Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques

James A. Bartholomai
Dept. of Bioengineering
University of Louisville
Louisville, KY

Hermann B. Frieboes*
Dept. of Bioengineering
University of Louisville
Louisville, KY
hbfrie01@louisville.edu

*Corresponding Author

***Abstract*— A regression model is developed to predict survival time in months for lung cancer patients. It was previously shown that predictive models perform accurately for short survival times of less than 6 months; however, model accuracy is reduced when attempting to predict longer survival times. This study employs an approach for which regression models are used in combination with a classification model to predict survival time. A set of de-identified lung cancer patient data was obtained from the Surveillance, Epidemiology, and End Results (SEER) database. The models use a subset of factors selected by ANOVA. Model accuracy is measured by a confusion matrix for classification and by Root Mean Square Error (RMSE) for regression. Random Forests are used for classification, while general Linear Regression, Gradient Boosted Machines (GBM), and Random Forests are used for regression. The regression results show that RF had the best performance for survival times ≤6 and >24 months (RMSE 10.52 and 20.51, respectively), while GBM performed best for 7-24 months (RMSE 15.65). Comparison plots of the results further indicate that the regression models perform better for shorter survival times than the RMSE values are able to reflect.**

Keywords—lung cancer, machine learning, supervised classification, SEER database, biomedical big data

I. Introduction

Currently, clinicians rely on personal knowledge and experience to estimate survival times of cancer patients. Yet these estimates have been shown to be inaccurate. In [1], physician consultants predicted a median survival time of 25.7 months, physician registrars predicted a median survival time of 21.4 months, and physician residents predicted a median survival time of 21.5 months for patients with an average of 11.7 months of actual survival. Additionally, for patients that were predicted to live longer than 3 months, only about 60% of them actually survived this long. Another study [2] showed that physicians predicted survival time to the month 10% of the time, within 3 months 59% of the time, and within 4 months 71% of the time. It was shown that short-term survival is overestimated and long-term survival is underestimated.

Machine learning is an appropriate application for this problem because the algorithms can quickly learn from a much larger volume of patients than any set of clinical expertise, which may allow the algorithms to yield more accurate predictions. This problem is important because lung cancer is the leading cause of cancer-related deaths [3]. Cancer patient survival models are generally classification models for either survival or survival time. Classification results may not be as useful as regression for this application since the output of time is continuous and not discrete. Thus, a regression based approach was selected for this study. The data set includes patient variables available at or near the time of diagnosis to represent a proactive set of survival predictors. We previously developed a regression model, but it is limited in that it yielded a large RMSE close to the standard deviation of the dataset [4]. This large RMSE is caused by both over-predictions of patients with lower survival times and the inability of the model to predict beyond 28 months due to the imbalance of training data.

Studies have evaluated lung cancer patient survival by analyzing large data sets, such as from the SEER database, with machine learning techniques, including logistic regression and SVM [5], and ensemble clustering-based approaches [6]. Data classification techniques were evaluated in [7] to assess the likelihood of patients with particular symptoms to develop lung cancer. A model was developed in [8] for prediction of survival of non-small cell lung cancer (NSCLC) patients by using artificial neural networks (ANNs). Data was used from the National Cancer Institute (NCI) caArray database. Several types of models were produced with different ANN architectures and ultimately yielded an optimal one with an overall accuracy of 83% by classification. In [9], Naïve-Bayes and Decision Trees were used to predict survival time with lung cancer, achieving an accuracy of 90%. In [10,11], an ensemble of five Decision Trees classification models produced

HBF acknowledges partial support from NIH/NCI R15CA203605.

the best prediction in terms of the area under the Receiver Operating Characteristic (ROC) curve. In [12], an association rule technique was used to create a tree of rules for lung cancer. Association rule mining techniques were used to determine correlation or association relationships among the data. Different standard criteria and techniques to extract the rules were proposed in [13]. The effectiveness of lung cancer treatment was considered in [14]. Other work analyzing the SEER database focusing on lung cancer via statistical and classification techniques have included [15-22] and [23-25], respectively, and has been reviewed in [26]. Recently, unsupervised methods were shown to yield comparable results to supervised techniques [27].

