2009 International Conference on Machine Learning and Applications

Application of Machine Learning Techniques for Prediction of Radiation Pneumonitis in Lung Cancer Patients

Jung Hun Oh, Rawan Al-Lozi, and Issam El Naqa

Division of Bioinformatics and Outcomes Research
Department of Radiation Oncology
Washington University School of Medicine
MO 63110, USA

Email:{joh, ielnaqa}@radonc.wustl.edu

Abstract

Lung cancer patients who receive radiotherapy as part of their treatment are at risk radiation-induced lung injury known as radiation pneumonitis (RP). RP is a potentially fatal side effect to treatment. Hence, new methods are needed to guide physicians to prescribe targeted therapy dosage to patients at high risk of RP. Several predictive models based on traditional statistical methods and machine learning techniques have been reported, however, no guidance to variation in performance has not been provided to date. Therefore, in this study, we compare several widely used classification algorithms in the machine learning field are used to distinguish between different risk groups of RP. The performance of these classification algorithms is evaluated in conjunction with several feature selection strategy and the impact of the feature selection on performance is further evaluated.

1 Introduction

Lung cancer is one of the most lethal diseases in both men and women in the world, resulting in a leading cause of cancer death with a low five-year survival rate of 15% [1]. About 8 to 9 out of 10 cases of all lung cancers are classified as non-small cell lung cancer (NSCLC). About 50% of lung cancer patients receive radiotherapy in addition to or instead of surgery and it is the main treatment for patients with advanced and inoperable stages. One of the potentially fatal side effects of radiotherapy in lung cancer is RP that results from a dose-limiting toxicity to surrounding normal tissues [2], [3], [4]. Thus, the optimization of dose distribution is very important for providing tumor tissues with sufficient doses while sparing normal tissues from excessive radiation. Highly advanced 3D treatment planning systems in conjunction with accurate estimates of tumor local control probability and complication risk to surrounding normal tissues has allowed for not only improvements of tumor localization and dose distribution but also individualized and patient-specific treatment planning decisions [5]. To identify factors associated with RP, in this study we use two machine learning approaches: feature selection and classification.

Feature selection strategies designed with different evaluation criteria are mainly divided into two categories: the filter approach and the wrapper approach. The criteria used by these approaches include distance measures [6], [7], dependency measures [8], consistency measures [9], [10], and information measures [11], [12]. The filter method selects relevant feature subsets based upon characteristics of the data without involving any classification algorithm. In contrast, the wrapper method employs a predetermined classification algorithm to evaluate the quality of features. Although it requires intensive computations, the wrapper method generally outperforms the filter method. In order to use advantages of both the filter and wrapper methods, hybrid approaches have been also proposed. These methods not only improve the performance but speed up the feature selection task.

Classification is the problem of assigning a sample to a predefined class based on conditional features. Many common classification techniques, including decision tree, neural networks, support vector machine (SVM), k-nearest neighbor (kNN), and Bayesian classifier, have been proposed in a variety of applications. SVM proposed by Vapnik and his colleagues is a novel approach for solving classification problems. It is based on the structural risk minimiza-

978-0-7695-3926-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ICMLA.2009.118

tion principle to minimize an upper bound of the generalization error [13], [14].

The remainder of this paper is organized as follows. In Sections 2 and 3, we introduce the feature selection and classification algorithms performed in this study. Experimental results with dose-volume data in lung cancer are shown in Section 4. Finally, we conclude our work in Section 5.

2 Feature Selection Techniques

2.1 SVM-Recursive Feature Elimination (SVM-RFE)

SVM-RFE, proposed by Guyon *et al.*, is a sequential backward feature elimination method based on SVM [15]. In SVM-RFE, features are ranked in a way that the least important feature is removed after iteratively training a SVM classifier with existing features. To determine the feature to be eliminated at each iteration, Eq. (8) is obtained and the feature with the smallest w_i^2 value in the weight vector is removed.

2.2 Correlation based Feature Selection

A correlation based feature selection method measures correlation between features and tries to find the best feature subset by using a heuristic search strategy in a manner of the forward best first search. The fundamental idea behind the method is that good features are highly correlated with the class but uncorrelated with each other.

2.3 Chi-square Feature Selection

A chi-square feature selection method is a simple algorithm based on the χ^2 statistic to discretize features repeatedly until some inconsistencies are found in the data. As a result of discretization, the feature selection is completed.

