Journal of Information
2015 Vol.1, No.1, pp.1-11
ISSN(e): 2520-7652
DOI: 10.18488/journal.104/2015.1.1/104.1.1.11
© 2015 Conscientia Beam. All Rights Reserved.

PREDICT SURVIVAL OF PATIENTS WITH LUNG CANCER USING AN ENSEMBLE FEATURE SELECTION ALGORITHM AND CLASSIFICATION METHODS IN DATA MINING

Mahdis Dezfuly[1] --- Hedieh Sajedi[2]†

[1]*Department of Electrical and Computer Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran*

[2]*Department of Computer Science, School of Mathematics, Statistics, and Computer Science, College of Science, University of Tehran, Tehran, Iran*

ABSTRACT

This research proposes an efficient model for predicting the survival rate of patients affected by lung cancer. The researchers collected data from four feature categories (population, recognition, treatment, and result) of cancer patients based on the importance of the survival of patients with lung cancer. Analyses of the predicted survival rates of the patients indicate that, among the classification algorithms, Decision Tree C5.0 results the highest accuracy. The models were created using algorithms based on the level of death risk in five stages: six months, nine months, one year, two years, and five years. In this paper, we proposed a mechanism for feature selection. Our mechanism combines the results of some feature section algorithm. The results illustrate that out mechanism outperform other feature selection algorithms. After applying the proposed mechanism for feature selection, the accuracy of the C5.0 algorithm was equivalent to 97.93%.

Keywords: Lung cancer, Data mining, Decision tree algorithm, Bayes network, Neural network.

Contribution/ Originality

This study proposes an Ensemble feature selection algorithm for predict survival of patients with lung cancer.

1. INTRODUCTION

Lung cancer is a type of cancer that occurs when malignant tissue develops in either one or both lungs. It is one of the most prevalent cancers all over the world, and more than 80% of patients die less than five years from the time the illness is diagnosed.

Data mining is used with a huge amount of data in order to discover unseen information and is available through various data stores. The data which exists in those stores is massive, but worthless because the data's value is in existent knowledge, and alone is not valuable. The aim of

† Corresponding author

discovering this knowledge is to have a true formative diagnosis in the data base process in order to achieve understandable models and samples of data.

A data base is made in order to preserve the clinical information which is used in the health industry. This data base is the information from the people affected. Every day, a huge amount of information about patients is saved in their data base. One of these stations is the authentic National Cancer Institute site that is called the "SEER[1] Site", which is used as a statistical cancer data base in the United States. Access to SEER's data is limited, and is only possible through the SEER website by sending agreement and confirmation to the site [1]. A review of cancer data items, and data collection are combined in SEER. SEER data features can be used with the various different categories considered, such population attributes (age, gender, location), recognition attributes (the early part of the disease, cancer grade), treatment attributes (surgery, radiation therapy), and result attributes (survival time record, cause of death). This information is added to the SEER data for scientific study and analysis of the data when appropriate [1]. The features of the SEER's data are considered in several different parts; Demographic features, diagnostic features, treatment features, and the result of the features as information. This information makes SEER data proper for scientific observation and data analysis. The data related to the period of the patient's survival is divided into five parts, to put each patient in a specific category (deceased or living), through this specification [2]. These five categories are named in order: (Y1, M9, M6, Y2, and Y5). For modeling, algorithms tailored to the lung cancer data used in this study have led to the production of various models for medical data mining. In this study, decision tree algorithms, Bayesian networks, and neural networks have been applied. Because studies in the field of data mining have been done to show that these algorithms outperformed other algorithms, this study tries to follow data analysis phase based on the existing data and predict the patient's survival by employing the classification methods.

In next the sections we explain the methods and experiments. Section 2 is about related works. Section 3 explains the methodology. Section 4 presents the experimental results and, Section 5 explains the conclusion and future goals of this research.