With respect to cancers other than lung, a study by [28] used a classification approach to predict survival at 6, 12, and 24 months. Models were developed using data from the Evaluation of Cancer Outcomes (ECO) registry, from electronic administrative records (EARs), or from both. An ensemble approach with 400 SVM with linear kernel (averaged output) was used for each model. The model using both sets of data performed best for genitourinary, head and neck, lung, skin, and upper gastrointestinal tumors. A study by [29] used artificial neural networks, Decision Trees, SVM, and logistic regression to predict prostate cancer survivability. SVM had the best test set accuracy (92.85%). Prostate cancer survivability was evaluated in a similar study [30] using the same models. Colon cancer patient data from the SEER database was used in [31] to predict patient survival, finding that neural networks were the most accurate. In a study by [32], an ensemble of the three best performing classifiers resulted in the best prediction for colon cancer survival according to the area under the ROC curve. A probabilistic model using Bayesian network (BN) was used to predict the short survivability of patients with brain metastasis from lung cancer [33].

Models have also been developed to predict survival of patients diagnosed with breast cancer, including Support Vector Machine (SVM), ANN, and semi-supervised methods (as reviewed in [26] and associated references). In early work, the concept of agglomerative clustering was applied to assemble groups of cancer patients [34]. A rule-based classification method was used in [35] to predict different types of breast cancer survival, showing that Trees Random Forest (TRF) had better results. Artificial neural networks, Decision Trees, and logistic regression were used in [36] to predict breast cancer survivability with accuracies of 93.6% 89.2%, respectively.

In this study, patients diagnosed with lung cancer during the years 2004-2009 were selected from de-identified information provided in the SEER database in order to predict their survival time. The SEER Program is an authoritative repository of cancer statistics in the United States [37]. Classification modeling was used to broadly separate validation data into three categories of survival time: less than or equal to 6 months, 7-24 months, and greater than 24 months. Each of these categories had its own respective regression model, with the goal to predict a more precise survival time for the validation data. The chosen algorithm for classification was Random Forests, and the chosen algorithms for regression included Random Forests, General Linear Regression, and Gradient Boosted Machines.

II. Methods

A. Data Analysis

The patient data was first analyzed using Minitab. **Table I** highlights the descriptive statistics.

TABLE I. Patient Data Statistics

Survival Max	Survival Min	N	Mean	St. Dev.	Q 1	Median	Q 3
71	0	10442	19.587	16.768	7	14	28

Next, an ANOVA was performed with survival time as the response and with 24 input parameters to select significant model parameters and to reduce dimensionality. Prior to analysis, outlier data (only one instance of a parameter level) were removed. The analysis was performed ensuring that parameters were appropriately coded as either numeric or non-numeric. For the ANOVA, a two-sided 95% confidence interval and adjusted (Type III) sum of squares for tests were used. A box-cox transformation was used with optimal λ in Minitab to improve the linearity of the normal probability plot. In order to perform the box-cox transformation, the survival times of 0 had to be replaced with 0.5. This replacement value was chosen since those data points represent less than 1-month survival. Out of the 24 parameters, 15 were determined to be significant (p-value<0.05). Parameters were not selected if they could not be known or estimated at the time of diagnosis or if they could not be estimated by ANOVA. Two of the significant parameters were not selected because they could not be known at the time of diagnosis due to being classifiers for cause of death. Parameters not selected were *CODtositerecode*, which includes classifiers for all cancer-related and cancer-unrelated causes of death, and *SEERcause.specificdeathclassification*. which includes classifiers that describe whether the cause of death was due to cancer. The 13 parameters selected for model training included: *Radiationsequencewithsurgery*, *Numberofprimaries*, *PrimarySite.labeled*, *HistologicTypeICD.O.3*, *CSlymphnodes.2004..*, *RXSumm..ScopeRegLNSur.2003..*, *RXSumm..SurgPrimSite.1998..*, *DerivedSS1977*, *TumorSizeNumeric*, *AgeNumeric*, *GradeNumeric*, *StageNumeric*, and *TNumeric*. Definitions are described in the SEER database documentation [38].

B. Model Design

The next step was to design the predictive model. Random sampling was used to select 75% of the data for training and the rest for validation. The first step was a classification-based model which separates the validation data into three separate categories: less than or equal to 6 months survival, 7-24 months survival, and greater than 24 months survival. This categorical split was selected to improve regression predictions for low and high survival times. This classification model was trained with all training data. Next, three regression models were made for each category, each using the same algorithms and parameters, but trained with a different portion of the data. The first category used data from 0-12 months, the second category was trained with all training data, and the third category was trained with data greater than or equal to 18 months. A wider range of data was used for training of each regression model to be able to predict data within or slightly outside the intended range. All data was used for training the second category since it was the largest category and contained data across the entire range of the data set.

