Colon cancer survival prediction using ensemble data mining on SEER data

Reda Al-Bahrani, Ankit Agrawal, Alok Choudhary

Dept. of Electrical Engg. and Computer Science

Northwestern University

Evanston, IL 60208, US

{rav650,ankitag,choudhar}@eecs.northwestern.edu

Abstract—**We analyze the colon cancer data available from the SEER program with the aim of developing accurate survival prediction models for colon cancer. Carefully designed preprocessing steps resulted in removal of several attributes and applying several supervised classification methods. We also adopt synthetic minority over-sampling technique (SMOTE) to balance the survival and non-survival classes we have. In our experiments, ensemble voting of the three of the top performing classifiers was found to result in the best prediction performance in terms of prediction accuracy and area under the ROC curve. We evaluated multiple classification schemes to estimate the risk of mortality after 1 year, 2 years and 5 years of diagnosis, on a subset of 65 attributes after the data clean up process, 13 attribute carefully selected using attribute selection techniques, and SMOTE balanced set of the same 13 attributes, while trying to retain the predictive power of the original set of attributes. Moreover, we demonstrate the importance of balancing the classes of the data set to yield better results.**

Keywords—Prediction, Ensemble, Colon Cancer, Machine Learning

I. INTRODUCTION

Cancers of the colon and rectum are of two of the most common types worldwide. Early diagnosis and treatment can greatly improve the chances of survivability [1].

The Surveillance, Epidemiology, and End Results (SEER) Program is a premier source of domestic statistics of cancer. The collected data from SEER represents 28% percent of the US population across several geographic regions. This data is available from the SEER website upon submitting a SEER limited-use data agreement form.

In this paper we analyze the colon cancer data available from the SEER program with the aim of developing accurate survival prediction models for colon cancer. The data analyzed in this study is from the surveillance, epidemiology and end results (SEER) Colon and Rectum cancer incidence data in the years of 1973-2009. The SEER Colon and Rectum cancer incidence data consist of four datasets named yr1973_2009.seer9, yr2000_2009.ca_ky_lo_nj_ga, yr2005.lo_2nd_half, and yr1992_2009.sj_la_rg_ak. The follow-up cutoff date of the datasets is December 31, 2009 [2].

Here we use supervised classification methods to predict survival of colon cancer patients, at the end of 1 year, 2 years and 5 years of diagnosis. We carried experiments with several classifiers to find that many meta classifiers used with decision trees and functions can give better results compared to basic classifiers. These results can be improved by adopting SMOTE to balance the survival and non-survival classes, and by combining the resulting prediction probabilities from several classifiers using an ensemble-voting scheme.

The rest of the paper is organized as follows: Section 2 summarizes related work, followed by a brief description of the prediction system used in this study in Section 3. A description of the data used in this work is described in Section 4. In Section 5 a list of the classification schemes used in the study is presented along side with a brief description. Experiments and results are presented in Section 6, followed by the conclusion and future work in Section 7.

II. BACKGROUND

With SEER data being publicly available, there is a mature literature on SEER data studies. SEER provides SEER*Stat a statistical software which provides convenience to analyze the data.

In addition, there have been data mining applications developed for various types of cancer based on SEER data. A number of techniques based on data mining have been proposed for the survivability analysis of various types of cancer. Zhou and Jiang [3] used decision trees and artificial neural networks for survivability analysis of breast cancer. Delen et al. [4] empirically compared neural networks, decision trees and, logistic regression for predicting breast cancer survivability. Comparisons of different methods were made by Endo et al. [5] on 5-year survivability of breast cancer diagnosed patients. They compared seven methods: artificial neural

network, naïve bayes, bayes network, decision trees with naive bayes, decision trees (ID3), decision trees (J48) and logistic regression models. Chen et al. [6] used SEER data to study survival patterns in of lung cancer. Also, Fradkin [7] used SEER data to study 8 months survivability of lung cancer for patients diagnosed between the years of 1988 and 2001.

Studies were also conducted on colorectal cancer. Fathy [8] studied colorectal cancer survivability prediction rates in relation to the number of hidden neurons in the Artificial Neural Networks (ANN). Stojadinovic et al. [9] describe how they utilize a data analytics suite named FasterAnalytics to build a machine-learned Bayesian Belief Network (ml-BBN) model for clinical decision support (CDS). Wang et al. [10] analyze colorectal cancer survival based on stage, age, gender, and race. Data mining applications and studies of colorectal cancer are not covered as much as breast or lung cancers.