2. RELATED WORKS

Researchers are always trying to predict things before they occur in order to reduce costs and risks, prevent the inevitable harmful events, and increase people's life span. The health industry is one of the most important fields in this regard. As a result, there are a great number of health research centers in many countries in which a lot of research and studies are being conducted. Some of these studies are as follows [2] analyzed data from the SEER database to provide a model for predicting survival in lung cancer patients. Their data were used in the 1998-2001 SEER databases. In this study, data mining techniques used to predict the survival of patients

[1] Surveillance, Epidemiology and End Results

with cancer of the respiratory tract at the end of 6 month, 9 month, 1 year, 2 year, and 5 year periods. In this study, 10 data mining algorithms were used, such as Support Vector Machine (SVM), Artificial Neural Network (ANN), J48Decision tree, and so forth. The results of this study showed that five of the algorithms, J48Decision Tree, Random Forests, Logit Boost, Random Subspace, and Alternating Decision Tree worked in a timely manner. They used polling methods to evaluate the algorithms, which resulted in greater accuracy for the Decision Tree algorithm. However, they suggested that SVM and Neural Network algorithm are not suitable for large data sets [2]. Lang, et al. [3] studied survival improvement factors in elderly patients suffering from cortical cancer in the United States based on population analysis. The data collected for their study, based on the SEER data base and data related to population, diagnosis, and treatment features of colon and rectal cancer patients was analyzed. In order to increase accuracy, each feature was analyzed individually. The five-year survival rate of patients was analyzed based on the Statistical Multivariable Logistic Regression Model. The results indicated that prediction of the survival in patients suffering from colon and rectal cancer improved, with a 95% degree of certainty in colon cancer from 3.46% to 43%, and in rectal cancer from 2.42% to 4.39% [3].

Palaniappan and Rafiah [4] presented a system for predicting heart disease based on designing techniques in three classifications. 909 records were under study and three methods; decision tree, neural network, and the simple Bayes, were used. The results indicated that the decision tree method results were easily identifiable. The simple Bayes method displayed better results in identifying the prediction factors more easily. Finally, the results of neural network method were very complicated [4]. Delen, et al. [5] compared the ANN method, Decision tree method, and logistic regression method for predicting breast cancer outcomes. They used 20 variables from the SEER data base in the prediction model and found that the decision tree method, with a precision of about 93.6 %, and the ANN method with 91.2% accuracy, are superior in comparison to the logistic method [5].

Lundin, et al. [6] analyzed the survival rates of patients suffering from breast cancer with the techniques and the SEER breast cancer data. They used the ANN method on 951 patient records in the central hospital of Turku University and the Turku City hospital in order to improve and check the precision of the neural network for periods of 5, 10, and 15 years of survival in breast cancer patients. The experimental values of the curve (cause of death), ROC was estimated for 5 years (0.909), for 10 years (0.086), and for 15 years (0.883). These values are a measure of the accuracy of the prediction model that was used. This research estimated the survival for 300/82 false predictions of logistic regression, with 300/43 compared to the ANN and found that the ANN accurately predicted survival Lundin, et al. [6].

3. METHODOLOGY

The research methodology is described in this section. A data mining process based on CRISP methodology includes business understanding, data understanding, data preparation, modeling, evaluation, and development. CRISP-DM [2] is a data mining process model that defines commonly used methods that data mining experts use to throw problems [7]. In this study, we employ a set of classification algorithms includes decision tree, Bayes network and neural network.

Decision tree Algorithm: Decision trees are used based on a set of decision making rules for prediction and classification. Because the results are considered to be binary, the outcome is a binary tree. We employ four algorithms, C&R, CHAID, QUEST, and C5.0 to analyze our data [8].

Neural network: In this algorithm, the processing units are put in order in different layers. There are usually tree parts in a neural network; a hidden layer, an internal layer, and an external layer. The units are attached to each other in different weights. There are six methods for making a neural network; Quick, Dynamic, Multiple, Prune, RFBN, and Exhaustive prune [8]. Since there is no limitation in the features of this method, the neural network can work with statistical, classification, and binary input and output. Neural networks are powerful speculators since they predict as well as the other predicators so in this study the two models, Multiple and Quick, are being used. Four other models of the neural network were not compatible with the input since a lot of time is spent on designing the model, so these methods were not efficient because of time consuming.

