Combining multiple models to generate consensus: Application to radiation-induced pneumonitis prediction

Shiva K. Das^{a)} and Shifeng Chen

Department of Radiation Oncology, Duke University Medical Center, Durham, North Carolina 27710

Joseph O. Deasy

Department of Radiation Oncology, Washington University School of Medicine, St. Louis, Missouri 63110

Sumin Zhou and Fang-Fang Yin

Department of Radiation Oncology, Duke University Medical Center, Durham, North Carolina 27710

Lawrence B. Marks

Department of Radiation Oncology, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599

(Received 24 January 2008; revised 8 September 2008; accepted for publication 15 September 2008; published 21 October 2008)

The fusion of predictions from disparate models has been used in several fields to obtain a more realistic and robust estimate of the “ground truth” by allowing the models to reinforce each other when consensus exists, or, conversely, negate each other when there is no consensus. Fusion has been shown to be most effective when the models have some complementary strengths arising from different approaches. In this work, we fuse the results from four common but methodologically different nonlinear multivariate models (Decision Trees, Neural Networks, Support Vector Machines, Self-Organizing Maps) that were trained to predict radiation-induced pneumonitis risk on a database of 219 lung cancer patients treated with radiotherapy (34 with Grade 2+ postradiotherapy pneumonitis). Each model independently incorporated a small number of features from the available set of dose and nondose patient variables to predict pneumonitis; no two models had all features in common. Fusion was achieved by simple averaging of the predictions for each patient from all four models. Since a model’s prediction for a patient can be dependent on the patient training set used to build the model, the average of several different predictions from each model was used in the fusion (predictions were made by repeatedly testing each patient with a model built from different cross-validation training sets that excluded the patient being tested). The area under the receiver operating characteristics curve for the fused cross-validated results was 0.79, with lower variance than the individual component models. From the fusion, five features were extracted as the consensus among all four models in predicting radiation pneumonitis. Arranged in order of importance, the features are (1) chemotherapy; (2) equivalent uniform dose (EUD) for exponent $a=1.2$ to 3; (3) EUD for $a=0.5$ to 1.2, lung volume receiving >20 –30 Gy; (4) female sex; and (5) squamous cell histology. To facilitate ease of interpretation and prospective use, the fused outcome results for the patients were fitted to a logistic probability function. © 2008 American Association of Physicists in Medicine. [DOI: [10.1118/1.2996012](https://doi.org/10.1118/1.2996012)]

Key words: radiation pneumonitis, modeling, prediction, multiple models, decision fusion

I. INTRODUCTION

The literature on causes of radiation-induced pneumonitis reflects a diversity of factors.^{1–22} For example, normal lung volumes above several different radiation dose thresholds, ranging from 5–50 Gy,^{1–11} have been identified as correlating to radiation pneumonitis. Three probable sources of this diversity are (1) differing radiotherapy treatment techniques in the respective patient populations, which, in some sense, yield truly different predictors; (2) differences between the outcome model methods used to analyze the data; and (3) strong correlation among the factors, with an element of chance as to which one is selected as most important in any given study. Typically, outcome models may take different linear or nonlinear approaches to approximating a

probability-of-outcome surface (more than one surface is possible) as a function of the selected variables (features). The models effectively make different assumptions about the shape (linear/nearly linear/highly nonlinear) and complexity (number of features used) of the desired outcome surface. In addition, the models have learning strategies that emphasize differing aspects of the input data. It seems reasonable to suppose that no one approach will, by chance, happen to identify the “true,” perfectly accurate radiobiological response surface, but that they may approximate the true surface with differing biases. If this is true, and if the biases from the different models are at least partially complementary, it should be possible to generate a less biased answer by combining their results via consensus (decision fusion).^{23–26}

A consequence of decision fusion is that strongly predictive factors can be identified as those that most strongly influence the consensus results.

Decision fusion has been widely utilized in several areas to combine statistical predictions from different models. For example, it has been used in the detection of explosive mines,²⁷ automatic brain MRI segmentation,²⁸ urban remote sensing,²⁹ blotch detection in digitized films,³⁰ motor fault diagnosis,³¹ pattern recognition of handwritten letters,²⁵ and various machine learning datasets.³² In all of these studies, the fusion of multiple models has been shown to improve performance over the individual component models. Decision fusion is typically achieved by pooling the predictions from the different models and extracting a single prediction value corresponding to the minimum, maximum, median, simple average, weighted average, logistic regression, or more complex Bayesian combinations.^{23–26} The fused prediction results are presumably strengthened by the partially complementary strengths (i.e., differing biases) of the component models. At the same time, the fused results may be more realistic, since features that the models agree/disagree on are implicitly strengthened/weakened via fusion.

In this work, we create a fused model from simple averaging of the predictions of four nonlinear multivariate models (Neural Networks, Decision Trees, Support Vector Machines, and Self-Organizing Maps), developed earlier.^{33–35} Each of these models attempted to independently predict the risk of Grade ≥ 2 radiation-induced pneumonitis in our database of thoracic radiotherapy patients (219 patients, 34 with grade ≥ 2 radiation pneumonitis). We use the fusion to extract the dose and nondose consensus features that have the strongest influence on the fused results, i.e., features that are most important to all four models in predicting radiation-pneumonitis risk.

While the fused model is intended for prospective prediction, it does not provide a quantitative closed-form equation that can be used to readily interpret the variation in predicted response with changes in the features. To facilitate this physical interpretation, we use the fusion results as the “ground truth” analog outcome, rather than the binary clinical outcome of radiation pneumonitis/no radiation pneumonitis. The fusion analog outcome is fitted to a logistic regression probability function of the important features.

II. METHODS AND MATERIALS

The aims of this study are to use the fusion of different models to attempt to extract the ground truth on patient predictions and important features (consensus features that are important to all four models), and to then generate a logistic probability function describing the fusion results. This section describes the patient database, briefly describes the four models (described in greater detail in the Appendix), the fusion methodology, extraction of important consensus features, and the fitting of a logistic probability function.