III. PREDICTION SYSTEM

The most important step of the process is to understand and clean the data. Our system consists of several stages. These stages consist of: SEER-related preprocessing, Problem specific preprocessing, Predictive modeling and Evaluation, as depicted in Figure 1.
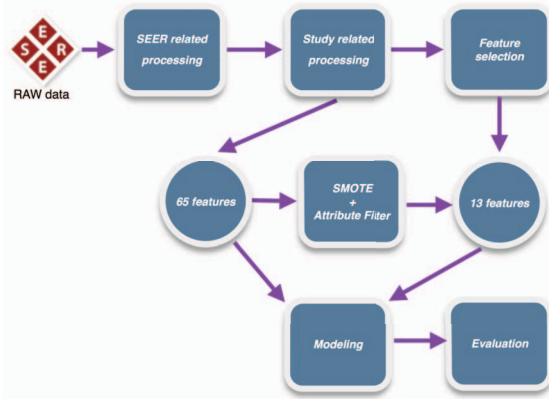

Fig. 1. Prediction System Flow

1) SEER-related preprocessing. The data provided by SEER is in raw format. A script was developed to convert the data into csv format. Also, in this stage the following conversions/calculations were performed on the datasets to format the raw data to appropriate values

a) Convert apparently numeric attributes to nominal, e.g., marital status, sex

b) Convert Size of Tumor to cm from SEER's coding. E.g. code 100 is equivalent to 10.0 cm

c) Calculate the survival time in months (numeric) from SEER's format of YYMM

2) Problem-specific preprocessing. In this stage the following filters were applied and class attributes were derived

a) Filter data records for the period of interest. Period of [1998, 2003]

b) Filter data records that are related to the cancer in study. Primary Site [C18.0-C18.9]

c) Filter all attributes that may indicate that vital status of the patient

d) Derive appropriate binary attributes for survival, e.g., 5-year survival

e) Remove attributes that do not vary at all or that vary too much. Constant attributes are removed, and attributes that exceed a maximum variance threshold e.g. 99%

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 data into training and testing sets or use cross-validation

b) Conducting experiments using the different classification schemes

4) Evaluation. In this stage the models were compared with respect to different metrics from the predictive modeling stage. These metrics include:

a) Percent of correct classified instances

b) Area Under the Receiver Operating Characteristic (ROC Curve)

IV. DATA CLEANUP

The data used in these experiments is from the period 1998 to 2003. Since the follow-up cutoff date of the datasets is December 31, 2009 and a large number of the attributes are applicable for records of years 2004+, we considered data until December 31, 2003. This decision was made to conduct the study for 5-year survivability. Also, to minimize the number of missing data due to the applicability of the attribute we have, our datasets start from January 1, 1998.

Since SEER data of colon and rectum cancers are represented together we had to filter the data to include only colon cancer cases. Any instances with cause of death not related to colon cancer were removed.

After finishing the cleanup process we had a total of 65 attributes plus the class. The class can be 1 year, 2 years, or 5 years survivability. The data from SEER consisted of 788,892 records it was truncated to 105,133 records after selecting the period and type of cancer. Table 1 shows the class distribution of the data used in our experiments.

patient at diagnosis for this cancer.

10) Diagnostic Confirmation: Records the best method used to confirm the presence of the cancer being reported. The data item is not limited to the confirmation at the time of diagnosis; it is the best method of confirmation during the entire course of the disease.

11) EOD-Tumor Size: Records the largest dimension of the primary tumor in millimeters.

12) Behavior (92-00) ICD-O-2: Behavior codes of the cancer.

13) Primary Site: Identifies the site in which the primary tumor originated.

V. CLASSIFICATION SCHEMES

The classification schemes used in our experiments are of two types: basic classifiers, and meta classifiers. The basic classifiers consist of trees, functions, and statistical methods. The meta classifiers are used to boost these basic classifiers and improve their performance. This section describes the classifiers used in our experiments.

A. Basic Classifiers

1) J48 decision tree: J48 (or C4.5) is a decision tree based classifier. While constructing the decision tree, the J48 algorithm [13] identifies the attribute that must be used to split the tree further based on the notion of information gain/gini impurity.

2) Reduced error-pruning tree: Commonly known as REPTree [14], 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.