Bayes network: The Bayes model enables us to design a model by blending observation and matching the knowledge with the evidence in order to arrange the occurrences with the miscellaneous features. Since, in this study, the network improves our cause and effect knowledge and prevents to connect the data, and illustrates the connections easily, the input can be from any kind of field study. And, the networks will be very durable and have the best prediction based on the data [8]. Because of the aforementioned reasons, both methods will be used in this study.

According to the above statements regarding the method, in the following, we will describe the data collection, data preparation, and data analysis.

3.1. Data Collection and Preparation

In order to collect data from the SEER data base, the current research used SEER State software, which enabled us to find the features as well. As mentioned earlier, SEER stands for Surveillance, Epidemiology and End Results [9], which is one of the national cancer institutes and a cancer data center in the United States [10]. Cancer patients' data is classified into four categories: population, recognition, treatment, and results. In the current research, selecting the

[2] Cross Industry Standard Process for Data Mining

parameters from the aforementioned categories is one of the most significant ways of obtaining the result. Also, data was selected from the records of lung cancer patients diagnosed from 1988 through 2011. The categories are as the follows:

Population attributes include: age at diagnosis and birth place.

Recognition attributes include: cancer grade, diagnostic confirmation, farthest extension of tumor, lymph node involvement, cancer stage, and number of malignant tumors in the past.

Treatment attributes include: type of surgery performed, reason for surgery not being performed, and order of surgery and radiation therapy, and scope of regional lymph node surgery.

Results attributes include: survival time record, cause of death, and vital status

The categories above are data features of the following four steps which outline a simple strategy for conducting effective research on data mining. The researchers clean up the data with the following steps:

1. Excluding the records of patients that died due to reasons other than lung cancer.
2. Eliminating features that are insignificant and do not affect the results, since the number of such features is limited.
3. In the results features, after filtering, the COD and VSR data are considered since it is very important in order to analyze the lung cancer survival features.
4. Eliminating the empty patient records.

Although, Patient's Survival Time Record (STR) is saved in the data base based on the year and month, the researchers used monthly information in this study since it is more precise. Cause of Death (COD) can also be different since a patient suffering from lung cancer can die for reasons other than the cancer itself. Vital Status Record (VSR) indicates that whether a patient is currently alive or not.The total number of data records collected was 296,716, and after cleaning up the data, was reduced to 73,961.Data on patient survival time after diagnosis was divided into five categories, and we also wanted to examine the patient characteristics in each category, living or deceased. The results of this study should be binary.

3.2. Data Analysis

In the modeling process, various modeling techniques are chosen and tested on the prepared data, the parameters are adjusted, and the prepared data is divided for education and testing purposes. Thus, the models are made and assessed. In order to design a model based on the medical data, the educational and experimental data should be specified beforehand. Classification and prediction precision is assessed in this section. In order to have good validity and reliability there are a variety of methods available; random sampling, cross validation, bootstrap, and holdout are common techniques used in assessments that are done on parts of the data based on random sampling [11]. The current study uses the holdout technique. This method randomly divides the data into two individual parts, Train and Test. In a particular division, 2/3 of the data is related to the train section, and 1/3 is related to the test section. In fact 66% of the data is used for the learning and 33% is used for the testing. With this technique, it is possible to increase the reliability and validity of the calculations, and it is also efficient for choosing the best model. In order to make the models, the crisp method and clementine software [12] are used.

4. PROPOSED MODEL

In order to increase the accuracy of the modeling in the process of data preparation, the study uses feature selection methods to identify important properties. The main purpose of feature selection is to find a minimum subset of features, and increase or preserve the accuracy of prediction. Also, feature selection in classification is important, because in these matters, there are many features that may not have much load information. While these features do not help the classification, they increase computational expense. In this regard, we propose a feature selection mechanism, which survives helpful information and useful data. Accordingly, this study suggests an ensemble model. Thus, the advantage of this suggested model is decrementing the errors. This method employs several feature selection algorithms and output characteristics, and then combines them into a final set based on the significance and importance of the particular properties, which are achieved by voting in the final stage of the ensemble model. The performance of this method is more than when an individual feature selection algorithm is used.