II.A. Patient database

The four models fused in this work [Neural Networks, Decision Trees, Support Vector Machines (SVM), Self-Organizing Maps (SOM)] were built from a database of 219 lung cancer patients who underwent thoracic radiotherapy (RT) at Duke University Medical Center and were subsequently assessed for radiation-induced pneumonitis at follow-up (typically at 1, 3, and then every 3–4 months postradiotherapy). Patient and treatment characteristics are listed in Table I (subsets of this patient group have been analyzed previously^{36,37}). The patients were typically treated with either anterior-posterior primary beams delivering 40–45 Gy, followed by off-cord parallel opposed boost beams, or multiple noncoplanar, nonaxial beams. Treatments were once daily (1.8–2.0 Gy per fraction) or twice daily (1.25 and 1.6 Gy per fraction concurrently to the clinical target volume and gross tumor volume, respectively; fractions separated by a minimum of 6 h). Radiation-induced pneumonitis was graded as: Grade 0 (no increase in pulmonary symptoms due to RT), Grade 1 (RT-induced symptoms not requiring initiation or increase in steroids and/or oxygen), Grade 2 (RT-induced pulmonary symptoms requiring initiation or increase in steroids), Grade 3 (RT-induced pulmonary symptoms requiring oxygen), and Grade 4 (RT-induced pulmonary symptoms requiring assisted ventilation or causing death). Of these 219 patients, 34 were diagnosed with Grade ≥ 2 radiation pneumonitis, the toxicity endpoint for this work (this study excluded 16 patients who were listed as hard-to-score for pneumonitis due to “exacerbation of preexisting lung disease (chronic obstructive pulmonary disease); tumor regrowth/progression; and cardiac disease”³⁸).

The set of database variables collected for the patients was as follows: lung dose-volume histogram and generalized equivalent uniform doses (EUD) derived from the lung dose volume histogram for exponent a ranging from 0.4 to 4 in increments of 0.1 ($EUD = (\sum_i V_i D_i^a / \sum_i V_i)^{1/a}$), mean heart dose, age, sex, race, histology (squamous cell cancer, adenocarcinoma, nonsmall cell cancer, small cell cancer, large cell cancer, other), tumor stage (1 through 4), central/peripheral tumor location in lung, anatomic upper/lower/middle tumor location in lung, right/left lobe tumor location in lung, chemotherapy schedule with respect to RT (prior, concurrent, prior and concurrent, post, concurrent and post, no chemo), surgery (yes or no), once/twice daily RT, pre-RT FEV1% (forced expiratory volume in 1 s, expressed as percent of predicted normal), and pre-RT DLCO% (diffusion capacity of carbon monoxide in lung expressed as percent of predicted normal). Note that lung was defined as whole lung minus the clinical target volume. The doses used to compute the dose-volume histograms were from tissue inhomogeneity corrected plans (doses were not radiobiologically corrected). Only a small subset of the variables, termed features, was selected by the four models to predict radiation pneumonitis, as briefly described in the next subsection.

TABLE I. Patient characteristics and treatment details. Abbreviations: SCLC=small cell lung cancer; NSCLC=nonsmall cell lung cancer; FEV1=forced expiratory volume in 1 s; DLCO=diffusion capacity of carbon monoxide.

Characteristic	Radiation pneumonitis	No radiation pneumonitis
Patient number	34	185
Age in years: mean [range]	62 [43–83]	64 [27–87]
Gender: male/female (%)	41/59	56/44
Race: white/black or asian (%)	91/9	82/18
Pre-RT FEV 1%: mean [range]	65 [15–99]	64 [16–127]
Pre-RT DLCO%: mean [range]	72 [32–111]	61 [11–129]
Fractionation: once/twice daily (%)	24/76	31/69
Histological type (%)		
SCLC	12	10
NSCLC (squamous, adeno)	88 (29, 21)	88 (26, 19)
Other	0	2
Tumor stage (%)		
I-II	15	17
III-IV	85	80
Recurrent	0	3
Tumor location (%)		
Central/peripheral location	65/35	60/40
Upper/lower/middle location	56/26/18	61/21/18
Right/left location	59/41	58/42
Chemotherapy schedule (%)		
Concurrent with RT	9	15
Pre RT	74	43
Pre RT+concurrent RT	6	6
Concurrent RT+Post RT	0	4
Post RT	0	1
No chemo	11	31
Radiotherapy dose in Gy: mean [range]		
Once daily treatment	66.0 [45.0–86.4]	62.5 [30.0–85.2]
Twice daily treatment	76.1 [56.4–86.2]	71.6 [45.0–86.4]

II.B. Description of models

Construction of the four models for radiation pneumonitis prediction (Neural Networks, Decision Trees, SVM, SOM) is described in detail in our previous work^{33–35,39} (see the Appendix for details). A brief description of the models is as follows. Neural Networks used activation functions of weighted features to trigger responses that identified patients who developed pneumonitis. Decision Trees subdivided the patient population to identify trends within each subpopulation that discriminated between patients with and without pneumonitis. Support Vector Machines separated the patient populations with and without pneumonitis using a multidimensional plane that had some tolerance to misclassified outliers. Self-Organizing Maps separated patients according to similarities in features within the patient populations with/without pneumonitis.

In general, each model was built with features that were sequentially chosen from the available set of database variables. Since the models are nonparametric, the possibility of overfitting is high. Overfitting implies using too many features to create a good fit to the training data and, consequently, faring poorly on unseen test data. To reduce overfitting, a portion of the training data was used to check features

selected by the remainder of the training data. This check dismissed features that were likely to overfit. Each model only selected some small subset of the total available set of database variables as features, and not all features were common to all four models. The features selected by the individual models were as follows. For neural networks:³³ V_{16} , EUD for $a=1.0$ and 3.5 , forced expiratory volume in 1 s (FEV_1), diffusion capacity of carbon monoxide (DLCO%), and pre-RT chemotherapy. For decision trees:³⁵ Lyman NTCP, V_{50} , V_{60} , EUDs for exponent $a=2, 3$, and 4 , female gender, squamous cell histology, pre-RT chemotherapy, and once daily treatment. For support vector machines:³⁴ EUD $a=1.2–1.4$, pre-RT chemotherapy, peripheral tumor location, female gender, and adenocarcinoma/small cell histology. For self-organizing maps:³⁹ EUD for $a=0.9–1.1$, pre-RT chemotherapy, squamous cell histology, and lower lobe tumor location.

