

Towards Prediction of Radiation Pneumonitis Arising from Lung Cancer Patients Using Machine Learning Approaches

Jung Hun Oh, Aditya Apte, Rawan Al-Lozi, Jeffrey Bradley, Issam El Naqa*
Division of Bioinformatics and Outcomes Research, Department of Radiation Oncology,
Washington University School of Medicine, MO 63110, USA

ABSTRACT

Purpose: Radiation pneumonitis (RP) is a potentially fatal side effect arising in lung cancer patients who receive radiotherapy as part of their treatment. For the modeling of RP outcomes data, several predictive models based on traditional statistical methods and machine learning techniques have been reported. However, no guidance to variation in performance has been provided to date.

Materials and methods: In this study, we explore several machine learning algorithms for classification of RP data. The performance of these classification algorithms is investigated in conjunction with several feature selection strategies and the impact of the feature selection strategy on performance is further evaluated. The extracted features include patient's demographic, clinical and pathological variables, treatment techniques, and dose-volume metrics. In conjunction, we have been developing an in-house Matlab-based open source software tool, called dose-response explorer system (DREES), customized for modeling and exploring dose response in radiation oncology. This software has been upgraded with a popular classification algorithm called support vector machine (SVM), which seems to provide improved performance in our exploration analysis and has strong potential to strengthen the ability of radiotherapy modelers in analyzing radiotherapy outcomes data. These tools are demonstrated on an institutional non-small cell lung carcinoma (NSCLC) dataset of patients who received radiotherapy.

Results: Our methods were applied to an NSCLC dataset that consists of 209 patients' information, each having 160 variables. Using several feature selection methods, relevant features were searched. Subsequently, with the selected features, various classification algorithms were tested. Through these experiments, we showed the usefulness of machine learning methods in the analysis of radiation oncology dataset.

Conclusions: We have presented an open-source software tool and several machine learning algorithms for analyzing radiotherapy outcomes. We demonstrated the tool on a lung cancer patient dataset. We believe that the improved tool will provide radiation oncology modelers with new means to analyze radiation response data.

Keywords: radiation pneumonitis, machine learning, informatics, DREES

Disclosure: The authors declare no conflicts of interest.

1. INTRODUCTION

Lung cancer is a leading cause of cancer-related death in both men and women in the world with a low five-year survival rate of 15% (American Cancer Society 2008). Two main types of lung cancer are small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). Approximately, 80% of lung cancer cases are classified as 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 (American Cancer Society 2008). One of the potentially fatal side effects of radiotherapy in lung cancer is radiation-induced lung injury known as radiation pneumonitis (RP) that results from over-dosage of surrounding normal tissues (Deasy et al. 2002; El Naqa et al. 2006a, 2006b; Spencer et al. 2009). Thus, the optimization of treatment planning dose distributions is crucial for providing tumor tissues with sufficient doses while sparing normal tissues from excessive radiation effects. Recent advances in radiotherapy and biotechnology such as highly advanced 3D treatment planning systems provide new opportunities to precisely estimate tumor local control probability and complication risk to surrounding normal tissues, which allows for not only improvements of tumor localization and dose distribution but also individualized and patient-specific treatment planning decisions (Hope et al. 2006). Nevertheless, the lack of dedicated informatics tools for extracting and analyzing metrics that could be related to

*elnaqa@wustl.edu; phone +1 314 362 0129; fax +1 314 362 8521; radonc.wustl.edu

radiotherapy outcomes such as RP poses challenges for prediction of such effects and customization of treatment plan design based on expected risk. We demonstrate the development of such tools in this study by comparing different machine learning strategies for identification of factors that could be associated with RP. This process is composed of three steps: (1) selecting relevant variables, (2) building appropriate classifiers based on supervised learning, and (3) presenting robust tools to the radiation oncology community through our open-source software.

Proper feature selection is a major challenge in machine learning and is posed as the ability to select a subset of features that will represent the dataset or distinguish one patients' group from another group. The objectives of feature selection are manifolds: to improve the learner (i.e., classifier) performance such as its accuracy or speed and to understand the underlying process that generates the data. 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 (Bins & Draper 2001; Sebban & Nock 2002), dependency measures (Yu & Liu 2004), consistency measures (Dash & Liu 2003; Lashkia & Anthony 2004), and information measures (Battiti 1994; Kwak & Choi 2002). 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. It tends to require intensive computations while it outperforms the filter method in general. 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. In a variety of bioinformatics areas, the feature selection methods have been used, including sequence analysis (Salzberg *et al.* 1998; Delcher *et al.* 1999), microarray analysis (Alon *et al.* 1999; Ben-Dor *et al.* 2000; Golub *et al.* 1999), mass spectra analysis (Petricoin & Liotta 2003; Oh *et al.* 2009), single nucleotide polymorphism (SNP) analysis (Daly *et al.* 2001), and text mining (Cohen & Hersch 2005; Jensen, Saric & Bork 2006; Saeys, Inza & Larrañaga 2007).

Classification is a problem of assigning a sample to a predefined class based on conditional features. Many common classification techniques, including linear discriminate analysis (LDA), decision tree, neural networks, SVM, k-nearest neighbor (kNN), and Bayesian classifiers, have been proposed in a variety of applications. LDA and SVM are two main kinds of linear classifiers. That is, they seek to find a hyperplane for which one group can be correctly separated from another group as much as possible (Lotte *et al.* 2007). SVM proposed by Vapnik and his colleagues is a novel approach for solving classification problems. It is based on the structural risk minimization principle to minimize an upper bound of the generalization error (Vapnik 1995; Jeng 2006).

A major part of our informatics efforts is focused towards providing better tools to the radiotherapy outcome analyst to gain a more insightful understanding of complex variable interactions that affect outcome and support treatment planning systems with improved predictive models of response. Therefore, we have upgraded our in-house software tool DREES (dose-response explorer system) with several statistical and graphical tools; in particular, we have added a new machine learning module based on SVM as discussed further below.