The classification model used Random Forests (RF) while the regression models used General Linear Regression (GL), RF, Gradient Boosted Machines (GBM), and a custom ensemble. RF were selected for classification and regression since they performed well in the classification models in literature. GL was selected as it is the simplest regression model and performs well with the data set. GBM was selected because it is similar to RF, but the model is built progressively in stages instead of all at once. RF was used over GBM for classification because it performed this task slightly better.

C. Classification, Random Forests

An arbitrary tree count of 1000 was selected; increasing the count beyond this count did not increase accuracy. The "mtry" parameter was 3, which is the constraint for the number of variables randomly sampled as candidates at each split since it improved accuracy over the default (the default value for classification is the square root of parameters). 50 was selected for the node size (minimum size of terminal nodes) to produce smaller trees (the default value for classification is 1) which are less precise but less prone to outlier data. Importance of predictors was assessed. *HistologicTypeICD.O.3* could not be used for training due to the high number of classifiers. A conditional inference random forest model was considered to handle the higher number of classifiers for HistologicTypeICD.O.3; however, it was ultimately rejected due to producing poorer results since a much smaller tree count was necessary for the model to converge. The "randomForest" package was used in R.

D. Regression, Random Forests

For the regression model, 500 trees were used with 10-fold cross validation. Increasing the tree count did not improve model accuracy. 50 was selected for the node size as in the other RF model (the default value for regression is 5). The 10-fold cross validation was also used to select between three possible values for the "mtry" parameter: 3, 5, 7 (the default for regression is one third of the number of parameters) in order to select the best possible parameter value. The "randomForest" package was used in R.

E. Regression, Gradient Boosted Machines

Similar to the RF model, 500 trees were used with 10-fold cross validation. 50 was selected for the node size as in the RF models. 1 was selected for the interaction depth (additive model without interaction). 0.1 was selected for the shrinkage, which is also known as the learning rate. The "gbm" package was used in R.

F. Regression, General Linear Regression

No unique parameters were used for the general linear model. The "glm" training method was used in R.

G. Regression, Custom Ensemble

The custom ensemble is a simple linear model of the other regression models (RF, GBM, and GL) where each model's predictions are weighted differently. The weights are determined by a linear regression with each individual model's predictions of the training data survival times and the actual training data survival times.

III. Results

The classification results in **Table II** indicate an overall accuracy of 50.3% and an error rate of 49.7%. Sensitivity was 31%, 76.4%, and 24.4% for Classes 1, 2, and 3, respectively, while precision was essentially 50% for all classes. This performance reflects the fact that these measures may not be the best to evaluate classification models, especially with unbalanced data. Further, the classification model is over-fitted on class 2, which is the class of 7-24 months with the bulk of the training data (46.8% of the actual data is in class 2).

TABLE II. Confusion Matrix

Table Head		Predicted Class		
		1	*2*	*3*
Actual Class	*1*	198	397	44
	2	149	933	139
	3	33	535	183

Table III shows the basic statistics of the separated validation data sets.

TABLE III. VALIDATION DATA DESCRIPTIVE STATISTICS (MONTHS)

Validation Survival Class	Standard Deviation	Mean Survival
1	9.70	9.58
2	16.22	19.74
3	19.16	28.21

For the regression model (≤6 months), RF showed the best performance (**Table IV**).

TABLE IV. REGRESSION RMSE CLASS 1

Method	RMSE	RMSE / Std. Deviation
General Linear Regression	10.63	1.10
Random Forests	10.52	1.08
GBM	10.65	1.10
Custom Ensemble	10.84	1.12

All the models primarily predicted within the range of roughly 3-7 months, and under-predicted large survival times (**Fig. 1**). In all figures, model predictions are on the x axis and actual survival times are on the y axis.

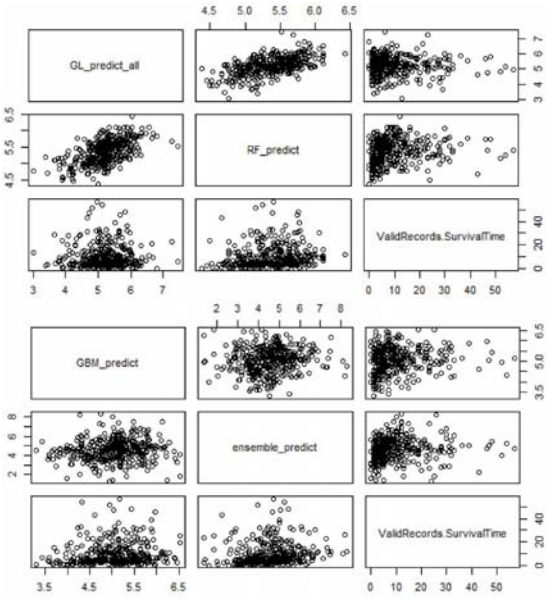


Fig. 1. Comparison plots for Class 1

For 7-24 months, GBM had the best RMSE (**Table V**), although RF performed better (**Fig. 2**).