Different feature selection algorithms have been applied on *n* features. Finally, resulted selected features will be investigated by counting. In the results of feature selection algorithms, more significant features are more frequent.

Hence, in order to remove the less important features, Backward Feature Elimination method is applied recursively. The overall structure of the model is shown in Figure 1.

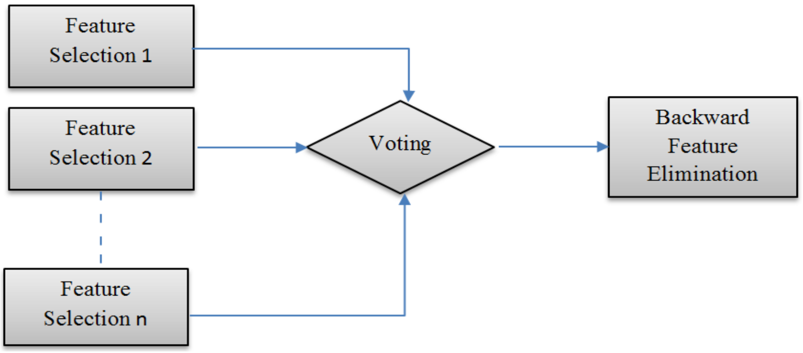


Figure-1. Structure of the proposed feature selection mechanism

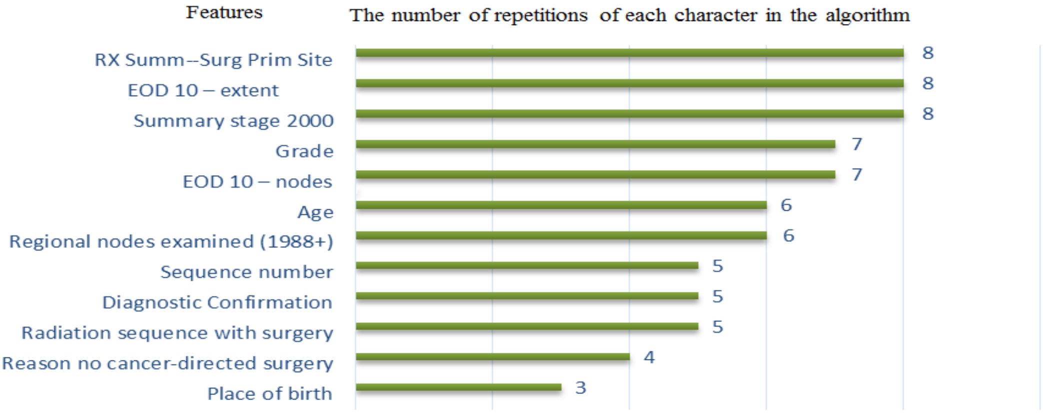


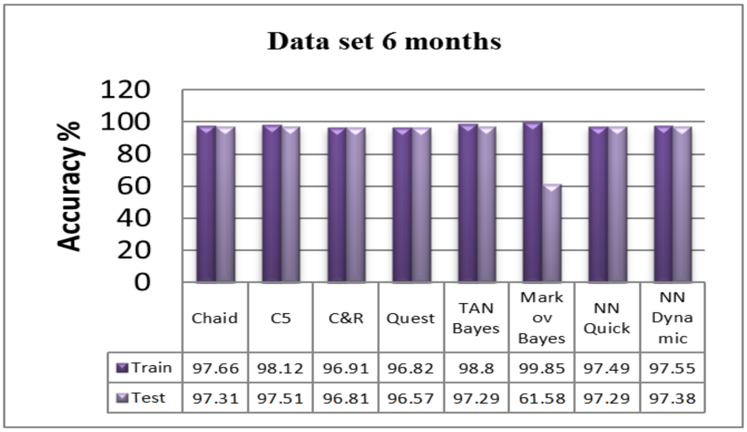
Figure-2. The number of repetitions of each feature by eight algorithms in data set of 6 months

In Figure 2, the number of repetitions of each feature by eight algorithms is shown. For example, the feature RX SUMM-SURG PRIM SITE is an important feature of all eight algorithms, and PLACE OF BIRTH is known to be significant in three algorithms. For this reason, the least important feature is removed, and the modeling is performed. As long as accuracy increases or remains constant, features are removed.