The remainder of this paper is organized as follows. In Sections 2 and 3, we introduce feature selection and classification algorithms investigated in this study. In Section 4, we present a new version of DREES that is equipped with SVM. Experimental results with dose-volume data in lung cancer are shown in Section 5. Finally, we summarize our conclusions in Section 6.

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 (Guyon *et al.* 2002). 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 a feature to be eliminated at each iteration, the weights (w_i) are estimated (see below) and a 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 correlations 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 (Hall & Smith 1999). The fundamental idea behind the method is that good features are highly correlated with the class, but uncorrelated with each other. The evaluation function of a subset of features is:

$$EV_S = \frac{h\overline{r_{cf}}}{\sqrt{h + h(h-1)\overline{r_{ff}}}} \quad (1)$$

where EV_S represents the heuristic evaluation of a feature subset S containing h features; $\overline{r_{cf}}$ and $\overline{r_{ff}}$ are the mean feature-class correlation and the mean feature-feature intercorrelation, respectively.

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 (Liu & Setiono 1995). As a result of discretization, the feature selection is completed. The measure of the Chi-square is defined to be:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^k \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

where,

k is the number of classes,

A_{ij} is the number of patterns in the i th interval, j th class,

R_i is the number of patterns in the i th interval = $\sum_{j=1}^k A_{ij}$,

C_j is the number of patterns in the j th class = $\sum_{i=1}^2 A_{ij}$,

N is the total number of patterns = $\sum_{i=1}^2 R_i$,

E_{ij} is the expected frequency of $A_{ij} = R_i \times C_j / N$.

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, \dots, c_k . The entropy of the class distribution in S is defined as follows:

$$I(S) = -\sum_{i=1}^k \frac{|c_i|}{|S|} \log_2 \frac{|c_i|}{|S|} \quad (3)$$

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

$$\begin{aligned} Gain(F_i) &= I(S) - I(S | F_i), \\ &= I(S) - \sum_{j=1}^t \frac{|S_j|}{|S|} \times I(S_j) \end{aligned} \quad (4)$$

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 (Oh *et al.* 2008).

3. CLASSIFICATION METHODS

3.1 Support Vector Machine

SVM is a supervised learning algorithm, originally designed to solve two-class classification problems (Burges 1998; El Naqa *et al.* 2002; El Naqa *et al.* 2009; Oh *et al.* 2006). 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:

$$f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b, \tag{5}$$

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

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \zeta_i, \tag{6}$$

subject to

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

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:

$$\min \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^n \alpha_i, \tag{7}$$

subject to

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

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. (7), $\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_j)$ is substituted with a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ by the kernel trick. Note that for the linear case $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$. Two typical kernels are polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (e + \mathbf{x}_i^\top \mathbf{x}_j)^d$ and radial basis function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-1/(2\sigma^2) \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ where e , d , and σ are adjustable kernel function parameters (El Naqa, Bradley & Deasy 2008).

3.2 Decision Trees

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 (Witten & Frank 2005).

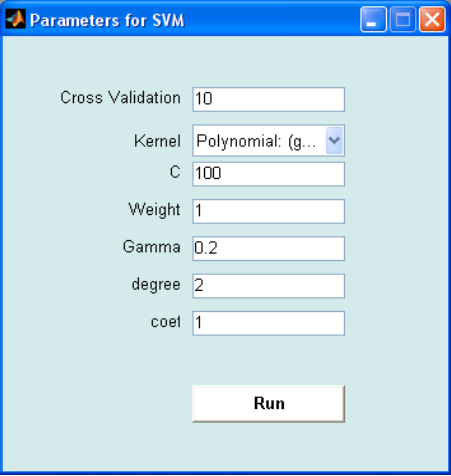
3.3 Random Forest

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 (Rodriguez, Kuncheva & Alonso 2006).

3.4 Naive Bayes

In a naive Bayes classifier, it is assumed that all features are mutually independent given a class label, that is, each feature has the class variable as its parent (Friedman, Geiger & Goldszmidt 1997). 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. In the naive Bayes classifier, the most probable class is obtained by using the Bayes' theorem:

$$c^* = \arg \max_c p(c) = \prod_{i=1}^n p(x_i | c). \tag{9}$$



(a) Parameters for SVM

```

Accuracy = 60.8696% (14/23) (classification)
Accuracy = 65.2174% (15/23) (classification)
Accuracy = 54.5455% (12/22) (classification)
Accuracy = 63.6364% (14/22) (classification)
Accuracy = 77.2727% (17/22) (classification)
Accuracy = 68.1818% (15/22) (classification)
Accuracy = 72.7273% (16/22) (classification)
Accuracy = 52.381% (11/21) (classification)
Accuracy = 80.9524% (17/21) (classification)
Accuracy = 57.1429% (12/21) (classification)
-----
Total samples: 219
0 class samples: 52
1 class samples: 167
accuracy: 0.65293
sensitivity: 0.70667
specificity: 0.63419
mcc: 0.3015
>>
                
```