II.C. Fusion of models

The models were fused by averaging their individual outcome predictions for each patient. The outcome predictions from a model are analog values, with higher values implying

greater risk of contracting radiation pneumonitis. For each model, patient outcome predictions were generated by testing 1/10 of the patients (test set), in turn, with a model built from the remaining 9/10 of the patients (training set). This approach is commonly termed tenfold cross validation,⁴⁰ where the patients take turns being part of the test set. Since the model has no knowledge of the patients in the test set, this was meant to be a realistic estimate of outcome prediction. However, the outcome prediction result could be dependent on the manner in which patients were divided into training and test sets. To reduce this dependency, the tests were repeated with 100 random divisions of the patients into 9/10 size training and 1/10 size test sets for tenfold cross validation (the time-intensive nature of computations restricted us to 100 tests). In essence, each of the four models predicted the outcome for a test patient 100 different times, using models constructed from 100 training sets that excluded the patient being tested.

Averaging these 100 predictions from all four models was used to generate a fusion predicted probability of pneumonitis. However, the predictions from different classifiers are signed real numbers (>0 for predicted pneumonitis, <0 for predicted nonpneumonitis, using the training set to determine a cutoff) that cannot be added together directly because of prediction scale differences between classifiers. To allow direct addition, the 100 patient predictions from each classifier were converted to a binary value (+1 for predicted pneumonitis, 0 for no predicted pneumonitis) prior to averaging. Thus, the fusion probability extremes of 1/0 imply that all 100 predictions from each of the four models agreed that the patient will/will not suffer radiation pneumonitis.

II.D. Extracting consensus prediction features

Each model independently selected a small number of features from the available set of variables (Sec. II A). While one expects some similarity in features selected by the models, methodological differences imply that not all features will be in common. In this work, a feature is considered to be predictive, in consensus, if it has a significant influence on the fused prediction.

The predictive strength of a selected feature (selected by one or more of the models) was based on rank reversal, as follows. The fused patient prediction results were used to rank patients from most-to-least severe risk of contracting radiation pneumonitis. The feature to be evaluated was exchanged in reverse-rank order between patients, i.e., the first-ranked patient exchanged its feature value with the last-ranked patient, the second-ranked patient with the second from last patient, and so on. Model predictions on these altered patient datasets with swapped features were then regenerated and fused, resulting in an altered patient ranking. Thus, a more strongly predictive feature would be more likely to reverse the overall patient ranking. Since the features selected by the models are not necessarily independent (e.g., highly correlated dose-volume histogram features), the potential for rank reversal was evaluated on one correlated group of features (Pearson correlation ≥ 0.90) at a time.

TABLE II. Spearman intra- and interclassifier correlation for classifiers: Neural Network, Decision Tree, Support Vector Machine (SVM), and Self-Organizing Map (SOM). Intra-classifier correlation (along-diagonal axis of table) is between any 2 of the 100 predictions from a single classifier. Interclassifier correlation (off-diagonal axis) is between any 2 of the 100 predictions from two different classifiers. The entry in each cell is the mean correlation, with minimum to maximum range shown in brackets.

	Neural Net	Decision Tree	SVM	SOM
Neural Net	0.90 (0.83 $\rightarrow$ 0.94)	0.63 (0.54 $\rightarrow$ 0.70)	0.25 (0.08 $\rightarrow$ 0.41)	0.56 (0.45 $\rightarrow$ 0.64)
Decision Tree		0.97 (0.94 $\rightarrow$ 0.99)	0.19 (0.04 $\rightarrow$ 0.30)	0.61 (0.52 $\rightarrow$ 0.69)
SVM			0.70 (0.55 $\rightarrow$ 0.83)	0.31 (0.16 $\rightarrow$ 0.50)
SOM				0.91 (0.86 $\rightarrow$ 0.94)

Quantitatively, the effectiveness of a grouped set of features in predicting radiation pneumonitis was evaluated by Spearman correlation between the pre- and post-exchange patient predictions. The correlation extremes of -1 and $+1$ indicate complete rank reversal (most effective) and no rank reversal (least effective), respectively.

II.E. Fitting a logistic probability function to the fusion results

While the fusion model is meant for prospective use, it does not directly provide a closed-form equation relating response to the selected features. Consequently, it does not lend itself to easy interpretation of the variation in response with variation of the feature values. To facilitate interpretation, the fused patient probabilities were fitted to a logistic probability function. The fusion was used to supply two pieces of information towards the logistic probability fitting: (a) an analog fusion probability value, rather than the binary clinical outcomes of radiation pneumonitis/no radiation pneumonitis; and (b) the features deemed most important by the fusion. Thus, the logistic function fitted the analog fusion probabilities (assumed to be the “ground truth”) to the fusion features.

III. RESULTS

III.A. Fusion of models

Table II shows the inter- and intraclassifier Spearman rank correlations between any 2 of the 100 predictions from the four models. Correlation was on the original (nonbinarized) model outputs. It is seen that, while intraclassifier prediction is highly correlated (except for SVM), interclassifier correlation is low. This indicates that the different models do indeed have different biases.

The fusion result from combining the 100 cross-validated predictions from each of the four models yielded an ROC area of 0.79. The variance in this result may be indirectly gauged by combining increasing numbers of randomly selected predictions from the 100 predictions for each model. Figure 1 compares the fusion of 1–10 predictions to each of

FIG. 1. ROC area from the fusion of 1–10 randomly selected predictions from each of the four models (95% confidence intervals are indicated by the error bars). For comparison, the ROC areas of the individual component models are also shown (NN-Neural Networks, DT-Decision Trees, SVM-Support Vector Machines, SOM-Self Organizing Maps).

the component models (95% confidence intervals are indicated by the error bars). This figure demonstrates that fusing even one prediction from each of the four models results in a higher mean ROC area, compared to the component models. There is little change in the mean ROC area when >1 prediction is combined; however, the 95% confidence interval progressively shrinks, rapidly improving over the corresponding confidence intervals from the component models.

