*. ADTree [18] is a decision tree classifier which supports only binary classification. It consists of two types of nodes: decision nodes (specifying a predicate condition, like 'age' $>$ 45) and prediction nodes (containing a single real-value number). An instance is classified by following all paths for which all decision nodes are true and summing the values of any prediction nodes that are traversed. This is different from the J48 decision tree algorithm in which an instance follows only one path through the tree.

(10) *Voting*. Voting is a popular ensemble technique for combining multiple classifiers. It has been shown that ensemble classifiers using voting may outperform the individual classifiers in certain cases [23]. Here we combine multiple classifiers by using the average of probabilities generated by each classifier. The base classifiers used for the voting scheme were LogitBoost (with DecisionStump), RandomSubSpace (with REPTree), J48 decision tree, Random Forests and ADTree.

4. Survival prediction system

Understanding and cleaning data is one of the most important steps for effective data mining. Appropriate preprocessing, therefore, becomes extremely crucial in any kind of predictive modeling, including that of cancer survival, as also widely accepted by numerous other related studies. The proposed respiratory cancer survival prediction system consists of four stages:

(1) *SEER-related preprocessing*. This is the first stage preprocessing designed according to the way SEER program records, codes, and releases the data. There are three principle steps in this stage:

  (a) Convert apparently numeric attributes to nominal, e.g., marital status, sex.
  (b) Split appropriate numeric attributes into numeric and nominal parts, e.g., tumor size. ('CS TUMOR SIZE' gives the exact size of the tumor in mm, if it is known. But in some cases, the doctor notes may say 'less than 2 cm', in which case the coder assigns a value of 992 to the field, which, if used as a numeric value, would correspond to 992 mm, which is incorrect.)
  (c) Construct survival time in months (numeric) from SEER format of YYMM.

(2) *Problem-specific preprocessing*. This is the second stage preprocessing which is specific to the problem of survival prediction. The following are the steps in this stage:

  (a) Select data records for a particular time period of interest.
  (b) Filter the attributes that vary too much or too little, since they do not have significant predictive power.
  (c) For cancer-specific survival analysis, remove records where the patient died because of something other than the cancer in study.
  (d) For cancer-specific survival analysis, remove attributes apart from survival time, which directly or indirectly specify the outcome, e.g., cause of death, whether the patient is still alive.

(e) For binary class prediction, derive appropriate binary attributes for survival, e.g., 5-year survival.

(3) *Predictive modeling*. This is where supervised classification methods are employed to construct predictive models for cancer-specific survival, on the preprocessed data. The two straightforward steps of this stage are:

(a) Split the preprocessed data in training and testing sets (or use cross validation).
(b) Construct a model on the training data using supervised classifiers, e.g., naive Bayes, logistic regression, decision trees, etc., including an ensemble of different classifiers.

(4) *Evaluation*. In this stage, the predictive model is evaluated on the testing data:

(a) Compare the survival predictions from the predictive model on unseen data (testing set) against known survival.
(b) Calculate performance metrics like accuracy (percentage of predictions that are correct), precision (percentage of positive predictions that are correct), recall/sensitivity (percentage of positive labeled records that were predicted as positive), specificity (percentage of negatively labeled records that were predicted as negative), area under the ROC curve (a measure of discriminative power of the model), etc.

Figure 1 presents the block diagram of the survival prediction system with carefully designed preprocessing steps followed by modeling and evaluation.

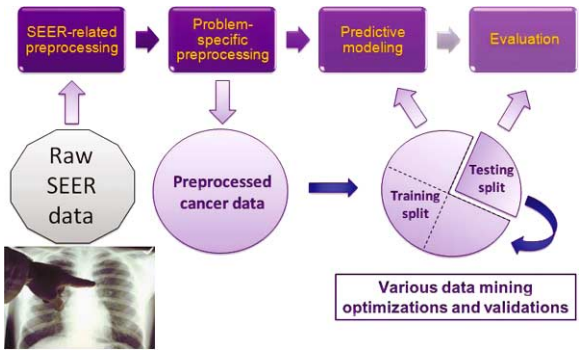

Fig. 1. Block-diagram of the survival prediction system. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/SPR-2012-0335.)

5. Experiments and results

In this study, we used the data in the SEER November 2008 Limited-Use Data files [25] (released in April 2009) from nine SEER registries (Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound and Utah). The SEER data used in this study had a follow-up cutoff date of December 31, 2006, i.e., the patients were diagnosed and followed-up upto this date. In our experiments, we used the WEKA toolkit for data mining [21]. All the models employed were run using the default parameters unless otherwise stated.

The SEER-related preprocessing resulted in modification and splitting of several attributes, many of which were found to have significant predictive power. In particular, 2 out of 11 newly created (derived) attributes were within the top 13 attributes that were selected to be used in the lung cancer outcome calculator. These were (a) the count of regional lymph nodes that were removed and examined by the pathologist; and (b) the count of malignant/in-situ tumors. These attributes were derived from 'Regional Nodes Examined' and 'Sequence Number-Central' respectively from raw SEER data, both of which had nominal values encoded within the same attribute, with the latter also encoding non-malignant tumors.