(b) The results of SVM classification

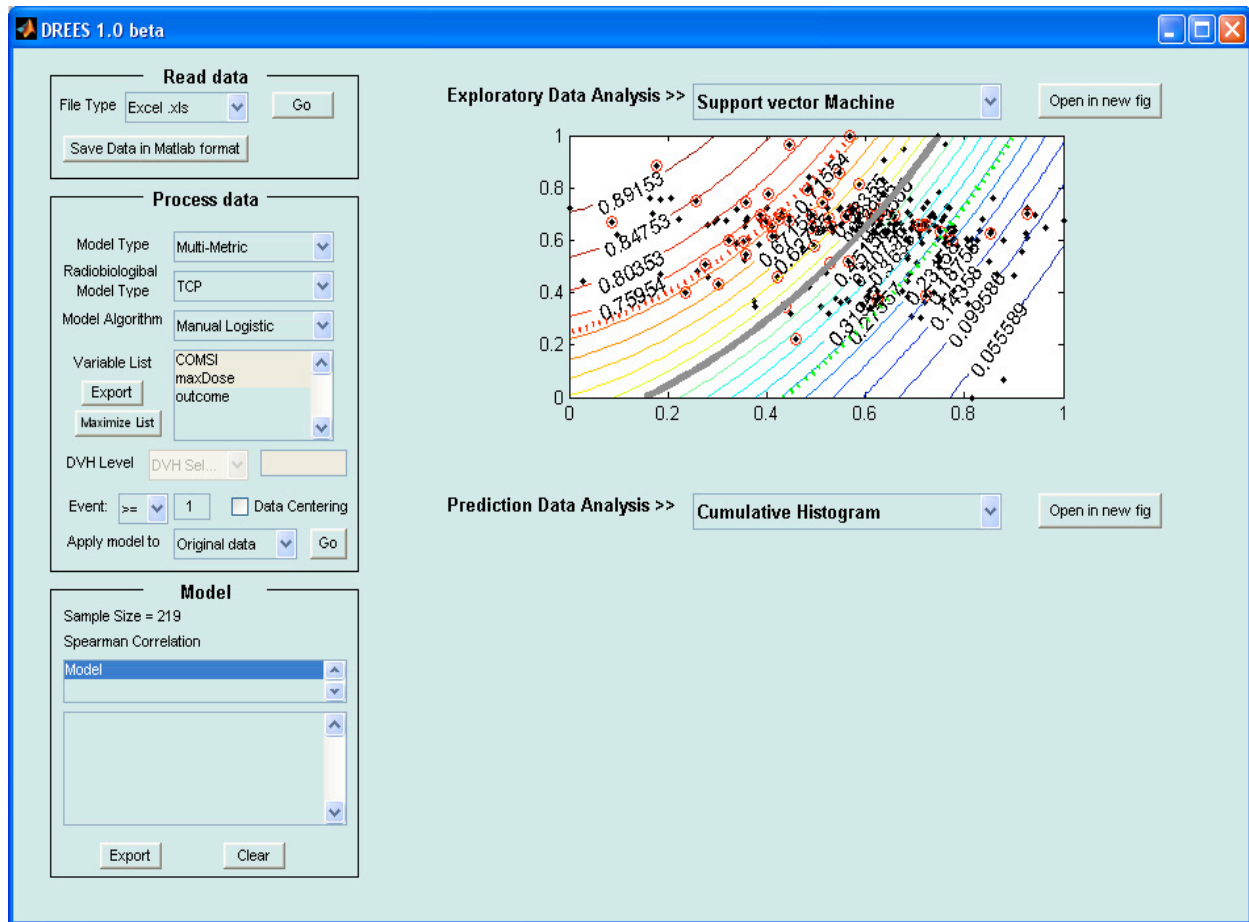


Figure 1. SVM classification in DREES

4. DREES SOFTWARE

Towards fulfilling our objective to provide clinicians and scientists with an accurate, flexible and user-friendly tool to explore radiotherapy outcomes data and model the statistical tumor control or normal tissue complication, we have developed an open-source software called DREES that enables clinical researchers to customize the function for radiotherapy outcome modeling (El Naqa et al. 2006b). DREES is available from <http://radium.wustl.edu/drees/>. Recently, we incorporated a popular SVM code called LibSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) into DREES so that it can provide more powerful ability for the analysis of radiotherapy data (Chang & Lin 2001). Figure 1 illustrates a screenshot of the interface for using SVM in DREES. The results of SVM classification are shown in 'Command Window' of Matlab. In addition, new visual representation using SVM was developed as shown in Figure 1. The figure shows the contour plot of SVM for two features, that is, the contour plot represents the kernel-based pneumonitis nonlinear prediction model. The gray line indicates the hyperplane of the SVM classifier. The software is shared based on the GNU General Public License (GPL) v3.

*elnaqa@wustl.edu; phone +1 314 362 0129; fax +1 314 362 8521; radonc.wustl.edu

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

Ranking	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	

Table 2. The performance when the correlation based feature selection is used.

Methods	MCC	Accuracy	Sensitivity	Specificity	AUC
Naïve Bayes	0.3881	0.7795	0.5253	0.8552	0.7615
Random Forest	0.2998	0.7671	0.3912	0.8790	0.6671
Decision Tree	0.2935	0.7944	0.2527	0.9555	0.5952
RBF-SVM	0.4131	0.7474	0.6957	0.7627	0.7292
P-SVM	0.4118	0.7362	0.7240	0.7400	0.7320
L-SVM	0.3222	0.6624	0.7300	0.6425	0.6863

5. EXPERIMENTAL RESULTS

5.1 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.), relative location of the tumor within the lung or nearby heart, and dosimetric features, including mean dose, maximum dose, minimum dose, V_x (volume getting at least x Gy), D_x (minimum dose to the hottest $x\%$ volume), MOH_x (mean dose to the hottest $x\%$ volume), MOC_x (mean dose to the coldest $x\%$ volume) and GEUD (generalized equivalent uniform dose). In this study, we included heart related variables due to the fact that in the radiotherapy of lung cancer, a portion of the heart is typically exposed to a relatively high dose of radiation that causes heart injury (Shafman et al. 2004). Recently, Deasy et al. have reported that heart dose-volume metrics may play a role in RP (Deasy et al. 2008).

Table 3. The performance when the Chi-square feature selection is used.