III.B. Extracting consensus prediction features

Table III summarizes the results of analyzing the importance of features in the fused model, i.e., the “consensus” opinion of the component models. The table lists Spearman correlation coefficients between the fused predictions before and after reverse-rank exchange of features. To assess variance in the correlation coefficients, the fused predictions before and after feature exchange were repeatedly sampled (10^5 samples, each sample fusing ten random predictions from each of the four models). Only significant groupings of fea-

TABLE III. Significant consensus features in fusion. Groups of correlated features were exchanged between patients in reverse-rank order, to determine their efficacy in reversing the patient ranking. A less important feature would be less capable of altering the patient (correlation ≈ 1) than a more important feature (correlation ≈ -1).

Grouped features	Spearman rank correlation
Dose group 1 (Pearson correlation ≥ 0.9): EUD $a=0.5 \rightarrow 1.2$, vol $> 20-30$ Gy	0.853 ± 0.007
Dose group 2 (Pearson correlation ≥ 0.9): EUD $a=1.2 \rightarrow 3.0$	0.604 ± 0.011
Chemotherapy prior to RT	0.371 ± 0.016
Female sex	0.939 ± 0.005
Squamous cell histology	0.955 ± 0.005
Dose group 2+chemo prior to RT	-0.373 ± 0.016
All dose features	0.191 ± 0.015
All dose features+chemo prior to RT	-0.641 ± 0.012

FIG. 2. Fusion computed probability of radiation-induced pneumonitis as a function of EUD $a=2$, the midpoint of the most important dose range, for cases with (solid circle) and without (open circle) radiation-induced pneumonitis.

tures ($p < 0.01$) are listed, i.e., significant drop in the correlation between any two pre- vs post-exchange sampled predictions, in relation to the correlation between any two pre-exchange sampled predictions. Two groups of dose features show significance: group 1: EUD for $a=0.5$ to 1.2 and lung volume above $20-30$ Gy, and group 2: EUD for $a=1.2$ to 3.0 . The latter group was more important in predicting radiation pneumonitis, with lower Spearman reverse-rank correlations. Chemotherapy prior to RT was more important in predicting radiation pneumonitis than either of these dose groups. All dose features, combined, were more important than chemotherapy prior to RT. However, the combination of group 2 dose features (EUD for $a=1.2$ to 3.0) with chemotherapy prior to RT was more important in predicting pneumonitis than all dose features combined (Spearman: -0.37 vs 0.19). Based on the correlation coefficients, female sex and squamous cell histology were less important predictors.

Figure 2 plots the fusion probability as a function of EUD $a=2$, the midpoint of the most important dose range. We have chosen to bundle the dose range into a single feature (EUD $a=2$), due to the very high correlation between the EUDs in the parameter range $a=1.2-3.0$. The fusion probability essentially “downgrades” patients with pneumonitis who had low EUD $a=2$ values; conversely, it “upgrades” patients without pneumonitis who had high EUD $a=2$ values. Figure 3 plots the fusion probability as a function of EUD $a=2$ for cases with and without pre-RT chemotherapy. Patients who received pre-RT chemotherapy are predicted by the fusion to be at generally higher risk for pneumonitis. Figures 4(a) and 4(b) further separate the cases with pre-RT chemotherapy [4(a)] and without pre-RT chemotherapy [4(b)] by gender. In cases with pre-RT chemotherapy, there appears to be some distinction by gender, with females predicted to be at greater risk; there appears to be little distinction by gender in cases without pre-RT chemotherapy.

FIG. 3. Fusion computed probability of radiation-induced pneumonitis as a function of EUD $a=2$ for cases with (x) and without (square) pre-RT chemotherapy.

III.C. Generating a logistic probability fit to the fusion results

Based on the trends observed in Figs. 2–4, we chose to fit the fusion probability to a logistic function of the variables EUD $a=2$, chemotherapy pre-RT, and gender, as follows:

Logistic probability fit

$$= 1 / (1 + e^{(a+b \cdot \text{dose} + c \cdot \text{chemotherapy} \cdot \text{gender} + d \cdot \text{chemotherapy})}), \quad (1)$$

where a, b, c, d are fit parameters, dose refers to EUD $a=2$ (Gy), chemotherapy is 1/0 for pre-RT/no pre-RT chemotherapy, and gender is 1/0 for female/male. Thus, the logistic fit probability for males/females without pre-RT chemotherapy is characterized by the slope b with respect to dose. Males with pre-RT chemotherapy are characterized by the slope b with offset d . Females with pre-RT chemotherapy are characterized by the same slope as for males with pre-RT chemotherapy, but with a higher offset c . The fitted values for these parameters, with 95% confidence intervals in brackets, are $a=8.630$ (7.101–10.158), $b=-0.204$ (–0.245––0.164), $c=-0.802$ (–1.216––0.388), $d=-2.406$ (–2.892––1.920). Figure 5(a) shows the logistic probability fits for males/females without pre-RT chemotherapy, males with pre-RT chemotherapy, and females with pre-RT chemotherapy; the fits are superimposed on the fusion probabilities for patients with and without pre-RT chemotherapy. Figure 5(b) shows the same fits superimposed on the fusion probabilities for patients with and without pneumonitis. Figure 6 shows the fits with 95% confidence intervals for the patient predictions. Using the fit, a 50% probability of contracting radiation pneumonitis for males/females without pre-RT chemotherapy corresponds to an incidental lung EUD=42 Gy (95% confidence interval: 40–45 Gy), for males with pre-RT chemotherapy: EUD=30 Gy (29–32 Gy), and for females with pre-RT chemotherapy: EUD=27 Gy (25–28 Gy).