3) Random Forest: The Random Forest [15] 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 ability to handle a very large number of input variables.

4) Alternating decision tree: ADTree [16] is decision tree classifier, which supports only binary classification. It consists of two types of nodes: decision and prediction.

5) Logistic Regression: Logistic Regression [17] is used for prediction of the probability of occurrence of an event by fitting data to a sigmoidal S-shaped logistic curve. Logistic regression is often used with ridge estimators to improve the parameter estimates and to reduce the error made by further predictions.

B. Meta Classifiers

1) Bagging: Bagging [15] is a meta-algorithm to improve the stability of classification and regression algorithms by reducing variance.

2) AdaBoost: AdaBoost [18] is a commonly used ensembling technique for boosting a nominal class classifier. In general, boosting can be used to significantly reduce the error of any weak learning algorithms.

3) Random SubSpace: The Random Subspace classifier [19] 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.

4) 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 [20].

VI. EXPERIMENTS AND RESULTS

In our experiments, we used the WEKA toolkit for data mining [21]. 10-fold cross-validation was used for evaluation. Cross-validation is used to evaluate the prediction performance of data mining models to avoid over-fitting. In k-fold cross-validation, the input data is divided into k random segments. k − 1 segments are used to build the model and the remaining segments are used to evaluate the model. In 10-fold cross-validation this process is repeated 10 times and the final validation result is the average of the 10 repetitions. We used prediction accuracy and area under the ROC curve to evaluate the models in our experiments. The area under the ROC curve is recommended as a performance metric to evaluate different machine learning algorithms [22].

A total of 20 classification schemes were used. 5 basic classifiers, and a combination of the 3 meta classifiers with the 5 basic classifiers as an underlying classifier. We also performed ensemble voting of 3 of the performing classification schemes.