Methods	Measurements	No. of features									
		1	2	3	4	5	6	7	8	9	10
Naive Bayes	MCC	-0.0117	-0.0121	0.1862	0.1769	0.2737	0.2831	0.2761	0.2888	0.2973	0.2960
	Accuracy	0.7654	0.7650	0.7525	0.7150	0.6988	0.6853	0.6591	0.6502	0.6514	0.6693
	Sensitivity	0.0000	0.0000	0.2398	0.3432	0.5573	0.6073	0.6572	0.7000	0.7123	0.6698
	Specificity	0.9933	0.9929	0.9051	0.8258	0.7406	0.7083	0.6595	0.6351	0.6330	0.6691
	AUC	0.6068	0.7008	0.7067	0.6698	0.6714	0.6929	0.6924	0.6926	0.6971	0.7060
Random Forest	MCC	0.1797	0.3337	0.2921	0.2828	0.2842	0.2648	0.2528	0.2468	0.2485	0.2488
	Accuracy	0.7095	0.7689	0.7570	0.7548	0.7557	0.7538	0.7515	0.7496	0.7503	0.7501
	Sensitivity	0.3643	0.4603	0.4215	0.4158	0.4078	0.3765	0.3660	0.3695	0.3623	0.3655
	Specificity	0.8126	0.8610	0.8572	0.8557	0.8591	0.8661	0.8661	0.8627	0.8658	0.8646
	AUC	0.6031	0.7181	0.6845	0.7020	0.6925	0.6704	0.6817	0.6789	0.6780	0.6794
Decision Tree	MCC	0.3179	0.3116	0.3030	0.2984	0.2959	0.2811	0.2736	0.2713	0.2785	0.2779
	Accuracy	0.8135	0.8100	0.8062	0.8029	0.8024	0.7956	0.7884	0.7879	0.7890	0.7889
	Sensitivity	0.1968	0.2000	0.2013	0.2053	0.2060	0.2192	0.2383	0.2377	0.2465	0.2458
	Specificity	0.9973	0.9917	0.9865	0.9809	0.9801	0.9673	0.9522	0.9518	0.9505	0.9505
	AUC	0.5898	0.5924	0.5917	0.5925	0.5927	0.5907	0.5930	0.5938	0.5970	0.5969
RBF-SVM	MCC	0.1583	0.3609	0.3525	0.3430	0.3294	0.3283	0.3124	0.3107	0.3350	0.3361
	Accuracy	0.6072	0.7363	0.7325	0.7310	0.7251	0.6489	0.6518	0.6556	0.6668	0.6699
	Sensitivity	0.5620	0.6250	0.6182	0.6012	0.5918	0.7708	0.7378	0.7268	0.7460	0.7413
	Specificity	0.6208	0.7694	0.7664	0.7698	0.7648	0.6125	0.6260	0.6343	0.6432	0.6486
	AUC	0.5914	0.6972	0.6923	0.6855	0.6783	0.6917	0.6819	0.6806	0.6946	0.6949
P-SVM	MCC	0.1567	0.3653	0.3511	0.3403	0.3268	0.3303	0.3112	0.3205	0.3311	0.2877
	Accuracy	0.6029	0.7405	0.7392	0.7299	0.7292	0.6529	0.6553	0.7232	0.6891	0.7148
	Sensitivity	0.5692	0.6210	0.5973	0.6000	0.5738	0.7662	0.7288	0.5793	0.6893	0.5408
	Specificity	0.6131	0.7760	0.7816	0.7686	0.7756	0.6191	0.6332	0.7658	0.6889	0.7664
	AUC	0.5912	0.6985	0.6895	0.6843	0.6747	0.6927	0.6810	0.6726	0.6891	0.6536
L-SVM	MCC	0.0913	0.2409	0.2780	0.2735	0.2687	0.2914	0.2793	0.2810	0.2546	0.2817
	Accuracy	0.5436	0.5756	0.6376	0.6366	0.6350	0.6329	0.6340	0.6355	0.6399	0.6869
	Sensitivity	0.5658	0.7553	0.7047	0.6992	0.6937	0.7387	0.7158	0.7158	0.6590	0.5992
	Specificity	0.5398	0.5224	0.6175	0.6180	0.6176	0.6015	0.6096	0.6115	0.6345	0.7131
	AUC	0.5528	0.6389	0.6611	0.6586	0.6556	0.6701	0.6627	0.6637	0.6468	0.6562

5.2 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.

Table 4. The performance when SVM-RFE is used.