IV. DISCUSSION

We have used the fusion of four methodologically different models to generate a consensus on the causes of

FIG. 4. (a) Fusion computed probability of radiation-induced pneumonitis for cases with pre-RT chemotherapy, separated by gender (asterisk-females, triangle-males). (b) Fusion computed probability of radiation-induced pneumonitis for cases without pre-RT chemotherapy, separated by gender (asterisk-females, triangle-males).

radiation-induced pneumonitis. This approach, employed in other fields, has been shown to be more robust than the individual component models. The premise is that methodologically different models are likely to differ in their biases and, hence, their fusion can be used to reduce overall bias. This was seen in Fig. 1, where the ROC area remained fairly constant for increasing numbers of randomly selected predictions fused from each of the four models, while the corresponding error bars (95% confidence intervals) progressively decreased. Both the mean ROC area and error bars were generally superior to the individual component models. Fusing all 100 predictions (obtained from 100 random divisions of the data into training and test sets) from each of the four models yielded an ROC area of 0.79. This synergistic improvement in prediction accuracy has been noted in several other fields.^{25,27–32}

In this work, the synergy is likely due to the methodologically different approaches of the component classifiers, i.e., the classifiers provide predictions that have some degree of complementary information or different biases. This is seen in Table II, where the mean interclassifier Spearman correlations were generally low, ranging from 0.19 to 0.63. By contrast, intraclassifier correlations were high (≥ 0.90), with the

FIG. 5. Fitted logistic probability of radiation-induced pneumonitis for males/females without pre-RT chemotherapy (1-long dashes), males with pre-RT chemotherapy (2-short dashes), and females with pre-RT chemotherapy (3-solid line). (a) The fits are superimposed on the fusion computed probability for patients with (x) and without (square) pre-RT chemotherapy. (b) The fits are superimposed on the fusion computed probability for patients with (solid circle) and without (open circle) pneumonitis.

exception of the SVM algorithm. The high intraclass correlation for Neural Networks, Decision Trees, and SOM seems to imply that the 100 different predictions selected very similar features, within each model. Conversely, the lower intraclass correlation for SVM seems to imply that the features selected in the 100 predictions were less similar.

The groupings of features extracted as most important in the fusion, shown in Table III, are to be contrasted against Table IV, which attempts to compile findings in literature on radiation-induced pneumonitis. Note that the patient groups analyzed by the citations in Table IV are not necessarily mutually exclusive. Our work extracted the correlated grouping of lung EUDs for $a=1.2$ to 3 as most important. Since EUD for $a=1$ is equivalent to the mean lung dose (MLD), this grouping corresponds to doses higher than MLD. Several studies have identified MLD as predictive for radiation pneumonitis.^{1,3,5,6,9–14,17,41} The correlated grouping of EUD for $a=0.5–1.2$ and lung volumes $>20–30$ Gy was less predictive. This latter correlated group includes the MLD and doses below MLD, together with the lung volume range $V_{20}–V_{30}$. Both V_{20} and V_{30} were frequently selected by other

FIG. 6. Fitted logistic probability of radiation-induced pneumonitis (see Fig. 5), with 95% confidence intervals for the patient predictions [males/females without pre-RT chemotherapy (1-long dashes), males with pre-RT chemotherapy (2-short dashes), and females with pre-RT chemotherapy (3-solid line)].

studies.^{1–6,8–10,12,42} The greater emphasis given by the fusion to the higher dose group, higher than MLD and less correlated to volumes above low doses, may be a result of the synergistic effect of the two nondose factors identified in this work: pre-RT chemotherapy and female sex. Our study extracted chemotherapy prior to RT as more important than either of the two previously mentioned dose groups in predicting pneumonitis. Chemotherapy has also been previously implicated in causing radiation pneumonitis.^{5,36,42,43} The fusion also extracted female sex and squamous cell histology as being predictive, though less influential than either chemotherapy or the two dose groups (Table III). Figure 4 appears to suggest that females undergoing pre-RT chemotherapy are more susceptible to radiation injury. Female sex has been implicated in previous work.^{21,41,44} A possible explanation for the greater sensitivity in the female population, alluded to by Robnett *et al.*,²¹ is that radiation pneumonitis is similar to an autoimmune response, which affects women to a greater degree. To the best of our knowledge, squamous cell histology has not been reported. Our fused model did not give significance to two other nondose factors that appear in literature: tumor location,^{12,19} and age.^{8,44} Mean heart dose, which has been shown to be an important factor affecting pulmonary function in a rat study,²² was not selected by any of the component models in the fusion. Previous work on a subset of the database analyzed here also noted the correlation between incidental cardiac irradiation and DLCO,⁴⁵ but the fusion did not select either heart dose or pulmonary function tests as predictors of pneumonitis. The fusion indicates that the combination of EUD for $a=1.2$ to 3 and chemotherapy prior to RT is particularly effective at predicting radiation-induced pneumonitis.

While the fusion model is promising for prospective use, it does not directly produce a generally usable closed-form equation. However, the fusion does provide important information that may be used towards generating such an equation: the outcome prediction for each patient as well as the

TABLE IV. Radiation pneumonitis factors in literature. Note that patient groups analyzed in these citations are not necessarily mutually exclusive.

Predictive factor	Study
Lung volume >5 Gy	Yorke (Ref. 11), Wang (Ref. 41)
Lung volume >10 Gy	Schallenkamp (Ref. 9), Yorke (Ref. 11)
Lung volume >13 Gy	Schallenkamp (Ref. 9), Yorke (Ref. 11), Lind (Ref. 8), Kong (Ref. 1), Seppenwoolde (Ref. 17)
Lung volume >15 Gy	Tsujino (Ref. 2), Schallenkamp (Ref. 9)
Lung volume >20 Gy	Tsujino (Ref. 2), Chang (Ref. 3), Tsujino (Ref. 4), Rancati (Ref. 5), Jenkins (Ref. 6), Graham (Ref. 12), Lind (Ref. 8), Schallenkamp (Ref. 9), Seppenwoolde (Ref. 42), Kong (Ref. 1)
Lung volume >30 Gy	Tsujino (Ref. 2), Rancati (Ref. 5), Hernando (Ref. 10), Lind (Ref. 8)
Lung volume >40 Gy	Rancati (Ref. 5)
Mean lung dose (MLD)	Kong (Ref. 1), Chang (Ref. 3), Graham (Ref. 12), Jenkins (Ref. 6), Hernando (Ref. 10), Martel (Ref. 13), Kwa (Ref. 14), Wang (Ref. 41), Yorke (Ref. 11), Schallenkamp (Ref. 9), Rancati (Ref. 5), Seppenwoolde (Ref. 17)
Lower lobe/inferior tumor location	Graham (Ref. 12), Hope (Ref. 19), Yorke (Ref. 11), Seppenwoolde (Ref. 55)
Age	Lind (Ref. 8), Hartsell (Ref. 44)
Chemotherapy	Wang (Ref. 43), Mao (Ref. 3), Seppenwoolde (Ref. 42), Rancati (Ref. 5)
Female sex	Robnett (Ref. 21), Wang (Ref. 41), Hartsell (Ref. 44)