5. EXPERIMENTAL RESULTS

In order to examine the developed model, the method is carried out, and compared to other algorithms used in the test section, in the six month data set. The C5.0 algorithm has the greatest accuracy, equal to 97.51%, and in the education section of this data set, the Markov algorithm has the greatest accuracy, equal to 99.85% (Figure 3).

In comparison to the other algorithms used in the test section, in the nine month data set, the Quick algorithm has the greatest accuracy, equal to 97.44%, and in the education section of this data set, the TAN algorithm has the greatest accuracy equal to 99.39% (Figure 4).



	Chaid	C5	C&R	Quest	TAN Bayes	Markov Bayes	NN Quick	NN Dynamic
▨ Train	97.66	98.12	96.91	96.82	98.8	99.85	97.49	97.55
▨ Test	97.31	97.51	96.81	96.57	97.29	61.58	97.29	97.38

Figure-3. Comparison of different techniques in the 6 month data set

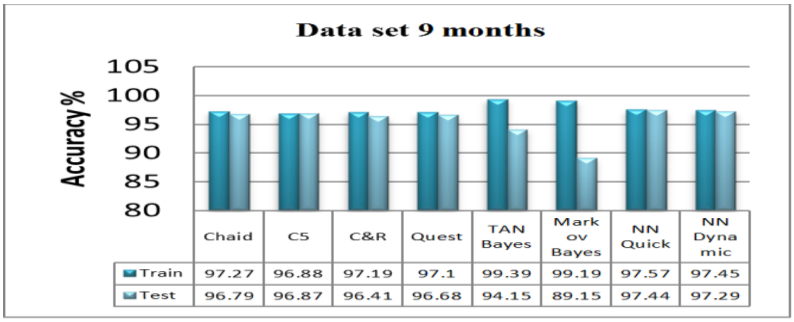


Figure-4. Comparison of different techniques in the 9 month data set

In comparison to the other algorithms used in the test section, in the one year data set, the Quick algorithm has the greatest accuracy, equal to 97.25%, and in the education section of this data set, the TAN algorithm has the greatest accuracy, equal to 99.51% (Figure 5).

In comparison to the other algorithms used in the test section, in the two year data set, the Quick algorithm has the greatest accuracy, equal to 95.81%, and in the education section of this data set, the Markov algorithm has the greatest accuracy, equal to 99.53% (Figure 6).