Methods	Measurements	No. of features									
		1	2	3	4	5	6	7	8	9	10
Naïve Bayes	MCC	0.0000	-0.0003	0.1069	0.1177	0.1686	0.1758	0.2241	0.2255	0.2376	0.2798
	Accuracy	0.7705	0.7331	0.6858	0.6727	0.6897	0.6971	0.7203	0.7104	0.7152	0.7069
	Sensitivity	0.0000	0.0592	0.3058	0.3530	0.3945	0.3888	0.4055	0.4390	0.4465	0.5498
	Specificity	1.0000	0.9339	0.7988	0.7676	0.7775	0.7890	0.8142	0.7913	0.7954	0.7537
	AUC	0.6959	0.6456	0.6477	0.6506	0.6650	0.6730	0.7050	0.6951	0.6920	0.6913
Random Forest	MCC	0.0805	0.1181	0.1290	0.0780	0.0321	0.0918	0.1011	0.0943	0.2791	0.2779
	Accuracy	0.6676	0.6898	0.6941	0.6932	0.6812	0.7019	0.7029	0.6999	0.7670	0.7631
	Sensitivity	0.3040	0.3123	0.3212	0.2543	0.2073	0.2442	0.2585	0.2497	0.3603	0.3602
	Specificity	0.7761	0.8027	0.8053	0.8245	0.8229	0.8384	0.8358	0.8347	0.8881	0.8829
	AUC	0.6093	0.6281	0.6250	0.6014	0.5825	0.6066	0.6099	0.6051	0.6904	0.6935
Decision Tree	MCC	0.0271	0.0207	0.0209	0.0269	0.0179	0.0120	0.0132	0.0093	0.3798	0.3774
	Accuracy	0.7631	0.7471	0.7473	0.7468	0.7420	0.7409	0.7379	0.7389	0.8028	0.8020
	Sensitivity	0.0448	0.0683	0.0683	0.0762	0.0760	0.0700	0.0803	0.0753	0.3967	0.3947
	Specificity	0.9767	0.9492	0.9494	0.9466	0.9404	0.9411	0.9338	0.9364	0.9233	0.9229
	AUC	0.5109	0.5236	0.5238	0.5251	0.5362	0.5382	0.5332	0.5677	0.6664	0.6651
RBF-SVM	MCC	0.3032	0.3313	0.2970	0.2861	0.2942	0.2975	0.3369	0.3256	0.3734	0.3669
	Accuracy	0.6657	0.7425	0.7168	0.7060	0.6368	0.6996	0.7137	0.7265	0.7519	0.7505
	Sensitivity	0.6910	0.5403	0.5542	0.5608	0.7358	0.5982	0.6368	0.5823	0.5967	0.5897
	Specificity	0.6580	0.8029	0.7654	0.7500	0.6072	0.7302	0.7368	0.7696	0.7983	0.7986
	AUC	0.6745	0.6716	0.6598	0.6554	0.6715	0.6642	0.6868	0.6760	0.6975	0.6941
P-SVM	MCC	0.3074	0.3210	0.3324	0.3064	0.2971	0.3169	0.3494	0.3426	0.3928	0.3782
	Accuracy	0.6710	0.7485	0.7395	0.7173	0.6371	0.7141	0.7418	0.7381	0.7642	0.7569
	Sensitivity	0.6875	0.5033	0.5563	0.5690	0.7405	0.5975	0.5855	0.5812	0.5960	0.5932
	Specificity	0.6658	0.8217	0.7944	0.7615	0.6062	0.7492	0.7888	0.7852	0.8145	0.8058
	AUC	0.6767	0.6625	0.6754	0.6652	0.6734	0.6734	0.6871	0.6832	0.7052	0.6995
L-SVM	MCC	0.2112	0.2123	0.2028	0.2828	0.2888	0.3181	0.3558	0.3456	0.3255	0.3197
	Accuracy	0.5431	0.5421	0.5404	0.6429	0.6492	0.7163	0.7310	0.7276	0.6940	0.6986
	Sensitivity	0.7663	0.7698	0.7578	0.7043	0.7022	0.5928	0.6233	0.6132	0.6653	0.6437
	Specificity	0.4767	0.4744	0.4759	0.6244	0.6332	0.7534	0.7633	0.7619	0.7027	0.7151
	AUC	0.6215	0.6221	0.6169	0.6644	0.6677	0.6731	0.6933	0.6875	0.6840	0.6794

5.3 Performance Metric

Our comparative experiments were performed using the WEKA software package (<http://www.cs.waikato.ac.nz/ml/weka/>). For the unbiased performance estimate, all measurements were averaged after 30 iterations of 10-fold cross-validation (CV) for each classification algorithm.

In the analysis of imbalanced data set, Matthew's correlation coefficient (MCC) is widely used as a performance evaluation metric. MCC is calculated as follows:

$$r = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

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. In addition, we measured accuracy, sensitivity, specificity, and AUC (area under the ROC curve) as performance evaluation metrics. The accuracy, sensitivity and specificity are defined as follows:

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}, Sen = \frac{TP}{TP + FN}, Spe = \frac{TN}{TN + FP}. \tag{11}$$

Table 5. The performance when the information gain based feature selection is used.