most important features. We have used this information as the “ground truth,” in place of the binary clinical outcome of pneumonitis/no pneumonitis, to fit a logistic probability function. While this function may be prospectively used, its application is limited to the scope of treatments used in our database.

V. CONCLUSIONS

Predicting radiation-induced pneumonitis by fusing the decisions of four different models (Neural Network, Decision Tree, Support Vector Machine, Self-Organizing Map) appears to produce results that are generally more robust (greater separation between cases with and without pneumonitis, lower variance) than any of the individual models. For physical interpretation and prospective use, the fusion output was fitted to a logistic probability function.

ACKNOWLEDGMENTS

This work was supported by grants NIH R01 CA 115748 and NIH R01 CA69579.

APPENDIX: DESCRIPTION OF MODELS

This appendix summarizes the methods used to build the four models (Neural Networks, Decision Trees, Support Vector machines, Self-Organizing Maps), extracted from our previous work.^{33–35,39}

I. Feed-forward neural network

Feed-forward neural networks is a widely used statistical model that is capable of deducing linear or nonlinear input/

output relationships. An example neural network is shown in Fig. 7. The nodes in the first layer (input layer) take in the inputs X ($X_1, X_2, \dots$) and nodes in the last layer (output layer) produce the output $f(X)$. Each node in the hidden and output layer sums (Σ) over all the weighted inputs, which is then passed through an activation function (σ) that outputs a response. The activation function used in our implementation is sigmoid: $\sigma(v) = 1 / (1 + e^{-v})$,⁴⁰ where v is the input and σ is the response.

FIG. 7. Typical architecture of a three-layer feed-forward neural network. Σ indicates summation over the bias and weighted inputs, and σ indicates a linear or nonlinear activation function.

Given the data $\langle(x_1, y_1) \dots (x_N, y_N)\rangle$, the neural network was trained by adjusting weights and biases to minimize the sum-of-squared errors, which was denoted by

$$R = \sum_{i=1}^N (y_i - f(x_i))^2, \quad (A1)$$

where i labels the N training cases, x_i is the vector of input features, and y_i is 1 (RP) or 0 (no RP). The error R was minimized by the backpropagation iterative learning procedure.⁴⁶ The weights and biases at the $(r+1)$ st iteration were updated as follows:

$$\beta^{(r+1)} = \beta^{(r)} + m(\beta^{(r)} - \beta^{(r-1)}) - (1-m)\gamma \frac{\partial R}{\partial \beta^{(r)}}, \quad (A2)$$

where β denotes weights or biases, γ is the learning rate, and m is a momentum parameter^{47,48} term used to accelerate learning. The learning rate γ and momentum parameter m were adaptively adjusted based on the change of error with iterations.

To avoid converging to the local minimum of the error function,⁴⁰ the network was trained multiple times with different initializations of weights and biases. The solution with the lowest error was chosen. At the global minimum of the training error R , the neural network can potentially detrimentally overfit the training data (overtraining). To avoid overfitting, an early stopping strategy was used.⁴⁰ This strategy stops network training before convergence to the global minimum. In this technique, patients with and without RP were randomly split into ten groups of approximately equal size. Eight groups of data were used as the training-construction data set to select input features and optimize the network weights and biases (network training). One group of data was used as the training-validation data set to monitor the evolution of a realistic estimate of error with training iterations and, thereby, signal early stopping in the training. Together, the training-construction and training-validation sets constitute the training set. During network training, the training-construction error can be expected to continuously decrease with iterations, whereas the training-validation error can be expected to initially decrease and then increase when the network starts overfitting.

The network was sequentially built, starting with an initial network consisting of one input node, a hidden layer with three nodes, and one output node. During the course of network building, three major steps were followed in sequence: pruning (removal of an input/hidden layer node), substitution (replacement of an input feature with another input feature), and addition (addition of an input/hidden node).

II. Decision trees

The model is composed of a sequence of weighted individual decision prediction units (indicated by U 's in Fig. 8). Each prediction unit in the sequence was built using patient weights that were computed based on the predictive capability of the previous unit in the sequence. The first prediction unit in the sequence was built with equal patient weights.

FIG. 8. Schematic of boosted model. The prediction units (U) are weighted (w) and summed to create the model. Each prediction unit, with the exception of the first, combines the LNTCP and a decision tree; the first prediction unit is composed of only the LNTCP. The output of the model is a value reflecting the extent of injury predicted—a higher value implies greater injury.

The process of changing patient weights for each prediction unit in the sequence, and computing the weight associated with each prediction unit, used the AdaBoost algorithm.⁴⁹ The AdaBoost algorithm attempts to sequentially increase model predictive capability (“boosting”) by adding on each successive predictive unit. In our model, each unit (with the exception of the first unit) was the sum of the Lyman NTCP (LNTCP) metric⁵⁰ and a decision tree; the first unit was composed of only the LNTCP metric. This choice of composition of the decision prediction unit was motivated by the rationale that the nonparametric decision tree component would augment the predictive capability of the more rigidly parametric, but widely used, LNTCP metric. Thus, the decision tree model, which is capable of using dose and nondose variables as inputs, would compensate for shortcomings in the LNTCP metric, which uses only dose in a parametric formulation.