Subsequently, we selected the data for the patients diagnosed between 1998 and 2001. This choice was made because of the following: since we wanted to do a survival prediction for upto 5-years, and the follow-up cutoff date for the SEER data in study was December 31, 2006, we used the data for cancer patients with year of diagnosis as 2001 or before. Moreover, since several important attributes were introduced to the SEER data in 1998 (like RX Summ–Surg Site 98-02, RX Summ–Scope Reg 98-02, RX Summ–Surg Oth 98-02, Summary stage 2000 (1998+)), we further restricted the patient data with year of diagnosis as 1998 or after. Thus, we selected the data of all cases of respiratory cancer patients in the above mentioned nine SEER registries diagnosed between 1998 and 2001. There were a total of 70,132 such instances. After removing the attributes which varied too much or too little (and hence did not have significant predictive power), we were left with a total of 68 attributes. We further removed all instances where the patient died because of something other than respiratory cancer, reducing the number of instances to 57,254. After removing cause of death and related attributes, we were left with 64 attributes (including survival time in months). Since the survival rate of res-

Table 1
Class distribution

Fraction	Survival classes				
	6-month	9-month	1-year	2-year	5-year
Not-survived	38.85%	49.12%	57.04%	72.79%	83.23%
Survived	61.15%	50.88%	42.96%	27.21%	16.77%

piratory cancer is extremely low, we derived binary attributes for 6-month, 9-month, 1-year, 2-year and 5-year survival. The number of attributes were thus reduced from 118 in the initial dataset to 64, i.e., 63 predictor attributes and 1 outcome attribute (which can be 6-month/9-month/1-year/2-year/5-year survival).

Table 1 presents the distribution of not-survived and survived patients at the end of 6 months, 9 months, 1 year, 2 years and 5 years of diagnosis. It clearly shows that the distribution can be quite lopsided for some classes.

For classification, we built predictive models using more than 30 different classification schemes, and of those which completed execution in reasonable time, the top 5 were selected:

(1) J48 decision tree.
(2) Random forest.
(3) LogitBoost (with Decision Stump as the underlying classifier).
(4) Random subspace (with REPTree as the underlying classifier).
(5) Alternating decision tree.

Because these 5 classification schemes gave good performance, we also decided to use the ensemble voting technique for combining the results from these classifiers. Voting can combine the probabilities generated by each classifier in different ways, like average, product, majority, maximum, minimum and median. After some initial experiments with the different ways of combining the probabilities (which gave similar results), we chose to calculate the resulting probability by taking the average of the probabilities generated by each classifier.

We conducted experiments with the above mentioned 6 ($= 5 + 1$) classification schemes, on each of the 5 datasets (with class variable as 6-month, 9-month, 1-year, 2-year and 5-year survival). 10-fold cross-validation was used for training and testing, and 10 runs of each $\langle$dataset, algorithm$\rangle$ were conducted (with different cross-validation folds) for statistical analysis of the performance comparison. Thus, there were a total of $5 \times 6 \times 10 \times 10 = 3000$ runs. Next, we present the results.

The ZeroR classifier is commonly used as a baseline classifier to measure the improvement in prediction performance due to modeling over simply going by statistical majority, i.e., always predicting the majority class. Figure 2 presents the overall prediction accuracy of the above-mentioned 6 classification schemes, along with ZeroR classifier, on each of the five datasets. Since, accuracy results can be often misleading due to imbalanced classes, the area under the ROC curve (AUC) is considered a better metric to measure the ability of the model to discriminate between the different class values. Figure 3 presents the area under the ROC curve (AUC) for the same. It is worth noting here that AUC for ZeroR classifier is technically not defined, as ZeroR provides no discrimination. But it can be thought of being equivalent to having an ROC curve with the line of no discrimination (from bottom-left to top-right corners of the ROC curve), and thus the baseline AUC for ZeroR is usually taken to be 50%. For completeness, Figs 2 and 3 also present the classification results obtained by using support vector machines (with RBF kernel) [9,14] and neural networks, although their results were found to be less accurate and inconsistent as compared to other classifiers. Moreover, the execution time for constructing SVM and neural network models was significantly larger as compared to other models. Therefore, instead of multiple runs of cross-validation, a single run of training-testing split (training on 66% data, testing on remaining) was conducted to measure the accuracy of these models. The SVM models required around 15 CPU hours for construction (slow training of SVM models is also acknowledged in [17]), and the neural network model construction did not complete after more than 1000 CPU hours of execution time. The results for neural network reported in this paper were obtained on the dataset with a reduced attribute set (from 63 to 13 attributes, used for the tool as described later), which, for the ensemble voting scheme was found to give similar prediction accuracy as with using all 63 attributes. Neural network modeling on this smaller dataset took about 80 CPU hours. Since for this data, better prediction quality was obtained