Methods	Measurements	No. of features									
		1	2	3	4	5	6	7	8	9	10
Naïve Bayes	MCC	-0.0117	-0.0071	0.0500	0.1687	0.2548	0.2800	0.2730	0.2948	0.2973	0.2862
	Accuracy	0.7654	0.7639	0.7478	0.7426	0.7212	0.6853	0.6757	0.6614	0.6514	0.6404
	Sensitivity	0.0000	0.0062	0.0923	0.2512	0.4562	0.6038	0.6113	0.6848	0.7123	0.7155
	Specificity	0.9933	0.9896	0.9430	0.8893	0.8001	0.7094	0.6950	0.6543	0.6330	0.6179
	AUC	0.6068	0.6740	0.6687	0.6794	0.6884	0.6905	0.6961	0.6977	0.6971	0.6972
Random Forest	MCC	0.1797	0.2524	0.1982	0.2240	0.2456	0.2446	0.2485	0.2417	0.2474	0.2372
	Accuracy	0.7095	0.7458	0.7326	0.7421	0.7463	0.7490	0.7471	0.7454	0.7503	0.7485
	Sensitivity	0.3643	0.3847	0.3270	0.3518	0.3670	0.3655	0.3703	0.3663	0.3602	0.3560
	Specificity	0.8126	0.8534	0.8535	0.8580	0.8592	0.8632	0.8593	0.8584	0.8662	0.8655
	AUC	0.6031	0.6938	0.6396	0.6557	0.6561	0.6902	0.6786	0.6822	0.6877	0.6726
Decision Tree	MCC	0.3179	0.3288	0.3153	0.2818	0.2833	0.2730	0.2763	0.2782	0.2770	0.2790
	Accuracy	0.8135	0.8159	0.8116	0.7983	0.7968	0.7887	0.7884	0.7890	0.7889	0.7893
	Sensitivity	0.1968	0.2062	0.1938	0.1980	0.2120	0.2367	0.2440	0.2457	0.2458	0.2472
	Specificity	0.9973	0.9975	0.9957	0.9772	0.9710	0.9530	0.9503	0.9507	0.9505	0.9507
	AUC	0.5898	0.5998	0.5985	0.5882	0.5865	0.5922	0.5967	0.5973	0.5964	0.5969
RBF-SVM	MCC	0.1583	0.3428	0.2814	0.3342	0.3200	0.3159	0.3221	0.3299	0.3350	0.3388
	Accuracy	0.6072	0.7494	0.7147	0.6811	0.6815	0.6687	0.6829	0.6621	0.6668	0.6706
	Sensitivity	0.5620	0.5448	0.5293	0.7135	0.6865	0.7078	0.6877	0.7467	0.7460	0.7452
	Specificity	0.6208	0.8108	0.7698	0.6715	0.6802	0.6571	0.6815	0.6368	0.6432	0.6482
	AUC	0.5914	0.6778	0.6495	0.6925	0.6834	0.6825	0.6846	0.6917	0.6946	0.6967
P-SVM	MCC	0.1567	0.3291	0.3003	0.3388	0.3279	0.3139	0.3187	0.3338	0.3311	0.3301
	Accuracy	0.6029	0.6815	0.6582	0.7256	0.6826	0.6575	0.6536	0.6889	0.6891	0.6893
	Sensitivity	0.5692	0.6987	0.7033	0.6115	0.6983	0.7288	0.7447	0.6940	0.6893	0.6873
	Specificity	0.6131	0.6767	0.6450	0.7594	0.6781	0.6362	0.6265	0.6874	0.6889	0.6897
	AUC	0.5912	0.6877	0.6742	0.6855	0.6882	0.6825	0.6856	0.6907	0.6891	0.6885
L-SVM	MCC	0.0913	0.2924	0.2324	0.2851	0.2907	0.2807	0.2545	0.2572	0.2546	0.2543
	Accuracy	0.5436	0.6998	0.6552	0.6369	0.6352	0.6351	0.6396	0.6407	0.6399	0.6393
	Sensitivity	0.5658	0.5862	0.5900	0.7198	0.7332	0.7158	0.6592	0.6618	0.6590	0.6597
	Specificity	0.5398	0.7343	0.6746	0.6124	0.6061	0.6111	0.6341	0.6347	0.6345	0.6335
	AUC	0.5528	0.6602	0.6323	0.6661	0.6697	0.6634	0.6466	0.6483	0.6468	0.6466

Table 6. The performance when all features without feature selection are used.

Methods	MCC	Accuracy	Sensitivity	Specificity	AUC
Naïve Bayes	0.2477	0.6411	0.6462	0.6399	0.6822
Random Forest	0.1562	0.7254	0.2850	0.8567	0.6401
Decision Tree	0.2564	0.7511	0.3730	0.8638	0.6191
RBF-SVM	0.2595	0.6388	0.6712	0.6298	0.6505
P-SVM	0.1568	0.6476	0.4758	0.6989	0.5874
L-SVM	0.1568	0.6476	0.4758	0.6989	0.5874

5.4 Feature Selection and Classification

Table 1 displays the top ten features selected by the 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, 'Followup', 'COMSI' (center-of-mass of tumor location in the superior inferior direction), and 'MOH10_heartMC' were selected in IG, Chi-square, and CFS. 'D10_heartMC' was found in IG, Chi-square, and SVM-RFE. It suggests that the features are important for distinguishing the disease (RP) group from the control (no RP) group.

Table 2 through Table 5 show 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 Table 2 illustrates the results obtained using all five features that were searched by CFS. Interestingly, in all cases kernel SVMs (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 Table 2, the best MCC was obtained when CFS was exploited with RBF-SVM, resulting in 0.4131. Also, P-SVM gained comparable MCC (0.4118). Table 6 shows the MCC values when all features were used without feature selection. As can be seen in the table, in all cases MCC values were much lower than those gained when only a few important features were utilized. It justifies the importance of feature selection in classification algorithms. Figure 2 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 (0.4131) 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 obtained with these parameters were 74.74%, 69.57%, 76.27%, and 0.7292, respectively.