The trees constituting the prediction units in the AdaBoost algorithm were built originating at a root node, which split the patient population into two branches with the aim of maximally separating the two classes according to the Gini index split threshold criterion.⁴⁰ The split was considered successful if there was any reduction in the Gini index postsplit, with respect to the Gini index prior to splitting. This is equivalent to a positive value for the expression below

$$N_S(1 - p_{inj,S}^2 - p_{uninj,S}^2) - N_L(1 - p_{inj,L}^2 - p_{uninj,L}^2) - N_R(1 - p_{inj,R}^2 - p_{uninj,R}^2) \quad (A3)$$

where the subscripts inj and uninj refer to the injured and uninjured patients, respectively; subscripts S, L, R , refer to the node being split, left and right branches, respectively; p and N are the proportion of patients and number of patients, respectively. The three bracketed terms in the expression above are the Gini indexes for the node being split, left node and right node, respectively. The left and right branches were designated as predicting for injury and no injury, respectively. The nodes at the ends of the branches were, in turn, split. A node that could not be split was designated as termi-

FIG. 9. Support vector machines for (a) separable and (b) nonseparable cases. The solid line is the optimal hyperplane separating the two categories. The dotted lines are margin edges with maximal width $2/\|\beta\|$. Points indicated by arrows are located within the margin or on the wrong side of the hyperplane (distance ξ'_i from the margin edge). These points determine the separation hyperplane and are hence termed “support vectors.”

nal. When all nodes were terminal, tree building was deemed complete. A patient that travels down the tree is predicted to incur a probability of injury corresponding to the terminal node at which the patient stops. This probability of injury is proportional to the number of training set injured patients that have stopped at the same terminal node. By virtue of the left and right branch designations, left branch terminal nodes have a higher probability of injury than their corresponding right branch terminal nodes.

To reduce the possibility of overfitting, not more than three node splits were allowed for each component decision tree. Also, certain rules were implemented to prevent the model from splitting variables in an “unrealistic” manner, for example, higher lung volume above a given dose predicting for reduced incidence of injury. These rules were set as a directional parameter during the tree building process: a direction of +1 for a variable indicated that the variable had to be above a split threshold to predict for injury, a direction of -1 indicated that the variable had to be below a split threshold to predict for injury, and a direction of 0 indicated that values either above or below a split threshold could be used to predict for injury. All dose variables and stage were assigned a direction of +1. Pre-RT FEV 1%, pre-RT DLCO, and pre-RT DLCO% were assigned a direction of -1. All other variables were assigned a direction of 0.

Model predictive capability was realistically evaluated via tenfold crossvalidation.⁴⁰ In this procedure, patients were randomly divided into ten groups of approximately equal size (each group contained approximately 1/10 of the injured and uninjured patients). The model was built using data from nine groups of patients, and then tested on the remaining group. Each group, in turn, served as the test group.

III. Support vector machine

The general idea behind support vector machines is to compute an optimal hypersurface (boundary) that maximizes the margin between data points in categories +1 (RP) and -1 (no RP). Figure 9 illustrates this concept using a simplified hyperplane. In Fig. 9, the hyperplane is denoted by $x^T\beta + \beta_0 = 0$, where β is the normal vector to the hyperplane, and

β_0 is an offset parameter. The two parallel margin edges are $x^T\beta + \beta_0 = \pm 1$, corresponding to a margin width of $2/\|\beta\|$. Ideally, the space within the margin is devoid of data points. In reality, Support Vector Machines maximize the margin [distance between dotted lines in Fig. 9(a)], while permitting some transgression of points into the margin area or the wrong side of the boundary [Fig. 9(b)]. In Fig. 9(b), the slack variable $\xi'_i > 0$ indicates that the point i has transgressed into the margin, and $\xi'_i > 1/\|\beta\|$ indicates that it is on the wrong side of the separation hyperplane. The sum total of transgressions is restricted to be below a certain limit, thereby allowing small transgressions from several points or, alternately, large “outlier” transgressions from a few points.

Mathematically, the concept of Support Vector Machines is described as

$$\max \frac{2}{\|\beta\|} \text{ subject to } \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\ \xi_i \geq 0, \sum \xi_i \leq \text{const}, \end{cases} \quad (\text{A4})$$

where $\xi_i = \|\beta\| \xi'_i$ [see Fig. 9(b)], (x_i, y_i) is the data point for patient i , with $y_i = \pm 1$ (+1: RP; -1: no RP), and vector x_i contains the selected patient features.

The solution of Eq. (A4) is equivalent to solving a quadratic programming problem⁴⁰ in which

$$\min_{\alpha_i} \left(\frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^N \alpha_i \right), \quad (\text{A5})$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^N \alpha_i y_i = 0$. Parameter C is user-defined. The coefficients α_i are computed via optimization.

The Support Vector Machine shown in Fig. 9 has a linear boundary. However, for cases with nonlinear boundaries, the original low-dimensional input space can be translated to a higher-dimensional feature space via a basis function.⁴⁰ The basis function does not explicitly appear in the calculation if $x_i^T x_j$ [Eq. (3)] is replaced by an appropriate kernel function $K(x_i, x_j)$.⁴⁰ For our modeling, the radial basis function, a popular kernel function in Support Vector Machine literature,^{40,51,52} was used,

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2), \quad (\text{A6})$$

where σ is a user-defined parameter. Parameters C [Eq. (3)] and σ [Eq. (6)] were determined prior to cross validation using grid search⁵³ and ninefold evaluation within each training set.

Input features were selected from the list of available variables as follows, within the framework of tenfold cross validation (nine groups were used for training and the remaining group was used for testing, with each group taking turns as the test group). For each variable that was a potential input feature, the SVM was trained using eight of the nine training groups and then evaluated on the one remaining training-evaluation group. Each of the nine training groups served as the training-evaluation group, in turn. The variable was added as input feature if the area under the collective ROC curve of the nine training-evaluation groups increased. Similarly, an already selected input feature was replaced by another variable if the training-evaluation ROC area in-

FIG. 10. An example of SOM grid topology: 4×3 array of 12 neurons. Each neuron contains two part of information: physical location and the reference vector.

creased. SVM construction was stopped if no new variable was accepted as input feature, after all unselected variables were evaluated through addition/substitution.