Each of these 20 classification schemes was evaluated for 1 year, 2 years, and 5 years. Figures 3 , 5, and 6 show the percentage accuracy for the classifiers that finished execution plus ensemble voting for 3 of the top classifiers on the dataset of 13 attributes obtained after SMOTE class blanacing. Figures 4, 6, and 8 respectively show the corresponding area under the ROC curve for results of 1 year, 2 years, and 5 years survivability. As described earlier, the original dataset consisted of 134 attributes that was reduced to 65 attributes by removing useless attributes related to the period and cancer related our study. Further attribute selection using CFS and Information Gain yielded a subset of 13 features. Moreover, another dataset of 13
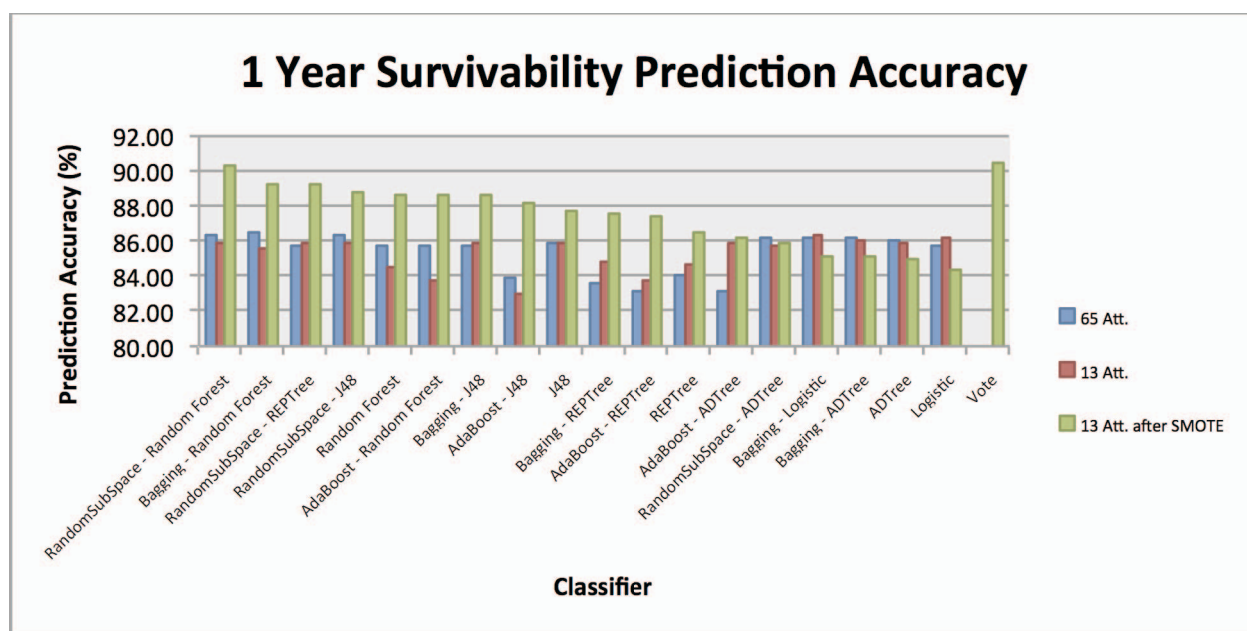
Fig. 3. 1 year survivability percentage accuracy comparison of 65 attributes, 13 attributes and 13 attributes after SMOTE class balancing
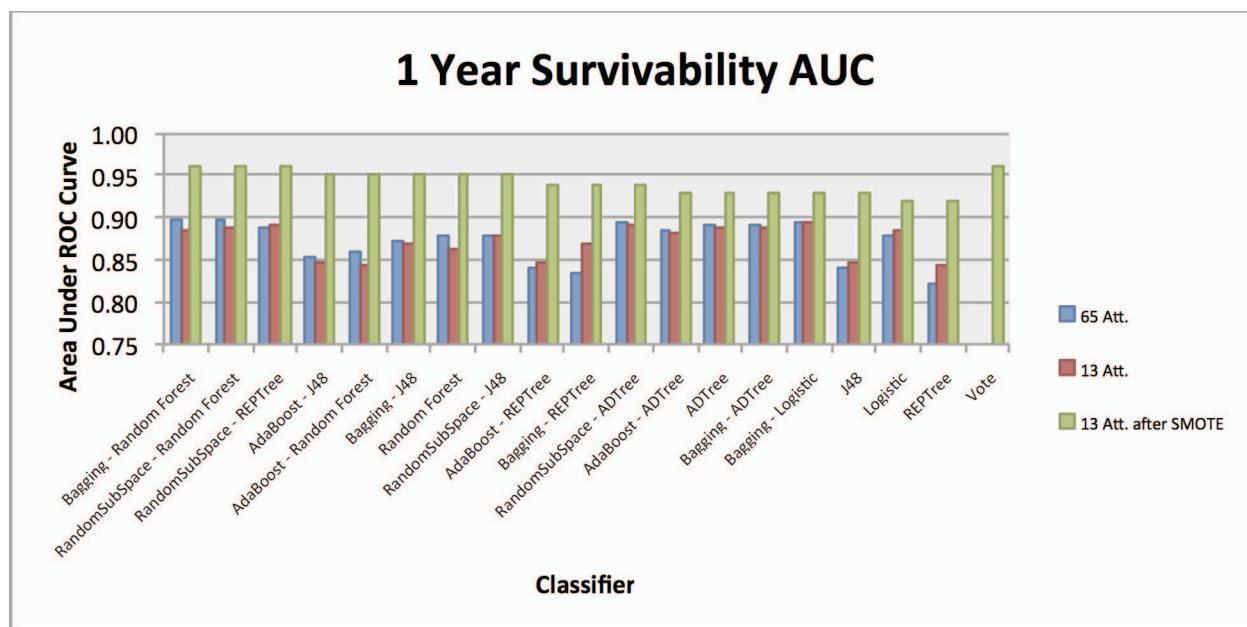

Fig. 4. 1 year survivability area under the curve (AUC) comparison of 65 attributes, 13 attributes and 13 attributes after SMOTE class balancing

features has been generated using SMOTE balancing of survived and non-survived classes.

As evident from the figures, there are many classification schemes that perform well. After combining the top 3 performing classification schemes on the SMOTE balanced dataset and using ensemble voting to combine their predictive powers we noticed that ensemble voting shows the best results in our study.

The ensemble voting model has predictive percentage accuracy of 90.38%, 88.01%, and 85.13% for 1 year, 2 years, and 5 years respectively and an AUC of 0.96, 0.95, and 0.92 for 1 year, 2 years, and 5 years respectively.

VII. Conclusion and Future Work

In this paper, we used different basic and meta classification schemes to construct models for survival prediction for colon cancer patients. Prediction accuracies of 90.38%, 88.01%, and 85.13%and an AUC of 0.96, 0.95, and 0.92 were obtained for the 1-year, 2-year and 5-year colon cancer survival prediction using the ensemble voting classification scheme. We have identified 13 attributes that have approximated the predictive power of 65 attributes. We also demonstrate how balancing the classes in the dataset yields better results if the imbalance is big.