TABLE V. REGRESSION RMSE CLASS 2

Method	RMSE	RMSE / Std. Deviation
General Linear Regression	15.77	0.97
Random Forests	15.70	0.97
GBM	15.65	0.96
Custom Ensemble	16.26	1.00

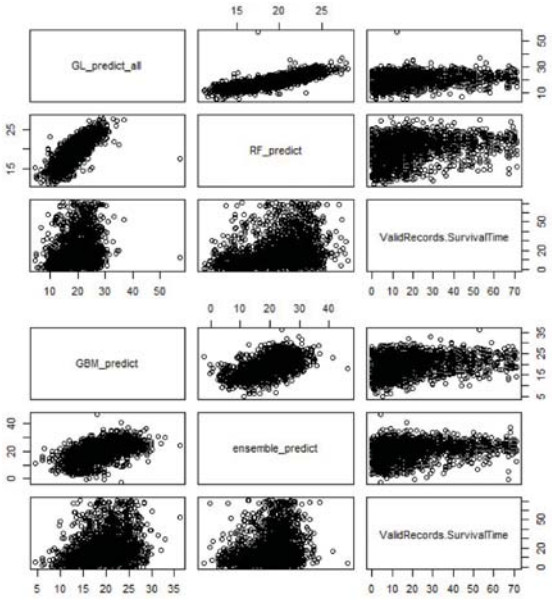
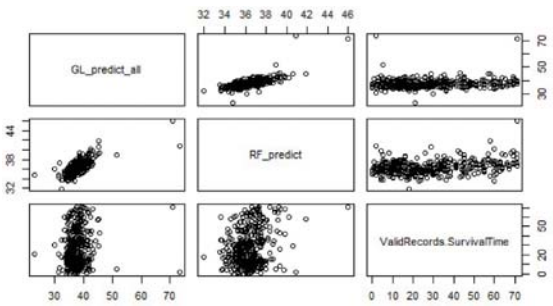


Fig. 2. Comparison plots for Class 2.

For time >24 months, RF had the best RMSE (**Table VI**) even though GBM performed better (**Fig. 3**).

TABLE VI. REGRESSION RMSE CLASS 3

Method	RMSE	RMSE / Std. Deviation
General Linear Regression	21.37	1.12
Random Forests	20.51	1.07
GBM	21.14	1.10
Custom Ensemble	21.18	1.11



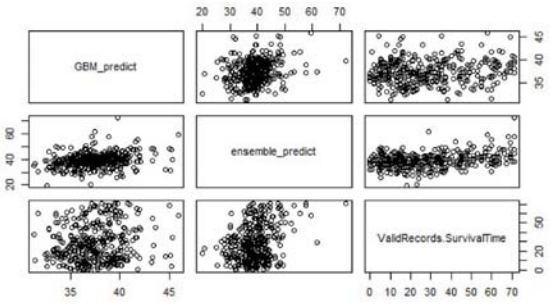


Fig. 3. Comparison plots for Class 3.

IV. Discussion

The goal of this study was to explore the possibility of more accurately predicting lung cancer patient survival by using a classification model along with three regression models. The purpose of the classification stage was to separate the validation data into smaller categories (survival ≤6, 7-24, or >24 months). Then, a regression model specific to each category could attempt to predict a more specific survival time. As expected, for the regression models, the predictions were more accurate when the actual survival time was lower. RMSE was used to measure accuracy because it is a common measurement of accuracy for regression models; however, it is misleading when trying to understand how well the models actually perform. Because of how RMSE is calculated, larger errors affect the value much more than smaller errors. This leads to a much larger apparent RMSE when predicting low survival times, such as in the first category (≤6 months), due to the wide range of data present. The comparison plots show that the regression models perform better for lower survival times, but the RMSE values do not properly represent the difference in performance.