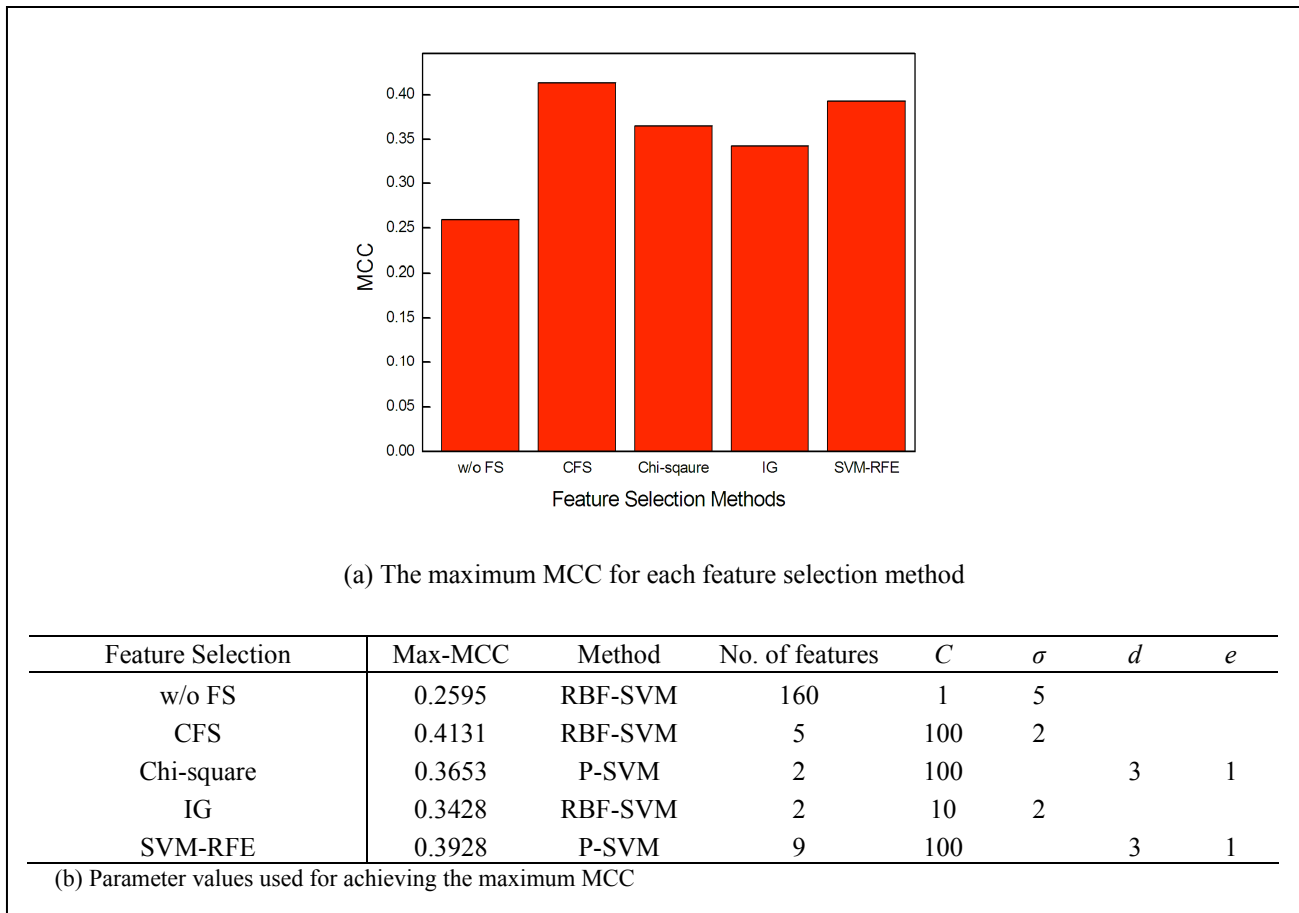


Figure 2. The comparison of the maximum MCC across all classification algorithms for each feature selection method. Note that 'w/o FS' means 'without feature selection'.

6. CONCLUSION

We have compared the performance of different machine learning algorithms in identifying significant features that are related to RP and could be used in building patients' classification risk models of this disease. In our classification experiments with the selected features, the kernel SVMs showed a higher MCC than not only linear SVM but also other competing classification algorithms after correction for imbalance effect. It is our expectation that the application of machine learning methods to the analysis of post-radiotherapy data will shed more light on a better understanding of underlying mechanisms in normal tissue toxicities and advance the clinical translational goal of individualizing radiotherapy in NSCLC patients. To support this goal, we have developed a graphical user interface (GUI) tool to explore radiotherapy outcomes data and build data-driven statistical tumor control or normal tissue complications. We will continue to develop DREES based on user's feedback, as an informatics tool to aid medical physicists and clinical researchers to build more predictive radiotherapy outcome models.

7. ACKNOWLEDGEMENTS

We thank Dr. Joseph Deasy for valuable suggestions. This work was partially supported by NIH grant 1K25CA128809.

8. REFERENCES

- [1] Alon, U, Barkai, N, Notterman, DA, Gish, K, Ybarra, S, Mack, D & Levine, AJ 1999, 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *Proc. Nat. Acad. Sci.*, vol. 96, 6745–50.
- [2] American Cancer Society 2008, *Cancer Facts and Figures*, Atlanta, GA: American Cancer Society.
- [3] Battiti, R 1994, 'Using mutual information for selecting features in supervised neural net learning', *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–50.
- [4] Ben-Dor, A, Bruhn, L, Friedmann, N, Nachman, I, Schummer, M & Yakhini, Z 2000, 'Tissue classification with gene expression profiles', *J. Comput. Biol.*, vol. 7, 559–84.
- [5] Bins, J & Draper, B 2001, 'Feature selection from huge feature sets', in *Proc. Int. Conference Computer Vision*, pp. 159–65.
- [6] Burges, C 1998, 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–67.
- [7] Chang, C & Lin, C 2001, LIBSVM: A Library for Support Vector Machines, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [8] Cohen, A & Hersch, W 2005, 'A survey of current work in biomedical text mining', *Brief. Bioinformatics*, vol. 6, 57–71.
- [9] Daly, MJ, Rioux, JD, Schaffner, SF, Hudson, TJ & Lander, ES 2001, 'High-resolution haplotype structure in the human genome', *Nat. Genet.*, vol. 29, 229–32.
- [10] Dash, M & Liu, H 2003, 'Consistency-based search in feature selection', *Artificial Intelligence*, vol. 151, pp. 155–76.
- [11] Deasy, JO, Niemierko, A, Herbert, D, Yan, D, Jackson, A, Haken, RT, Langer, M, Sapareto, S & AAPM/NIH 2002, 'Methodological issues in radiation dose-volume outcome analyses: summary of a joint AAPM/NIH workshop', *Medical Physics*, vol. 29, no. 9, pp. 2109–27.
- [12] Deasy, JO, Trovo, M, Huang, EX, Mu, Y, El Naqa, I, Bradley, JD 2008, 'High-dose heart irradiation is a statistically significant risk factor for radiation pneumonitis within logistic-multivariate modeling', ASTRO 2008.
- [13] Delcher, A, Delcher, AL, Harmon, D, Kasif, S, White, O & Salzberg, SL 1999, 'Improved microbial gene identification with GLIMMER', *Nucleic Acids Res.*, vol. 27, 4636–41.
- [14] El Naqa, I, Yang, Y, Wernick, M, Galatsanos, N & Nishikawa, R 2002, 'A support vector machine approach for detection of microcalcifications', *IEEE Trans. Medical Imaging*, vol. 21, no. 12, pp. 1552–63.
- [15] El Naqa, I, Bradley, J, Blanco, A, Lindsay, P, Vicic, M, Hope, A & Deasy, J 2006, 'Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors', *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 64, no. 4, pp. 1275–86.
- [16] El Naqa, I, Suneja, G, Lindsay, P, Hope, A, Alaly, J, Vicic, M, Bradley, J, Apte, A & Deasy, J 2006, '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. 5719–35.
- [17] El Naqa, I, Bradley, J & Deasy, J 2008, 'Nonlinear kernel-based approaches for predicting normal tissue toxicities', in *Proc. of 7th Int. Conference on Machine Learning and Applications*, pp. 539–44.