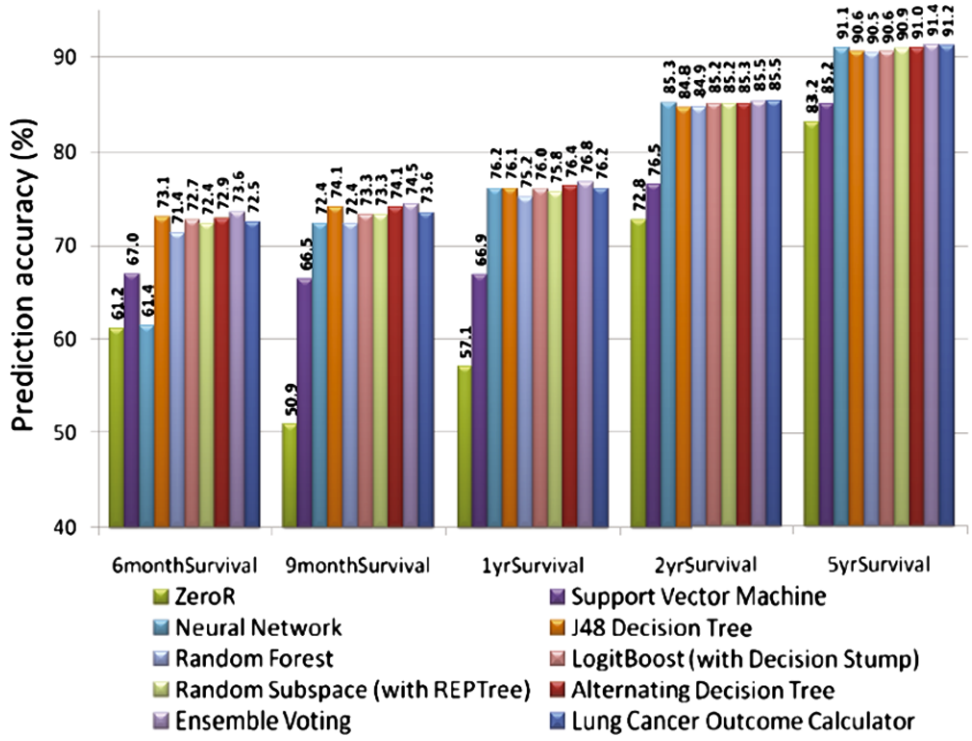

Fig. 2. Prediction accuracy comparison amongst different classification techniques. The lung cancer outcome calculator uses ensemble voting scheme using just 13 predictor attributes. The bars in the figure from left to right are in the same order as the legend (when read row-wise). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/SPR-2012-0335.)

by other models that could be constructed faster than SVM and neural network models, these models were not investigated further. Ability to construct the models in reasonable time is crucial to enable regular model updates by incorporating new data as and when it becomes available.

From Figs 2 and 3, it is clear that ensemble voting classification scheme gives the best prediction performance, both in terms of prediction accuracy and AUC, which was also found to be (statistically) significantly better than the J48 decision tree as the base learner, at 5% significance level. Some important observations from these figures are as follows. For 5-year survival prediction, the baseline classifier (ZeroR) classifies all records as 'not survived' (majority class), achieving a prediction accuracy of 83.2% because of the imbalanced class distribution, which seems quite impressive, but is clearly uninformative and not useful in practice. Model-driven prediction for the same 5-year prediction boosts the prediction accuracy up to 91.4%, which means an effective reduction of error rate from 16.8% to 8.6%, thereby reducing the error rate almost by a factor of 2. Apart from prediction accuracy, an excellent discriminative power (discrimination between death and survival) of 5-year survival prediction model was also obtained with a high AUC of 0.94.

In general, it is not straightforward to compare prediction results on different datasets with different class distributions. The work in [10] had applied ensemble clustering to study survival patterns of obtained clusters, but no test results were reported. Moreover, the study used only 4 attributes with popular values of those attributes. The predictive models used in the current study are more general using all available attributes. The work in [17] studied 8 months survivability of lung cancer using variations of logistic regression and SVM techniques, and reported results in terms of sensitivity and specificity. Again, their results are not directly comparable to ours, since both the dataset and the target class are different. More specifically, we use a more recent release of the SEER database with newer attributes, and a different time period of the diagnosed cases, as compared to [17]. They report sensitivity and specificity as measures of the quality of prediction. Some of the best ⟨sensitivity, specificity⟩ combinations in their experiments were: ⟨74.62, 70.57⟩, ⟨74.84, 68.26⟩, ⟨75.44, 63.27⟩. We had conducted experiments for 9-month survival, and the ⟨sensitivity,

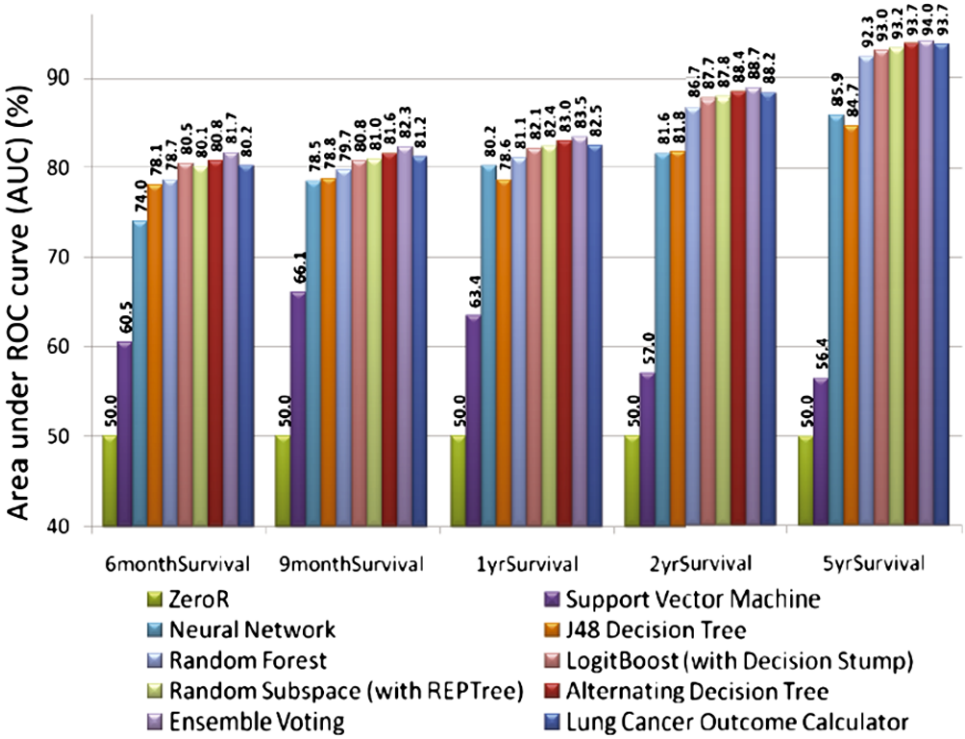

Fig. 3. Prediction performance comparison in terms of area under the ROC curve (AUC). The lung cancer outcome calculator uses ensemble voting scheme using just 13 predictor attributes. The bars in the figure from left to right are in the same order as the legend (when read row-wise). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/SPR-2012-0335.)