Several different approaches were attempted when initially designing the models to address two specific issues: not being able to predict outside a narrow range and having a large RMSE. The first approach was to oversample data beyond 28 months for training and validating; however, this only allowed the models to predict slightly further. The oversampling approach was not used because it leads to an overall lower RMSE and could lead to further overfitting to the data set. The second approach was to remove data beyond 28 months for training and validating. This led to a much lower RMSE (~7 months), but introduced a similar issue of not being able to predict beyond approximately 15 months. The next approach was to create a random forest classification model using the original survival time classifiers assigned by the SEER database: C0, C1, C2, C3, C4, C5. With these classifiers, each number represents the survival time year threshold exceeded. This classification model would not predict any data points other than C0 or C1 due to the imbalance of the data set.

In the approach that was implemented, the classification model directly affects the regression model accuracy. The classification model could be improved by choosing a different set of classes. The models used in this study are prone to overfitting, so each class should contain approximately the same amount of data. The current design is able to predict a wider range of survival times with varying RMSE values. The most ideal regression models were for the range of ≤6 months, which had both a more linear plot and a lower RMSE. There are likely two reasons for this: there is more data for lower survival times, and lower survival times are more predictable. For our previous design in [4], the GBM model achieved the highest single-model RMSE of 15.32. The best RMSE value achieved in the design outlined in this study was 10.52 by the RF model. Although both RMSE values are inflated by large errors due to the range of the data, the result here improve upon the previous study.

In conclusion, predictive models can perform accurately with lung cancer patient data in the SEER database for short survival times of ≤6 months; however, the accuracy is reduced as the models attempt to predict higher survival times. A primary limiting factor is that the range of data is too wide. This may be a problem because not all patient deaths are due to cancer, especially for longer survival times, Although regression models have potential for more accurately predicting short term survival, more work is necessary to improve and validate such methods for meaningful clinical application.

References

[1] C. Clément-Duchêne, C. Carnin, F. Guillemin, and Y. Martineta, “How Accurate Are Physicians in the Prediction of Patient Survival in Advanced Lung Cancer?” The Oncologist, vol. 15, pp. 782–789, 2010.

[2] M. F. Muers, P. Shevlin, and J. Brown. “Prognosis in lung cancer: physicians' opinions compared with outcome and a predictive model,” Thorax, vol 51, pp. 894-902, 1996.

[3] American Cancer Society, Cancer Facts & Figures 2018.

[4] C. M. Lynch, B. Abdollahi, J. D. Fuqua, et al., “Prediction of lung cancer patient survival via supervised machine learning classification techniques,” Int. J. Med. Inform., vol. 108, pp. 1-8, 2017.

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

[6] D. Chen, K. Xing, D. Henson, L. Sheng, A. M. Schwartz, and X. Cheng, “Developing prognostic systems of cancer patients by ensemble clustering,” J. Biomed. Biotechnol., , vol. 2009, Article ID 632786, 7 pages, 2009.

[7] V. Krishnaiah, G. Narsimha, and N. S. Chandra, “Diagnosis of lung cancer prediction system using data mining classification techniques,” International Journal of Computer Science and Information Technologies, vol. 4, pp. 39-45, 2013.

[8] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu, “Risk classification of cancer survival using ANN with gene expression data from multiple laboratories.” Comput. Biol. Med., vol. 48, pp. 1-7, 2014.

[9] G. Dimitoglu, J. A. Adams, and C. M. Ji, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability," Journal of Computing, vol. 4, pp. 1-9, 2012.

[10] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, editors. "A lung cancer outcome calculator using ensemble data mining on SEER data," Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics, ACM 2011.

[11] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on SEER data," Scientific Programming vol. 20, pp. 29-42, 2012.

[12] A. Agrawal and A. Choudhary, editors. "Identifying hotspots in lung cancer data using association rule mining," 11th International Conference on Data Mining Workshops (ICDMW), IEEE 2011.

[13] R. Agrawal and T. Imieliński, and A. Swami, editors, "Mining association rules between sets of items in large databases.,"SIGMOD Record, ACM, 1993.

[14] Y. Wu, "Propensity score analysis to compare effects of radiation and surgery on survival time of lung cancer patients from National Cancer Registry (SEER)," [Master's], Epidemiology and Biostatistics: School of Public Health, SUNY-Albany; 2006.

[15] S. Ramalingam, K. Pawlish, S. Gadgeel, R. Demers, and G. Kalemkerian, "Lung cancer in young patients: analysis of a Surveillance, Epidemiology, and End Results database," Journal of Clinical Oncology, vol. 16, pp. 651-657, 1998.