Future work includes exploring more techniques to deal with imbalanced data. Also, we plan to build a colon cancer outcome calculator. An outcome calculator can accurately estimate survivability of a colon cancer patient. Moreover, it can aid doctors in decision-making and provide a better understanding of the risks involved in a particular treatment procedure, based on patient-specific attributes. Further more we also plan to do similar analysis for other cancers.

References

[1] Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB. Cancer incidence in five continents. Lyon: International Agency for Research on Cancer. Vol. VIII. IARC Scient. Publ. No. 155. 2002.

[2] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2009), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012, based on the November 2011 submission.

[3] 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

[4] 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.

[5] 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.

[6] 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.

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

[8] Fathy, Sherif Kassem. "A predication survival model for colorectal cancer." In Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications, pp. 36-42. World Scientific and Engineering Academy and Society (WSEAS), 2011.

[9] Stojadinovic, Alexander, John S. Eberhardt, Elizabeth Ben Ward, Aviram Nissan, Eric K. Johnson, Mladjan Protic, George E. Peoples, Itzhak Avital, and Scott R. Steele. "Clinical Decision Support and Individualized Prediction of Survival in Colon Cancer: Bayesian Belief Network Model." Annals of surgical oncology 20, no. 1 (2013): 161-174.

[10] Wang, Samuel J., Clifton D. Fuller, Rachel Emery, and Charles R. Thomas Jr. "Conditional survival in rectal cancer: a SEER database analysis." Gastrointestinal cancer research: GCR 1, no. 3 (2007): 84.

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

[12] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *arXiv preprint arXiv:1106.1813* (2011).

[13] J. Quinlan. C4. 5: programs for machine learning. Morgan Kaufmann, 1993.

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

[15] L. Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[16] 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.

[17] 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.

[18] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. 1996.