specificity⟩ combination with ensemble voting scheme was ⟨78.90, 70.15⟩.

6. On-line lung cancer outcome calculator

Further, for the purpose of building an on-line tool for lung cancer outcome prediction, we used correlation-based feature subset selection technique [20] to identify a smaller non-redundant subset of attributes which were highly correlated with the outcome variable while having low inter-correlation amongst themselves. The goal here was to make the tool convenient to use by reducing the number of attributes, while trying to retain the predictive power of the original set of attributes in the preprocessed data. The attribute subsets obtained for the five different outcome variables were combined, and clearly redundant attributes were manually removed. SEER-specific attributes were further removed to make the calculator more easily applicable to new patients. The calculator uses the resulting 13 input variables as shown in Fig. 4 (with relative predictive power) to estimate lung-cancer-specific mortality risk using the ensemble voting scheme. Following is a brief description of these attributes. The original SEER names of the attributes are also mentioned wherever significantly different from the names used in the calculator:

(1) *Age at diagnosis*. Numeric age of the patient at the time of diagnosis for lung cancer.
(2) *Birth place*. The place of birth of the patient. There are 198 options available to select for this attribute (based on the values observed in the SEER database).
(3) *Cancer grade*. A descriptor of how the cancer cells appear and how fast they may grow and spread. Available options are – well-differentiated, moderately differentiated, poorly differentiated, undifferentiated, and undetermined.
(4) *Diagnostic confirmation*. The best method used to confirm the presence of lung cancer. Available options are – positive histology, positive cytology, positive microscopic confirmation (method unspecified), positive laboratory test/marker study, direct visualization, radiology, other clinical diagnosis, and unknown if microscopically confirmed.

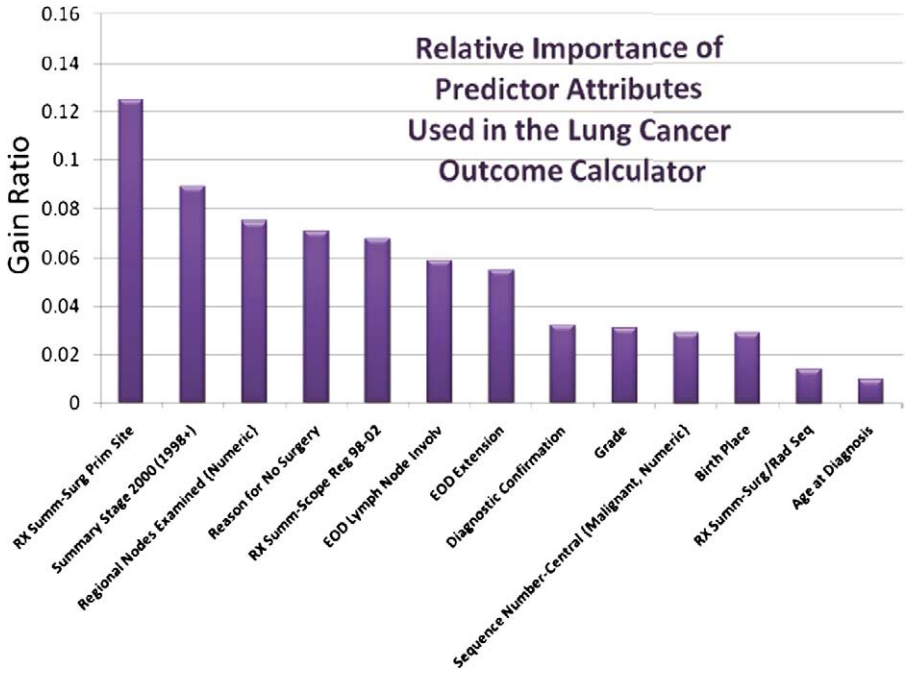

Fig. 4. The attributes used in the lung cancer outcome calculator along with their relative predictive power. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/SPR-2012-0335.)

(5) *Farthest extension of tumor*. The farthest documented extension of tumor away from the lung, either by contiguous extension (regional growth) or distant metastases (cancer spreading to other organs far from primary site through bloodstream or lymphatic system). There are 20 options available to select for this attribute. The original SEER name for this attribute is 'EOD extension'.

(6) *Lymph node involvement*. The highest specific lymph node chain that is involved by the tumor. Cancer cells can spread to lymph nodes near the lung, which are part of the lymphatic system (the system that produces, stores, and carries the infection-fighting-cells. This can often lead to metastases. There are 8 options available for this attribute. The original SEER name for this attribute is 'EOD Lymph Node Involv'.