- [18] El Naqa, I, Bradley, JD, Lindsay, PE, Hope, AJ & Deasy, JO 2009, 'Predicting radiotherapy outcomes using statistical learning techniques', *Physics in Medicine and Biology*, vol. 54, pp. S9–S30.
- [19] Friedman, N, Geiger, D & Goldszmidt, M 1997, 'Bayesian network classifiers', *Mach. Learn.*, vol. 29, no. 2, pp. 131–64.
- [20] Golub, TR, Slonim, DK, Tamayo, P, Huard, C, Gaasenbeek, M, Mesirov, JP, Coller, H, Loh, ML, Downing, JR, Caligiuri, MA, Bloomfield, CD & Lander, ES 1999, 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, vol. 286, 531–7.
- [21] Guyon, I, Weston, J, Barnhill, S & Vapnik, V 2002, 'Gene selection for cancer classification using support vector machines', *Machine Learning*, vol. 46, pp. 389–422.
- [22] Hall, MA & Smith, LA 1999, 'Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper', in *Proc. of FLAIRS Conference*, pp. 235–9.
- [23] Hope, A, Lindsay, P, El Naqa, I, Alaly, J, Vicic, M, Bradley, J & Deasy, J 2006, 'Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters', *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 65, no. 1, pp. 112–24.
- [24] Jeng, JT 2006, '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.
- [25] Jensen, LJ, Saric, J & Bork, P 2006, 'Literature mining for the biologist: from information retrieval to biological discovery', *Nat. Rev. Genet.*, vol. 7, 119–29.
- [26] Kwak, N & Choi, CH 2002, 'Input feature selection for classification problems', *IEEE Trans. Neural Networks*, vol. 3, no. 1, pp. 143–59.
- [27] Lashkia, G & Anthony, L 2004, 'Relevant, irredundant feature selection and noisy example elimination', *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 34, no. 2, pp. 888–97.
- [28] Liu, H & Setiono, R 1995, 'Chi2: feature selection and discretization of numeric attributes', in *Proc. of 7th Int. Conference on Tools with Artificial Intelligence*, pp. 388–91.
- [29] Lotte, F, Congedo, M, Lécuyer, A, Lamarche, F & Arnaldi, B 2007, 'A review of classification algorithms for EEG-based brain-computer interfaces', *J. Neural Eng.*, vol. 4, R1–13.
- [30] Oh, JH, Nandi, A, Gurnani, P, Knowles, L, Schorge, J, Rosenblatt, KP & Gao, J 2006, 'Proteomic biomarker identification for diagnosis of early relapse in ovarian cancer', *J. of Bioinformatics and Computational Biology*, vol. 4, no. 6, pp. 1159–79.
- [31] Oh, JH, Kim, YB, Gurnani, P, Rosenblatt, KP & Gao, J 2008, 'Biomarker selection and sample prediction for multicategory disease on MALDI-TOF data', *Bioinformatics*, vol. 24, pp. 1812–8.
- [32] Oh, JH, Gurnani, P, Schorge, J, Rosenblatt, KP & Gao, JX 2009, 'An extended Markov blanket approach to proteomic biomarker detection from high-resolution mass spectrometry data', *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, pp. 195–206.
- [33] Petricoin, E & Liotta, L 2003, 'Mass spectrometry-based diagnostic: the upcoming revolution in disease detection', *Clin. Chem.*, vol. 49, 533–4.
- [34] Rodriguez, JJ, Kuncheva, LI & Alonso, CJ 2006, 'Rotation forest: a new classifier ensemble method', *IEEE. Trans. Pattern Analysis and machine Intelligence*, vol. 28, pp. 1619–30.
- [35] Saeys, Y, Inza, I & Larrañaga, P 2007, 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, vol. 23, 2507–17.
- [36] Salzberg, SL, Delcher, AL, Kasif, S & White, O 1998, 'Microbial gene identification using interpolated Markov models', *Nucleic Acids Res.*, vol. 26, 544–8.
- [37] Sebban, M & Nock, R 2002, 'A hybrid filter/wrapper approach of feature selection using information theory', *Pattern Recognition*, vol. 35, no. 4, pp. 835–46.
- [38] Shafman, T, Yu, X, Vujaskovic, Z, Anscher, MA, Miller, K, Prosnitz, RG & Marks, LB 2004, 'Radiation-induced lung and heart toxicity', *Advances in Radiation Oncology in Lung Cancer*. New York: Springer-Verlag.
- [39] Spencer, S, Bonnin, D, Deasy, J, Bradley, J & El Naqa, I 2009, 'Bioinformatics methods for learning radiation-induced lung inflammation from heterogeneous retrospective and prospective data', *J. of Biomedicine and Biotechnology*.
- [40] Vapnik, V 1995, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.
- [41] Witten, I & Frank, E 2005, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann.
- [42] Yu, L & Liu, H 2004, 'Efficient feature selection via analysis of relevance and redundancy', *J. Machine Learning Research*, vol. 5, pp. 1205–24.