2.4 Information Gain based Feature Selection

An information gain based feature selection is an algorithm based on information theory for feature selection in multi-class problems. Let S be the set of instances from k classes, $i.e.$, $c_1, c_2, \ldots, c_k$. The entropy of the class distribution in S is defined as follows:

$$I(S) = -\sum_{i=1}^{k} \frac{|c_i|}{|S|} \log \frac{|c_i|}{|S|}. \tag{1}$$

Then, the information gain of instance set S based on attribute F_i is calculated as

$$\begin{aligned} Gain(F_i) &= I(S) - I(S|F_i), & (2) \\ &= I(S) - \sum_{j=1}^{t} \frac{|S_j|}{|S|} \times I(S_j) \end{aligned}$$

where t is the set of all the possible values of feature F_i. The information gain reflects the reduction in uncertainty about the overall class entropy when a certain feature F_i is given. In other words, features with zero information gain indicate the inability to reduce such uncertainty and should be removed [16].

3 Classification Methods

3.1 SVM

SVM is a supervised learning algorithm originally designed to solve two-class classification problems [17], [18], [19]. The basic idea behind SVM is to find an optimal hyperplane for which a given training data are well separated. It is achieved by maximizing the margin between the two classes after mapping the training data $\mathbf{x}$ into a higher dimensional space via a mapping function $\Phi(\mathbf{x})$. As a result, a decision function is as follows:

$$f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b, \tag{3}$$

where $\mathbf{w}$ is a weight vector and b is a scalar.

Suppose that there are n training samples $\{(\mathbf{x}_i, y_i), 1 \leq i \leq n\}$ where $\mathbf{x}_i$ is the i^{th} training sample consisting of an m-dimensional feature vector and $y_i \in$ {-1, 1} is the class label of $\mathbf{x}_i$. The problem of finding the optimal hyperplane can be formulated as the following optimization problem

$$\min_{\mathbf{w},\zeta_i} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \zeta_i, \tag{4}$$

subject to

$$y_i f(\mathbf{x}_i) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, \tag{5}$$

where ζ_i is a slack variable and C is a user defined soft-margin constant which regularizes the trade-off between training error and margin maximization. This optimization problem can be solved in its *Wolfe dual form* with respect to Lagrange multipliers and can be reduced to a quadratic programming problem:

$$\min_{\alpha_i} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i, \tag{6}$$

subject to

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{7}$$

Here, we can compute the weight vector as $\mathbf{w}$:

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i^* y_i \Phi(\mathbf{x}_i), \tag{8}$$

where α_i^* is Lagrange multipliers and l is the number of support vectors. In Eq. (6), $\Phi(\mathbf{x}_i)^{\mathrm{T}}\Phi(\mathbf{x}_j)$ is substituted with a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ by the kernel trick. Note that for the linear case $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$. Two typical kernels are polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (e + \mathbf{x}_i \cdot \mathbf{x}_j)^d$ and radial basis function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2\sigma^2}||\mathbf{x}_i - \mathbf{x}_j||^2)$ where e, d and σ are adjustable kernel function parameters [20].

3.2 Decision Tree

A decision tree classifier has a hierarchical structure in which the data set is recursively partitioned until each partition consists entirely or almost entirely of samples from one class. In the tree, leaf nodes represent classes and non-leaf nodes indicate selected decision rules. Starting at the root node, one sample is evaluated by the decision rule. It keeps moving down the tree branch until it reaches a leaf node. We used J48 that is implemented as a decision tree classifier in WEKA software package (http://www.cs.waikato.ac.nz/ml/weka/) [21].

3.3 Random Forest

A random forest classifier is an ensemble of classification trees grown on bootstrap samples of the training data in conjunction with a random feature selection in the tree induction process. Given a new input, each tree casts a vote and the class having the most votes is chosen.

3.4 Naive Bayes

In a naive Bayes classifier, it is assumed that all features are mutually independent given a class label, namely that each feature has the class variable as its parent. In practice, despite its simplified assumption, the naive Bayes classifier has often shown good performance compared to sophisticated classification methods in a variety of applications.

4 Experimental Results

4.1 The Data Set

In this study, we analyzed an NSCLC dataset that consists of information obtained from 209 patients at Washington University School of Medicine, who had received radiotherapy with median doses around 70 Gy as part of their treatment. The dose distribution was recalculated using Monte Carlo methods (MC). The number of patients diagnosed with RP was 48 patients called the disease group. The remaining 161 patients belong to the control group. The data obtained from each patient is composed of clinical features (age, gender, race, chemo, stage, smoke, treatment, etc.), dosimetric features such as V_x (volume getting at least x Gy) and D_x (minimum dose to the hottest $x\%$ volume), and relative location of the tumor within the lung or nearby heart.

Table 1. Top ranked 10 features for each feature selection strategy. For CFS, only 5 features were found by its criterion.