(7) *Type of surgery performed*. The surgical procedure that removes and/or destroys cancerous tissue of the lung, performed as part of the initial work-up or first course of therapy. There are 25 options available for this attribute, like cyrosurgery, fulguration, wedge resection, laser excision, pneumonectomy, etc. The original SEER name for this attribute is 'RX Summ–Surg Prim Site'.

(8) *Reason for no surgery*. The reason why surgery was not performed (if not). Available options are – surgery performed, surgery not recommended, contraindicated due to other conditions, unknown reason, patient or patient's guardian refused, recommended but unknown if done, and unknown if surgery performed.

(9) *Order of surgery and radiation therapy*. The order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation. Available options are – no radiation and/or surgery, radiation before surgery, radiation after surgery, radiation both before and after surgery, intraoperative radiation therapy, intraoperative radiation with other radiation given before/after surgery, and sequence unknown but both surgery and radiation were given. The original SEER name for this attribute is 'RX Summ–Surg/Rad Seq'.

(10) *Scope of regional lymph node surgery*. It describes the removal, biopsy or aspiration of regional lymph node(s) at the time of surgery of the primary site or during a separate surgical event. There are 8 options available for this attribute. The original SEER name for this attribute is 'RX Summ–Scope Reg 98-02'.

(11) *Cancer stage*. A descriptor of the extent the cancer has spread, taking into account the size of the tumor, depth of penetration, metastasis, etc. Available options are – *in situ* (noninvasive neoplasm), localized (invasive neoplasm confined to the lung), regional (extended neoplasm), distant (spread neoplasm) and unstaged/unknown. The original SEER name for this attribute is 'Summary Stage 2000 (1998+)'.

(12) *Number of malignant tumors in the past*. An integer denoting the number of malignant tumors in the patient's lifetime so far. This attribute is derived from the SEER attribute 'Sequence Number-Central', which encodes both numeric and categorical values for both malignant and benign tumors within a single attribute. As part of the preprocessing, the original SEER attribute was split into numeric and nominal parts, and the numeric part was further split into 2 attributes representing number of malignant and benign tumors respectively.

(13) *Total regional lymph nodes examined*. An integer denoting the total number of regional lymph nodes that were removed and examined by the pathologist. This attributed was derived by extracting the numeric part of the SEER attribute 'Regional Nodes Examined'.

Prediction performance with just 13 attributes used in the calculator is also presented in Figs 2 and 3, which shows only marginal decrease in prediction performance as compared to using all 63 variables. A careful selection of attributes for the calculator has therefore resulted in a decrease in the number of attributes from 63 to 13, while incurring only a marginal cost on prediction accuracy (Prediction accuracy $= 91.2\%$ for 5-year survival prediction with 13 attributes, as compared to 91.4% with 63 attributes; AUC $= 0.937$ for 5-year survival prediction with 13 attributes, as compared to 0.94 with 63 attributes). It seems that these 13 attributes were able to reasonably encode the information available in the previously used 63 attributes, which prevents any significant drop in accuracy. It is also interesting that the birth place shows up as a significant attribute in the set of 13 attributes. Figure 5 shows a screenshot of the lung cancer outcome calculator. A preliminary version of the calculator was earlier reported in a poster abstract [4].

Further, we performed a consistency analysis on the predictions generated by the model used in the calculator. Since we have survivability predictions for increasing time durations, it would be interesting to see if the predictions are consistent. For example, if the 6-month survival model predicts that the patient will not survive but any other model (9-month, 1-year, 2-year or 5-year) predicts that the patient will survive, the predictions are inconsistent. Essentially, if the survivability model for time T predicts that the patient will survive (will not survive), then any other model predicting survivability for time less than (greater than) T should predict that the patient will survive (will not survive). 10-fold cross validation predictions were checked for consistency and out of 57,254 instances, only 168 instances (0.29%) were found to be inconsistent.

7. Conditional survival prediction

Survival prediction from time of diagnosis is important, but for patients who have already survived a period of time since diagnosis, conditional survival is a more clinically relevant and useful measure, as it tries to incorporate the changes in risk over time. For this reason, we further used the ensemble voting scheme to create conditional survival prediction models on the data used in the lung cancer outcome calculator. Since 5-year survival rate is the most commonly used measure to estimate the prognosis of cancer, we built models to estimate patient-specific risk of mortality after 5 years of diagnosis of lung cancer, given that the patient has already survived for 6, 9, 12, 18 and 24 months, i.e., five new models were constructed.

For constructing a model for estimating mortality risk after 5 years of diagnosis of patients already survived for time T, only those patients were included in the modeling data, that survived at least time T. Note that this is equivalent to taking the data used in the calculator to build 5-year survival prediction model, and removing the instances where the survival time was less than T. Thus, 5 new datasets were created for 5 different values of T (6, 9, 12, 18 and 24 months), and the same ensemble voting scheme was used to build 5 new models. Table 2 presents the class distribution for the conditional survival datasets.

Figures 6 and 7 present the overall prediction accuracy and AUC for the 5 new models constructed for conditional survival prediction. Again, 10-fold cross-validation was used to obtain model accuracy and AUC. The figures confirm the significant improvement in prediction accuracy and AUC over the baseline values. The resulting models have also been incorporated in the on-line lung cancer outcome calculator.