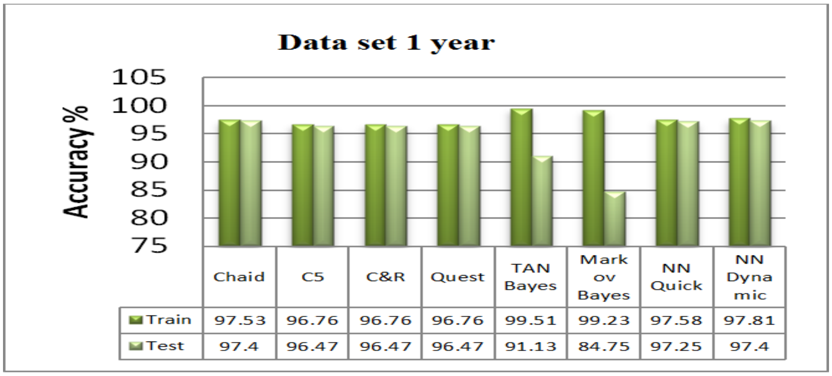


Figure-5. Comparison of different techniques in the 1 year data set

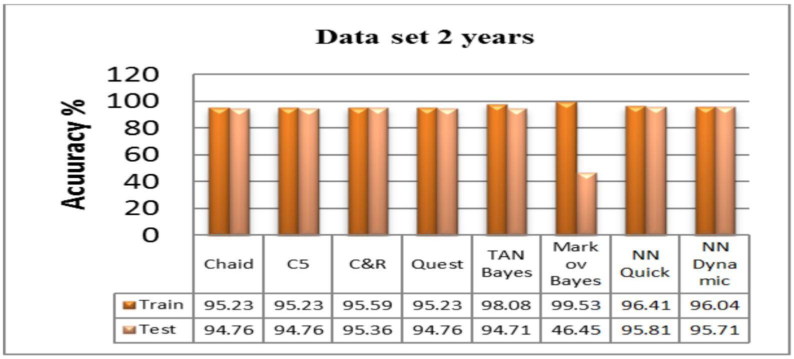
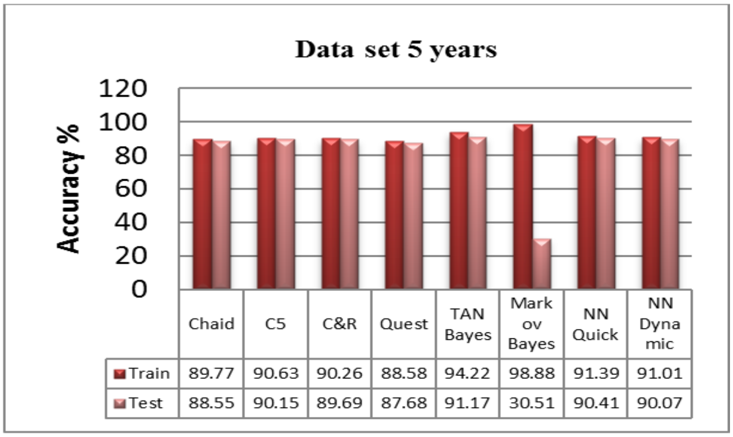


Figure-6. Comparison of different techniques in the 2 year data set

In comparison to the other algorithms used in the test section, in the five year data set, the TAN algorithm has the greatest accuracy, equals to 91.15%, and in the education section of this data set, the Markov algorithm has the greatest accuracy, equal to 98.88% (Figure 7).



	Chaid	C5	C&R	Quest	TAN Bayes	Mark ov Bayes	NN Quick	NN Dyna mic
Train	89.77	90.63	90.26	88.58	94.22	98.88	91.39	91.01
Test	88.55	90.15	89.69	87.68	91.17	30.51	90.41	90.07

Figure-7. Comparison of different techniques in the 5 year data set

In the education section of the six month data set, by considering the complete results, it is clear that the decision tree algorithm and C5.0 test and data set have the greatest accuracy, and this accuracy in is the test related to the model. The analysis in the Markov education, related to the model, shows that among the algorithms in the test section, the model made with the decision tree (CHAID, C5.0, C&R, and Quest), has the greatest accuracy.

Table-1. Comparison of initial results and the proposed model in C5.0, C&R, and CHAID

Data Set	**C5.0**		**C&R**		**CHIAD**	
	Accuracy before applying the proposed model	Accuracy after applying the proposed model	Accuracy before applying the proposed model	Accuracy after applying the proposed model	Accuracy before applying the proposed model	Accuracy after applying the proposed model
6 months	97.51%	97.93%	96.82%	96.82%	96.50%	96.56%
9 months	96.87%	96.94%	96.41%	96.41%	96.79%	96.87%
1 year	96.47%	96.91%	96.47%	96.47%	96.40%	96.67%
2 years	94.76%	96.31%	95.36%	95.58%	94.76%	95.34%
5 years	90.75%	93.12%	89.86%	89.86%	89.43%	89.54%

In C5.0, the creation model method (Quick, Dynamic) of neural network algorithms has the greatest accuracy, and among the two made "Quick" by Bayes network, the model created (TAN, Markov) by the algorithm, shows the greatest accuracy. The results related to the education section demonstrate that in all of the algorithms under study, Bayes network has done the best job. In Tables (1) and (2), the modeling, before and after applying the proposed method, are compared. The modeling accuracy of the data before deleting the least important features, and then after removing the least important features is shown.