Ranking \ Methods	IG	Chi-square	SVM-RFE	CFS
1	Followup	Followup	D10_heartMC	COMSI
2	COMSI	MOH10_heartMC	V40_heartMC	PerformanceStatus
3	V55_heartMC	MOH5_heartMC	V5_heartMC	Followup
4	MOH5_heartMC	V55_heartMC	DCOMSI_heart	D20_lungMC
5	MOH10_heartMC	D5_heartMC	D55_lungMC	MOH10_heartMC
6	D10_heartMC	COMSI	maxDose	
7	D35_lungMC	D10_heartMC	PerformanceStatus	
8	D5_heartMC	MOH20_heartMC	V30_heartMC	
9	MOH20_heartMC	D35_lungMC	TimeAxis	
10	MOH15_heartMC	V65_heartMC	D15_heartMC	

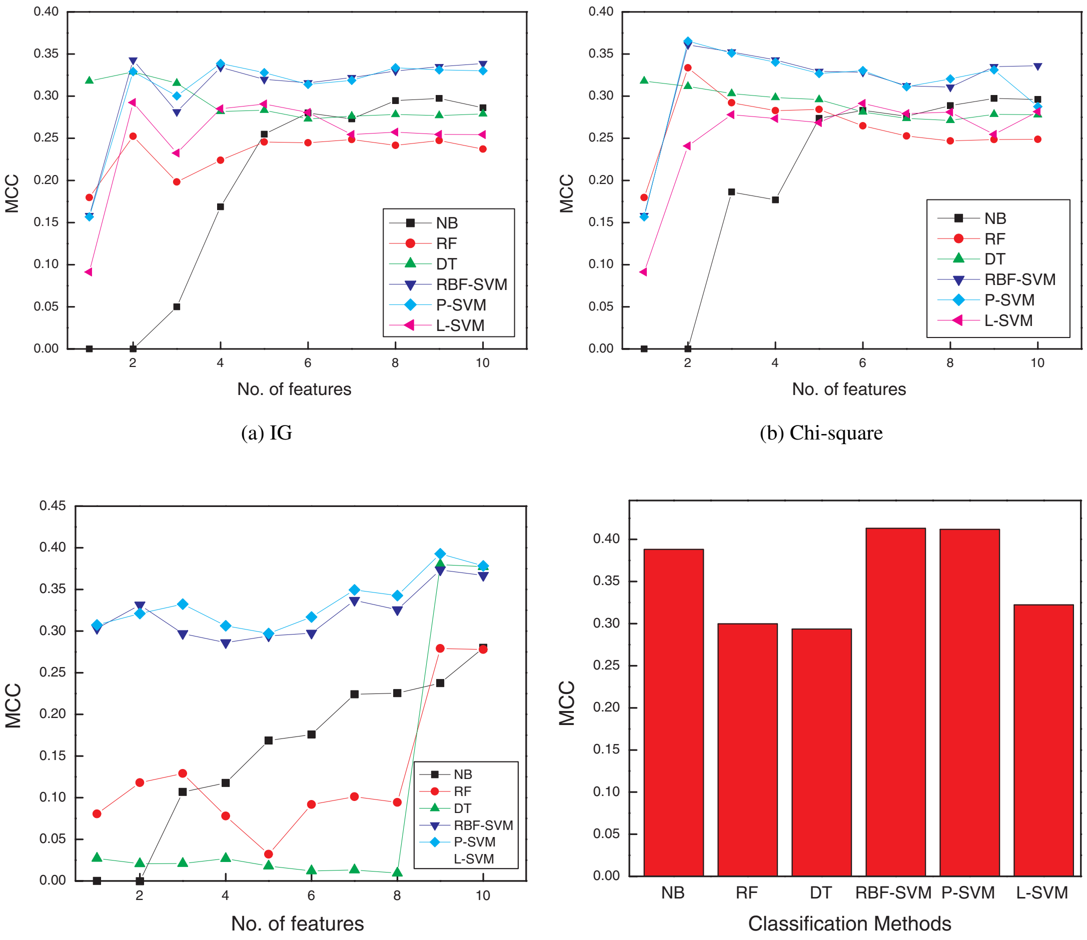
4.2 Machine Learning Methods

For analysis of the dataset, a variety of machine learning methods for feature selection and classification were tested. For feature selection, information gain (IG) based feature selection, chi-square feature selection, correlation based feature selection (CFS), and SVM-RFE were used. For classification, random forest (RF), naive Bayes (NB), decision tree (DT), and SVM were employed. In SVM, the experiments were carried out changing parameters. The parameter values used in this study are as follows: σ in radial basis function SVM (RBF-SVM) varies in $\{0.5, 1, 2, 3, 4, 5\}$; degree d and coefficient e in polynomial SVM (P-SVM) vary in $\{1, 2, 3, 4\}$ and $\{0, 1\}$, respectively; for C, $\{1, 10, 100\}$ are set. By combining these parameters, 18 RBF-SVMs, 24 P-SVMs, and 3 linear SVMs (L-SVMs) are formed. Since the dataset is imbalanced in size, in SVMs weighting values of 3 and 1 were placed into the disease group and control group, respectively.