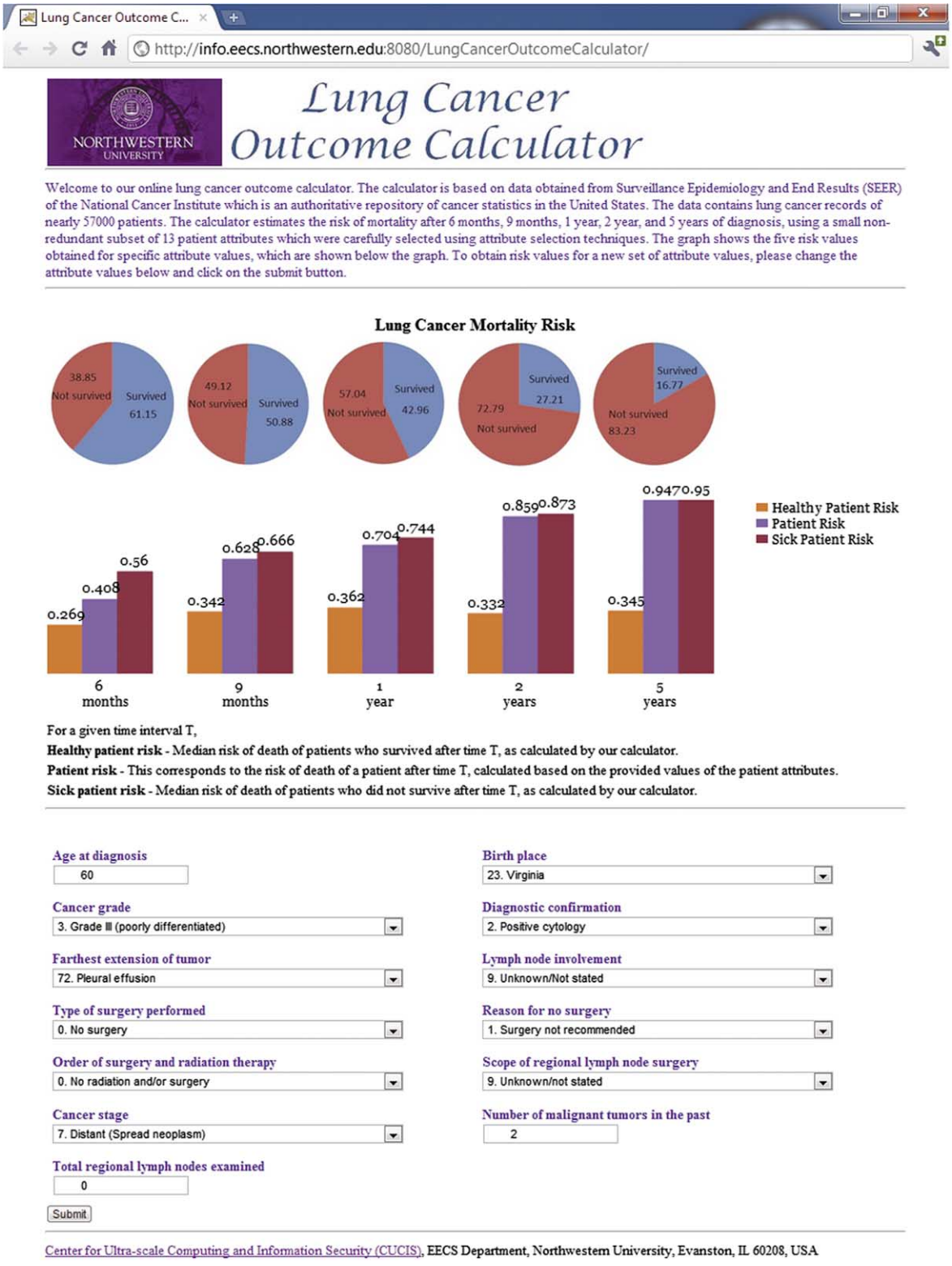

Fig. 5. Screenshot of the lung cancer outcome calculator. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/SPR-2012-0335.)

Table 2
Class distribution for conditional survival (5-year survival given that the patient has already survived for time T)

Fraction	Time T for which patient has already survived				
	6 months	9 months	12 months	18 months	24 months
Not-survived	72.58%	67.04%	60.97%	48.89%	38.38%
Survived	27.42%	32.96%	39.03%	51.11%	61.62%

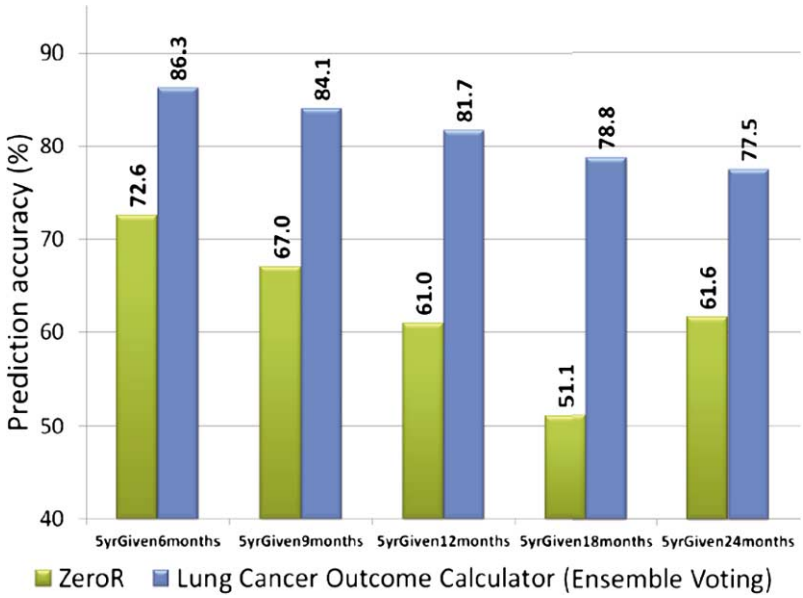

Fig. 6. Prediction accuracy comparison for different conditional survival prediction models. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/SPR-2012-0335.)