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

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

[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] Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." Pattern recognition 30, no. 7 (1997): 1145-1159.
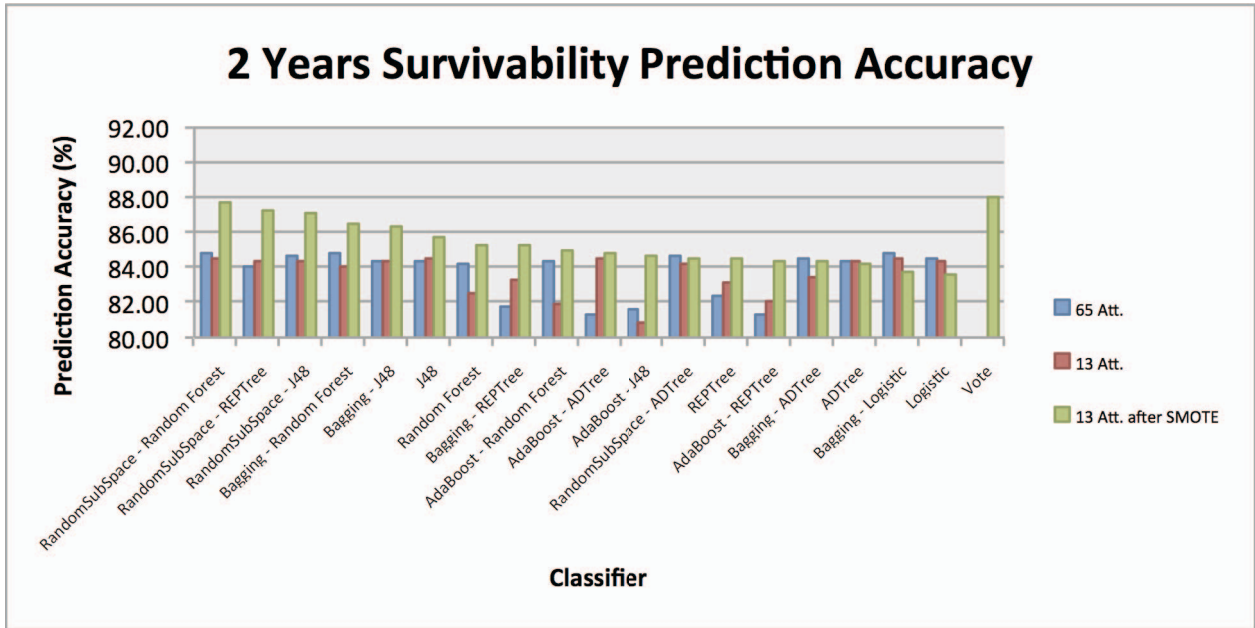
Fig. 5. 2 years survivability percentage accuracy comparison of 65 attributes, 13 attributes and 13 attributes after SMOTE class balancing
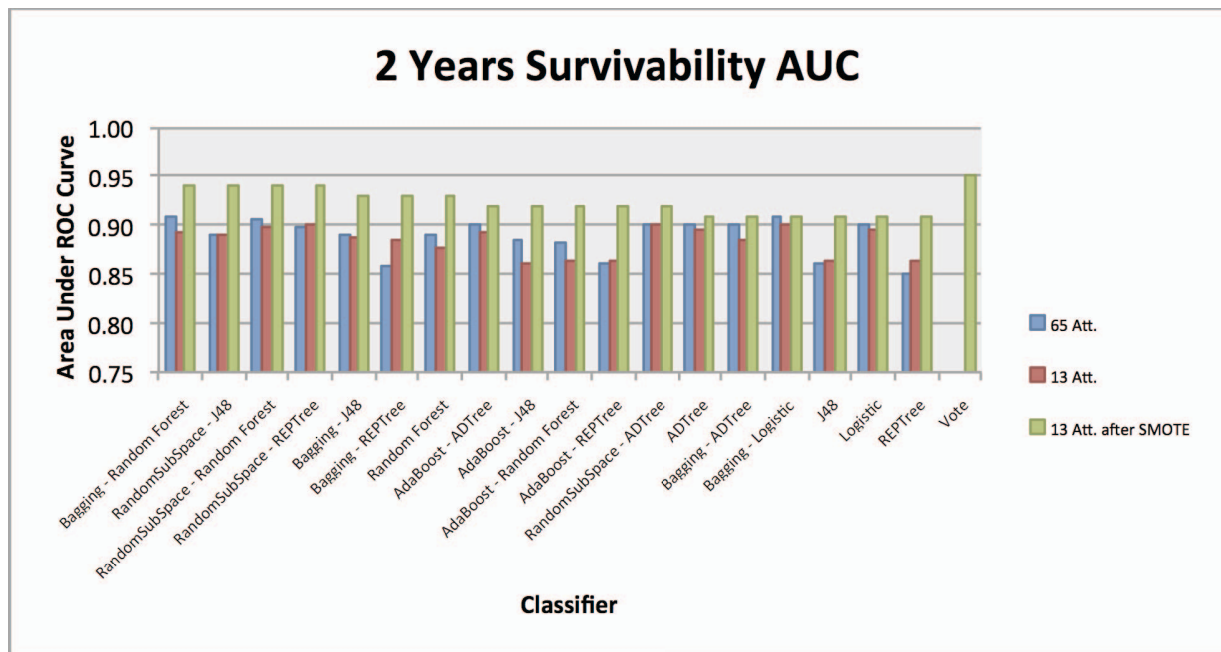