4.3 Performance Metric

All our experiments were performed using the WEKA software package. For unbiased performance estimate, all measurements were averaged after 30 iterations of 10-fold cross-validation (CV) for each classification algorithm.

In the analysis of an imbalanced dataset, Matthew's correlation coefficient (MCC) is widely used as a performance



(a) IG

(b) Chi-square

(c) SVM-RFE

(d) CFS

Figure 1. Comparison of MCC for four feature selection algorithms.

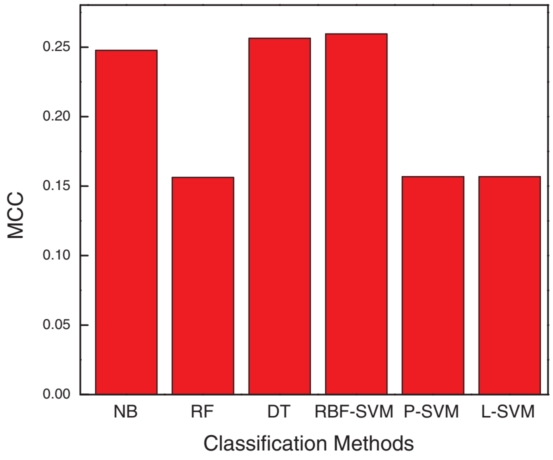


Figure 2. Comparison of MCC when all features were used.

evaluation metric. MCC is calculated as follows:

$$r = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{9}$$

where TP and TN are the number of patients correctly classified in the disease and control group, and FN and FP are the number of patients falsely classified in the disease and control group, respectively. r takes a real value in [-1.0, 1.0]. A coefficient of +1 means a perfect classification. In contrast, -1 represents a perfect inverse prediction. A coefficient of zero indicates an average random prediction.

4.4 Feature Selection and Classification

Table 1 displays the top ten features selected by three feature selection algorithms. For CFS, only five features were chosen by its criterion. It is worthy to note that some features were commonly found in different feature selection methods. For example, $'\mathrm{Followup}','\mathrm{COMSI}'$ (center-of-mass of tumor location in the superior inferior direction), and $'\mathrm{MOH10\_heartMC}'$ were selected in IG, Chi-square, and CFS. $'\mathrm{D10\_heartMC}'$ was found in IG, Chi-square, and SVM-RFE.

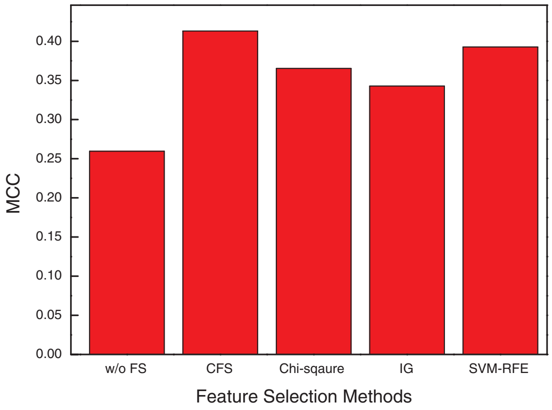
Figure 1 shows the performance in classification algorithms for four different feature selection strategies with the top one, then the top two, and so forth up to the top ten features. Note that Figure 1-(d) illustrates the results obtained using all five features that were searched by CFS. Interestingly, in all cases RBF-SVM and P-SVM achieved the best MCC on this dataset. In particular, with features found by SVM-RFE, the performance of RBF-SVM and P-SVM outperformed considerably other methods. As shown in the figure, the best MCC was obtained when CFS was exploited with RBF-SVM and P-SVM, resulting in 0.413 and 0.412, respectively. Figure 2 shows the MCC values when all features were used without feature selection. As can be seen in the figure, in all cases MCC values were lower than when only a few important features were utilized. It justifies the importance of feature selection in classification algorithms. Figure 3 displays the maximum MCC values across all classification algorithms for each feature selection method. The first bar in the figure represents the MCC value when all features were used. The highest MCC value (0.413) was achieved when RBF-SVM with $C = 100$, $\sigma = 2$, and the five features in conjunction with CFS were employed. Also, accuracy, sensitivity, specificity, and AUC (area under the ROC curve) obtained with these parameters were 74.74%, 69.57%, 76.27%, and 0.729, respectively.