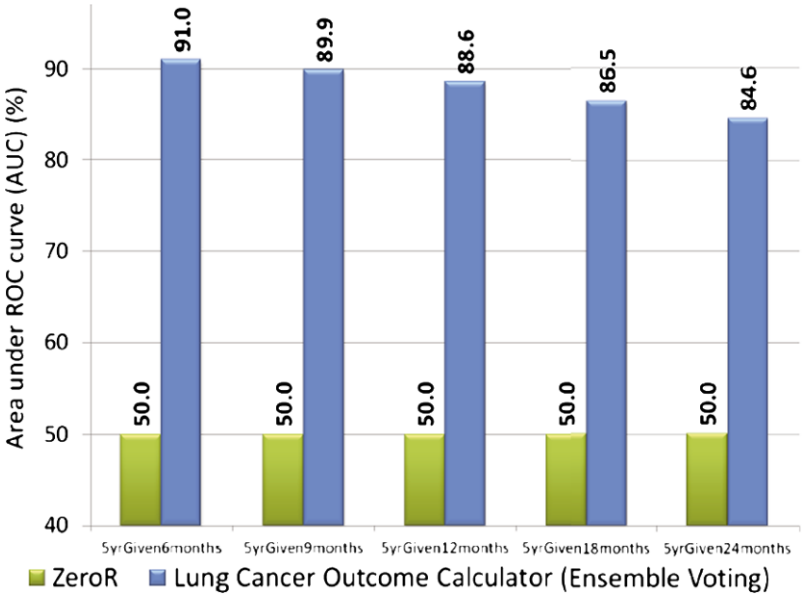

Fig. 7. Prediction performance comparison in terms of area under the ROC curve (AUC) for conditional survival models. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/SPR-2012-0335.)

8. Conclusion and future work

In this paper, we used different meta classification schemes with underlying decision tree classifiers to construct models for survival prediction for respiratory cancer patients. Prediction accuracies of 73.61%, 74.45%, 76.80%, 85.45% and 91.35% was obtained for the 6-month, 9-month, 1-year, 2-year and 5-year respiratory cancer survival prediction using the ensemble voting classification scheme, and a lung cancer outcome calculator was developed using carefully selected 13 attributes, while retaining the prediction quality. Further, conditional survival models were also constructed to estimate patient-specific mortality risk after 5 years of diagnosis for patients who have already survived for 6, 9, 12, 18 and 24 months.

Given the prediction quality, we believe that the calculator can be very useful to not only accurately estimate survivability of a lung cancer patient, but also aid doctors in decision making and improve informed patient consent by providing a better understanding of the risks involved in a particular treatment procedure, based on patient-specific attributes. Accurate risk prediction can potentially also save valuable resources by avoiding high risk procedures that may not be necessary for a particular patient.

Future work includes exploring the use of undersampling/oversampling to deal with unbalanced data, and using multi-stage classification [33]. We also plan to do similar analysis for other cancers, and developing on-line cancer outcome calculators for them.

Acknowledgements

This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, OCI-1144061, CNS-0551639, IIS-0536994 and in part by DOE grants DE-FC02-07ER25808, DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309 and DE-SC0005340.

References

[1] A. Agrawal and A. Choudhary, Association rule mining based hotspot analysis on seer lung cancer data, *International Journal of Knowledge Discovery in Bioinformatics* **2**(2) (2011), 34–54.

[2] A. Agrawal and A. Choudhary, Identifying hotspots in lung cancer data using association rule mining, in: *Proc. 2nd IEEE ICDM Workshop on Biological Data Mining and Its Applications in Healthcare*, IEEE Computer Society, 2011, pp. 995–1002.

[3] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi and A. Choudhary, A lung cancer outcome calculator using ensemble data mining on seer data, in: *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics*, ACM Press, New York, NY, 2011, pp. 5:1–5:9.

[4] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi and A. Choudhary, Poster: a lung cancer mortality risk calculator based on seer data, in: *Proceedings of IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences*, IEEE Computer Society, 2011, pp. 233–233.

[5] A. Bellaachia and E. Guven, Predicting breast cancer survivability using data mining techniques, in: *Proceedings of 9th Workshop on Mining Scientific and Engineering Datasets in Conjunction with the 6th SIAM International Conference on Data Mining*, Bethesda, MD, 2006.

[6] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

[7] L. Breiman, Random forests, *Machine Learning* **45**(1) (2001), 5–32.

[8] Centers for Disease Control and Prevention, Lung cancer statistics, available at: http://www.cdc.gov/cancer/lung/statistics/, accessed: April 29, 2010.

[9] C.-C. Chang and C.-J. Lin, *LibSVM – A Library for Support Vector Machines*, 2001, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] D. Chen, K. Xing, D. Henson, L. Sheng, A. Schwartz and X. Cheng, Developing prognostic systems of cancer patients by ensemble clustering, *Journal of Biomedicine and Biotechnology* **2009** (2009), 632786.

[11] M. Choi, C.D. Fuller, C.R. Thomas Jr. and S.J. Wang, Conditional survival in ovarian cancer: results from the seer dataset 1988–2001, *Gynecologic Oncology* **109**(2) (2008), 203–209.