Table-2. Comparison of initial results and the proposed model in Quest, NN Quick, and Bayes_TAN

Data Set	**Quest**		**NN_Quick**		**Bayes_TAN**	
	Accuracy before applying the proposed model	Accuracy after applying the proposed model	Accuracy before applying the proposed model	Accuracy after applying the proposed model	Accuracy before applying the proposed model	Accuracy after applying the proposed model
6 months	96.58%	96.57%	97.33%	97.41%	97.29%	97.29%
9 months	96.68%	96.68%	97.44%	97.44%	94.15%	97.02%
1 year	96.47%	96.47%	97.21%	97.40%	91.18%	97.11%
2 years	94.76%	95%	95.53%	95.86%	94.71%	95.51%
5 years	88.91%	88.91%	91.32%	91.32%	91.91%	91.91%

According to the results of this study, the concluding remarks will increase the accuracy of the feature selection method. By eliminating the least important features of a small number of features, the accuracy of the models is increased or remains constant. Then, having a small number of features, physicians can make better decisions to improve the health of the patients. According to studies by other researchers done using the same features, it can be concluded that the proposed method has greater accuracy than methods used in previous studies, and with a smaller number of features, accuracy has been increased in comparison to previous methods.

6. CONCLUSION

To improve the accuracy of the survival prediction of patients with Lung cancer, we examined the importance of the features, and the characteristics that were less important. It is shown that by employing our ensemble of feature selection methods and eliminating less important features we are able to achieve greater accuracy. The experimental results illustrate that our proposed feature selection model is more effective than previous feature selection algorithms and improves the prediction accuracy more. In future studies, in order to make a comparison to the model proposed in this study, it would be possible to make new models with the help of data analysis on lung cancer data related to the SEER site.

REFERENCES

[1] L. GloecklerRies, A. M. Reichman, D. Lewis, R. B. F. Hankey, and B. K. Edwards, "Cancer survival and incidence from the surveillance, epidemiology, and end results (SEER) Program," *Oncologist*, 2003.

[2] A. Ankit, M. Sanchit, N. Ramanathan, P. Lalith, and C. Alok, "A lung cancer outcome calculator using ensemble data mining on SEER data," *Electrical Engg. and Computer Science Northwestern University*, 2011.

[3] K. Lang, J. Korn, D. W. Lee, L. M. Lines, C. C. Earle, and J. Menzine, "BMC Cancer, USA," 2009.

[4] S. Palaniappan and A. Rafiah, "Intelligent heart disease prediction system using data mining techniques. Department of information technology Malaysia university of science and technology," 2008.

[5] 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.

[6] M. Lundin, J. Lundin, H. BurkeB, S. Toikkanen, L. Pylkkänen, and H. Joensuu, "Artificial neural networks applied to survival prediction in breast cancer," *Oncology International Journal for Cancer Resaerch and Treatment*, vol. 57, pp. 281-286, 1999.

[7] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *J. Data Warehousing*, vol. 5, pp. 13-22, 2000.

[8] M. Kantardzic, *Data mining: Concepts, models, methods, and algorithms*, 2nd ed. Simltaneously in Canada: WILEY, 2011.

[9] SEER, "Surveillance, epidemiology, and end results (SEER) program (www.seer.cancer.gov) limited-use data (1973-2006)," National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch2009.

[10] SEER, "Overview of the seer program," *Surveillance Epidemiology and End Results.* Available http://seer.cancer.gov/about/, 2014.

[11] M. Green and M. Ohlsson, "Comparision of standard resampling methods for performance estimation of artificial neural network ensembles. Computational biology and biological physics group, department of theoretical physics, Lund University," 2006.

[12] Clementine® 12.0 Algorithms Guide, *SPSS Inc. 233 South Wacker Drive*, 11th ed. Chicago, IL 60606-6412 Copyright © by Integral Solutions Limited, 2007.

Views and opinions expressed in this article are the views and opinions of the author(s), Journal of Information shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.