[16] T. K. Owonikoko, C. C. Ragin, C. P. Belani, .et al., "Lung cancer in elderly patients: an analysis of the Surveillance, Epidemiology, and End Results database," Journal of Clinical Oncology, vol. 25, pp. 5570-5577, 2007.

[17] A. Bhaskarla, P. C. Tang, T. Mashtare, et al., "Analysis of second primary lung cancers in the SEER database," Journal of Surgical Research, vol. 162, pp. 1-6, 2010.

[18] M. J. Hayat, N. Howlader, M. E. Reichman, and B. K. Edwards. "Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program," The Oncologist, vol. 12, pp. 20-37, 2007.

[19] M. J. Thun, L. M. Hannan, L. L. Adams-Campbell, et al.. "Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies," PLOS Medicine, vol. 5, e185, 2008.

[20] J. B. Fu, T. Y. Kau, R. K. Severson, and G. P. Kalemkerian, "Lung Cancer in Women: Analysis of the National Surveillance, Epidemiology, and End Results Database," CHEST Journal, vol. 127, pp. 768-77, 2005.

[21] X. Wu, V. Chen, J. Martin, et al., "Comparative Analysis of Incidence Rates Subcommittee, Data Evaluation and Publication Committee, North American Association of Central Cancer Registries. Subsite-specific colorectal cancer incidence rates and stage distributions among Asians and Pacific Islanders in the United States, 1995 to 1999," Cancer Epidemiol. Biomarkers Prev., vol. 13, pp. 1215-1222, 2004.

[22] S. J. Wang, C. D. Fuller, R. Emery R, and C. R. Thomas Jr, "Conditional survival in rectal cancer: a SEER database analysis," Gastrointestinal Cancer Research: GCR, vol. 1, p. 84, 2007.

[23] I. Skrypnyk, editor, "Finding Survival Groups in SEER Lung Cancer Data. Machine Learning and Applications (ICMLA), 2012 11th International Conference on Machine Learning and Applications.

[24] A. Agrawal, and A. Choudhary. "Association Rule Mining Based HotSpot Analysis on SEER Lung Cancer Data," International Journal of Knowledge Discovery in Bioinformatics (IJKDB), vol. 2, pp. 34-54, 2011.

[25] N. Kapadia, F. Vigneau, W. Quarshie, A. Schwartz, and F. Kong, "Patterns of Practice and Outcomes for Stage I Non-Small Cell Lung Cancer (NSCLC): Analysis of SEER-17 Data, 1999-2008," International Journal of Radiation Oncology* Biology* Physics, vol. 84, p. S545, 2012.

[26] K. Korou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8-17, 2015.

[27] C. M. Lynch, V. H. van Berkel, and H. B. Frieboes, "Application of unsupervised analysis techniques to lung cancer patient data," PLoS ONE, vol. 12,. e0184370, 2017.

[28] S.Gupta, T.Tran, W.Luo, et al., "Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry," BMJ Open, vol. 4, e004007, 2014.

[29] D. Delen. "Analysis of cancer data: a data mining approach," Expert Systems, vol. 26, pp. 100-112, 2009.

[30] D. Delen and N. Pat, editors. "Knowledge Extraction from Prostate Cancer Data," 39th Hawaii International Conference on System Sciences, Hawaii 2006.

[31] N. A. Noohi, M. Ahmadzadeh, and M. Fardaer, "Medical Data Mining and Predictive Model for Colon Cancer Survivability," International Journal of Innovative Research in Engineering & Science, vol. 2, 2013.

[32] R. Al-Bahrani, A. Agrawal, and A. Choudhary, editors. "Colon cancer survival prediction using ensemble data mining on SEER data," IEEE Big Data Workshop on Bioinformatics and Health Informatics, 2013.

[33] B. Makond,, K.Wang, and K.Wang. "Probabilistic modeling of short survivability in patients with brain metastasis from lung cancer," Computer Methods and Programs in Biomedicine, vol. 119, pp. 142-162, 2015.

[34] K. Xing, D. Chen, D. Henson, and L. Sheng, editors, "A clustering-based approach to predict outcome in cancer patients," Sixth International Conference on Machine Learning and Applications (ICMLA), IEEE 2007.

[35] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," Technology and Health Care, vol. 24, pp. 31-42, 2016

[36] D. Delen, G. Walker, and A. Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods," Artif, Intell, Med., vol. 34, pp. 113-127, 2005.

[37] SEER_Program. 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.

[38] SEER Research Data Record Description. Retrieved July 16, 2017.