Fig. 6. 2 years survivability area under the curve (AUC) comparison of 65 attributes, 13 attributes and 13 attributes after SMOTE class balancing
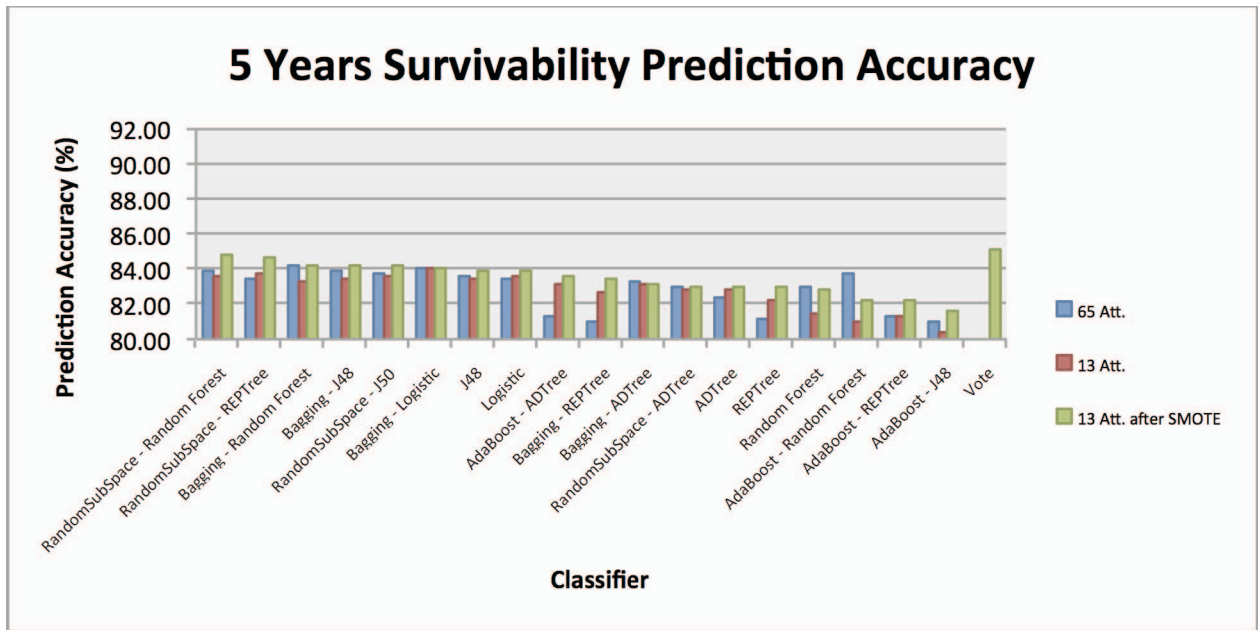
Fig. 7. 5 years survivability percentage accuracy comparison of 65 attributes, 13 attributes and 13 attributes after SMOTE class balancing
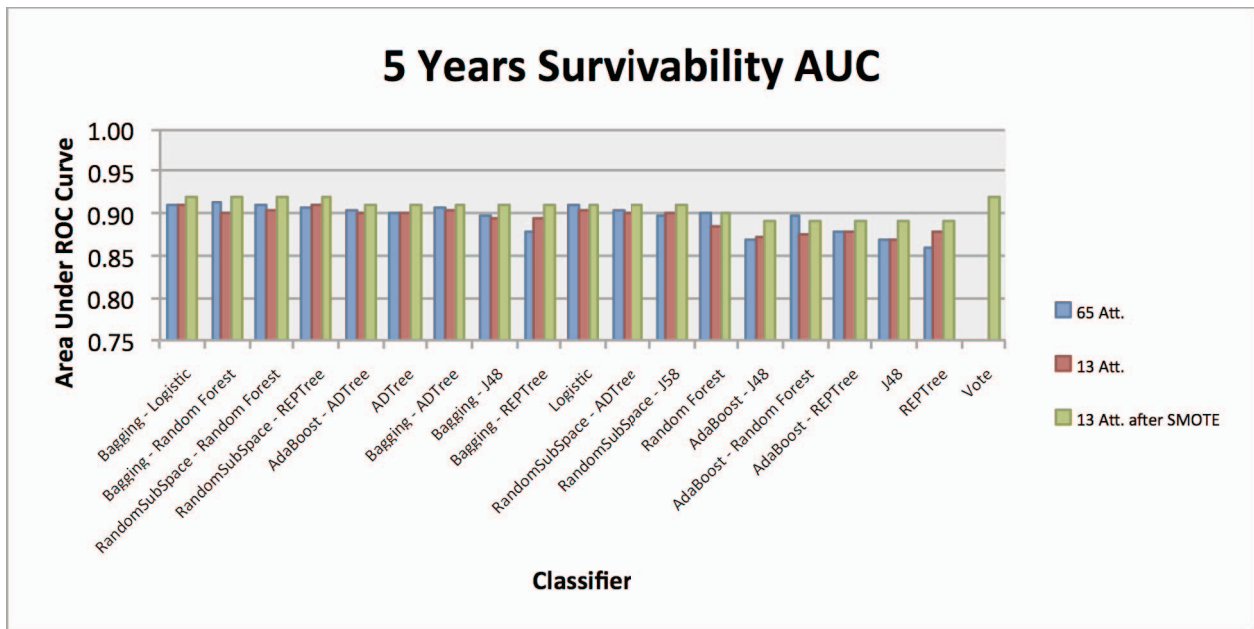

Fig. 8. 5 years survivability area under the curve (AUC) comparison of 65 attributes, 13 attributes and 13 attributes after SMOTE class balancing