[12] N. Coburn, A. Govindarajan, C. Law, U. Guller, A. Kiss, J. Ringash, C. Swallow and N. Baxter, Stage-specific effect of adjuvant therapy following gastric cancer resection: a population-based analysis of 4,041 patients, *Annals of Surgical Oncology* **15** (2008), 500–507.

[13] D. Delen, G. Walker and A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, *Artificial Intelligence in Medicine* **34**(2) (2005), 113–127.

[14] Y. El-Manzalawy and V. Honavar, *WLSVM: Integrating LibSVM into Weka Environment*, 2005, available at: http://www.cs.iastate.edu/~yasser/wlsvm.

[15] A. Endo, T. Shibata and H. Tanaka, Comparison of seven algorithms to predict breast cancer survival, *Biomedical Soft Computing and Human Sciences* **13**(2) (2008), 11–16.

[16] L. Fausett, *Fundamentals of Neural Networks*, Prentice Hall, New York, NY, 1994.

[17] D. Fradkin, Machine learning methods in the analysis of lung cancer survival data, DIMACS Technical Report 2005-35, February 2006.

[18] Y. Freund and L. Mason, The alternating decision tree learning algorithm, in: *Proceeding of the 16th International Conference on Machine Learning*, Morgan Kaufmann, Citeseer, 1999, pp. 124–133.

[19] J. Friedman, T. Hastie and R. Tibshirani, Special invited paper. Additive logistic regression: a statistical view of boosting, *Annals of Statistics* **28**(2) (2000), 337–374.

[20] M. Hall, Correlation-based feature selection for machine learning, PhD thesis, Citeseer, 1999.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The weka data mining software: an update, *SIGKDD Explorations* **11**(1) (2009), 10–18.

[22] T. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8) (1998), 832–844.

[23] J. Kittler, Combining classifiers: a theoretical framework, *Pattern Analysis and Applications* **1**(1) (1998), 18–27.

[24] M. Lundin, J. Lundin, H.B. Burke, S. Toikkanen, L. Pylkkanen and H. Joensuu, Artificial neural networks applied to survival prediction in breast cancer, *Oncology* **57** (1999), 281–286.

[25] National Cancer Institute, Surveillance, epidemiology and end results (seer) program (www.seer.cancer.gov) limited-use data (1973–2006). DCCPS, Surveillance Research Program, Cancer Statistics Branch, 2008. released April 2009, based on the November 2008 submission.

[26] National Cancer Institute, Introduction to lung cancer, SEER training modules, available at: http://training.seer.cancer.gov/lung/intro/, accessed: April 29, 2010.

[27] Overview of the seer program, Surveillance Epidemiology and End Results, available at: http://seer.cancer.gov/about/, accessed: April 29, 2010.

[28] J. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

[29] L.A.G. Ries and M.P. Eisner, Cancer of the lung, *SEER Survival Monograph: Cancer Survival Among Adults: US SEER Program, 1988–2001, Patient and Tumor Characteristics*, SEER Program, National Cancer Institute, Bethesda, MD, 2007.

[30] L.A.G. Ries and M.P. Eisner, *Cancer of the Lung*, Chapter 9, SEER Program, National Cancer Institute, 2007.

[31] L.A.G. Ries, M.E. Reichman, D.R. Lewis, B.F. Hankey and B.K. Edwards, Cancer survival and incidence from the surveillance, epidemiology, and end results (SEER) program, *Oncologist* **8**(6) (2003), 541–552.

[32] K. Rusthoven, T. Flaig, D. Raben and B.D. Kavanagh, High incidence of lung cancer after non-muscle-invasive transitional cell carcinoma of the bladder: implications for screening trials, *Clin. Lung Cancer* **9** (2008), 106–111.

[33] T.E. Senator, Multi-stage classification, in: *Proceedings of the 5th IEEE International Conference on Data Mining*, IEEE Computer Society, Washington, DC, 2005, pp. 386–393.

[34] J. Thongkam, O. Xu, Y. Zhang and F. Huang, Breast cancer survivability via adaboost algorithms, in: *Proceedings 2nd Australasian Workshop on Health Data and Knowledge Management*, Wollongong, NSW, Australia, 2008.

[35] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[36] S.J. Wang, R. Emery, C.D. Fuller, J.-S. Kim, D.F. Sittig and C.R. Thomas, Conditional survival in gastric cancer: a seer database analysis, *Gastric Cancer* **10** (2007), 153–158.

[37] S.J. Wang, C.D. Fuller, R. Emery and C.R. Thomas, Conditional survival in rectal cancer: a seer database analysis, *Gastrointest. Cancer Res.* **1** (2007), 84–89.

[38] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, CA, 2005.

[39] L. Ya-Qin, W. Cheng and Z. Lu, Decision tree based predictive models for breast cancer survivability on imbalanced data, in: *Proceedings of the 3rd International Conference on Bioinformatics and Biomedical Engineering*, Beijing, 2009, pp. 1–4.

[40] J.C. Yao, M. Hassan, A. Phan, C. Dagohoy, C. Leary, J.E. Mares, E.K. Abdalla, J.B. Fleming, J.-N. Vauthey, A. Rashid and D.B. Evans, One hundred years after 'carcinoid': epidemiology of and prognostic factors for neuroendocrine tumors in 35,825 cases in the United States, *Journal of Clinical Oncology* **26**(18) (2008), 3063–3072.

[41] Z.-H. Zhou and Y. Jiang, Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble, *IEEE Transactions on Information Technology in Biomedicine* **7**(1) (2003), 37–42.