5 Conclusion

We have demonstrated machine-learning application in finding significant features with risk of RP patients. In our classification experiments with the selected features, as expected kernel SVMs showed greatly higher MCC values than not only linear SVM but also other competing classification algorithms after correction for imbalance. In our future work, we will aim to develop more sophisticated kernel functions and include other variables such as biological markers [1]. It is our expectation that this will shed more light on a better understanding of underlying mechanisms in RP onset and advance the individualization of radiotherapy in NSCLC patients.

Acknowledgments

This work was supported in part by NIH grant 5K25CA128809-02. The authors would like also to thank Dr. Jeffrey D Bradley, Dr. Joseph O Deasy, and Dr. Ellen Huang for their assistance in the study.



(a) Maximum MCC for each feature selection method

Feature Selection \ Parameters	Max-MCC	Method	No. of features	C	σ	d	e
w/o FS	0.260	RBF-SVM	160	1	5		
CFS	0.413	RBF-SVM	5	100	2		
Chi-sqaure	0.365	P-SVM	2	100		3	1
IG	0.343	RBF-SVM	2	10	2		
SVM-RFE	0.393	P-SVM	9	100		3	1

(b) Parameter values used for the maximum MCC

Figure 3. Comparison of the maximum MCC across all classification algorithms for each feature selection method.

References

[1] S. Spencer, D. Bonnin, J. Deasy, J. Bradley, and I. El-Naqa, "Bioinformatics methods for learning radiation-induced lung inflammation from heterogeneous retrospective and prospective data," *J. of Biomedicine and Biotechnology*, 2009.

[2] I. El-Naqa, J. Bradley, A. Blanco, P. Lindsay, M. Vicic, A. Hope, and J. Deasy, "Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 64, no. 4, pp. 1275–1286, 2006.

[3] I. El-Naqa, G. Suneja, P. Lindsay, A. Hope, J. Alaly, M. Vicic, J. Bradley, A. Apte, and J. Deasy, "Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships," *Physics in Medicine and Biology*, vol. 51, no. 22, pp. 5719–5735, 2006.

[4] J. Deasy, A. Niemierko, D. Herbert, D. Yan, A. Jackson, R. T. Haken, M. Langer, S. Sapareto, and AAPM/NIH, "Methodological issues in radiation dose-volume outcome analyses: summary of a joint

aapm/nih workshop," *Medical Physics*, vol. 29, no. 9, pp. 2109–2127, 2002.

[5] A. Hope, P. Lindsay, I. El-Naqa, J. Alaly, M. Vicic, J. Bradley, and J. Deasy, "Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 65, no. 1, pp. 112–124, 2006.

[6] J. Bins and B. Draper, "Feature selection from huge feature sets," in *Proc. Int. Conference Computer Vision*, 2001, pp. 159–165.

[7] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognition*, vol. 35, no. 4, pp. 835–846, 2002.

[8] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[9] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial Intelligence*, vol. 151, pp. 155–176, 2003.

[10] G. Lashkia and L. Anthony, "Relevant, irredundant feature selection and noisy example elimination," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 34, no. 2, pp. 888–897, 2004.

[11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[12] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Networks*, vol. 3, no. 1, pp. 143–159, 2002.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[14] J.-T. Jeng, "Hybrid approach of selecting hyperparameters of support vector machine for regression," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 36, no. 3, pp. 699–709, 2006.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[16] J. Oh, Y. Kim, P. Gurnani, K. Rosenblatt, and J. Gao, "Biomarker selection and sample prediction for multicategory disease on maldi-tof data," *Bioinformatics*, vol. 24, pp. 1812–1818, 2008.

[17] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[18] I. El-Naqa, Y. Yang, M. Wernick, N. Galatsanos, and R. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Medical Imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.

[19] J. Oh, A. Nandi, P. Gurnani, L. Knowles, J. Schorge, K. Rosenblatt, and J. Gao, "Proteomic biomarker identification for diagnosis of early relapse in ovarian cancer," *J.of Bioinformatics and Computational Biology*, vol. 4, no. 6, pp. 1159–1179, 2006.

[20] I. El-Naqa, J. Bradley, and J. Deasy, "Nonlinear kernel-based approaches for predicting normal tissue toxicities," in *Proc. of 7th Int. Conference on Machine Learning and Applications*, 2008, pp. 539–544.

[21] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.