IV. Self-organizing map

Self-organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning.⁵⁴ The basic idea behind SOM is to construct a nonlinear transformation to project high dimensional input data onto a low, often two-dimensional space, termed a feature map (see Fig. 10). For our modeling, similar patient input vectors in the higher dimensional input space were clustered into the same region or neuron of the feature map. The entries in the input vector, termed features, are a subset of all available factors. For example, if the input features constituting the input vector are “ a ” and “ b ,” patients with similar values of a and b were clustered into the same region (neuron). Each neuron in the map contains two parts of information. The first part is its relative physical location, i.e., proximity to other neurons, and the second part is its reference vector or weight vector, which can be thought of as the “typical” feature values for that neuron.

During the course of SOM training, patients were assigned to neurons with reference vectors most similar to the patient input vector. Following each such patient assignment, the reference vector of the assigned neuron and its immediate neighboring neurons were updated to reflect the change to the patient population at the neuron. The neighboring neurons were identified as neurons whose distances to the assigned neuron were less than the neighborhood distance (nd). The neighborhood distance (nd) decreased from a large number to 0 during training, such that the neighborhood neurons included all neurons at the beginning and only the assigned neuron at the end. The learning rate (lr) determined the amount of information passed from the patient input vector to the assigned neuron, which decreased from 1 to 0 during training. Similarity was judged by the Euclidean distance between a patient input vector and the neuron reference vector (a smaller distance implies greater similarity). Since

SOM is an unsupervised learning method, patients were assigned to each neuron without knowledge of their toxicity classification (RP vs no RP).

Input features were selected by a trial addition/substitution process, as follows. Using tenfold cross validation, nine groups of data were used to train the SOM. To optimally select the input features, the SOM was trained (MATLAB, The MathWorks, Inc., Natick, MA) using eight of the nine training groups and then evaluated on the remaining training group by ROC analysis. The eight training groups used to build the SOM were collectively termed the training-construction set, and the one training group used for evaluation was termed the training-evaluation set. This evaluation process was internal to the training group and only served to build the list of selected features, i.e., it was not used as an unbiased test of SOM performance. The added or substituted factor was accepted as an input feature if the area under the training-evaluation ROC curve increased. SOM construction was stopped if no new factor was accepted as an input feature.

^{a)} Author to whom correspondence should be addressed. Electronic mail: shiva.das@duke.edu; Telephone: (919) 681-5424; Fax: (919) 681-7183.

¹F.-M. Kong *et al.*, “Final toxicity results of a radiation-dose escalation study in patients with non-small-cell lung cancer (nsccl): Predictors for radiation pneumonitis and fibrosis,” *Int. J. Radiat. Oncol., Biol., Phys.* **65**(4), 1075–1086 (2006).

²K. Tsujino *et al.*, “Radiation pneumonitis following concurrent accelerated hyperfractionated radiotherapy and chemotherapy for limited-stage small-cell lung cancer: Dose-volume histogram analysis and comparison with conventional chemoradiation,” *Int. J. Radiat. Oncol., Biol., Phys.* **64**(4), 1100–1105 (2006).

³D. T. Chang *et al.*, “The impact of heterogeneity correction on dosimetric parameters that predict for radiation pneumonitis,” *Int. J. Radiat. Oncol., Biol., Phys.* **65**(1), 125–131 (2006).

⁴K. Tsujino *et al.*, “Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer,” *Int. J. Radiat. Oncol., Biol., Phys.* **55**(1), 110–115 (2003).

⁵T. Rancati *et al.*, “Factors predicting radiation pneumonitis in lung cancer patients: A retrospective study,” *Radiother. Oncol.* **67**(3), 275–283 (2003).

⁶P. Jenkins *et al.*, “Radiation pneumonitis following treatment of non-small-cell lung cancer with continuous hyperfractionated accelerated radiotherapy (CHART),” *Int. J. Radiat. Oncol., Biol., Phys.* **56**(2), 360–366 (2003).

⁷V. Moiseenko *et al.*, “Dose-volume analysis of lung complications in the radiation treatment of malignant thymoma: A retrospective review,” *Radiother. Oncol.* **67**, 265–274 (2003).

⁸P. A. Lind *et al.*, “ROC curves and evaluation of radiation-induced pulmonary toxicity in breast cancer,” *Int. J. Radiat. Oncol., Biol., Phys.* **64**(3), 765–770 (2006).

⁹J. M. Schallenkamp *et al.*, “Incidence of radiation pneumonitis after thoracic irradiation: Dose-volume correlates,” *Int. J. Radiat. Oncol., Biol., Phys.* **67**(2), 410–416 (2007).

¹⁰M. L. Hernando *et al.*, “Radiation-induced pulmonary toxicity: A dose-volume histogram analysis in 201 patients with lung cancer,” *Int. J. Radiat. Oncol., Biol., Phys.* **51**(3), 650–659 (2001).

¹¹E. D. Yorke *et al.*, “Correlation of dosimetric factors and radiation pneumonitis for non-small-cell lung cancer patients in a recently completed dose escalation study,” *Int. J. Radiat. Oncol., Biol., Phys.* **63**(3), 672–682 (2005).

¹²M. V. Graham *et al.*, “Clinical dose-volume histogram analysis for pneumonitis after 3D treatment for non-small cell lung cancer (NSCLC),” *Int. J. Radiat. Oncol., Biol., Phys.* **45**(2), 323–329 (1999).

¹³M. K. Martel *et al.*, “Dose-volume histogram and 3-D treatment planning evaluation of patients with pneumonitis,” *Int. J. Radiat. Oncol., Biol., Phys.* **28**(3), 575–581 (1994).

- ¹⁴S. L. S. Kwa *et al.*, "Radiation pneumonitis as a function of mean lung dose: An analysis of pooled data of 540 patients," *Int. J. Radiat. Oncol., Biol., Phys.* **42**(1), 1–9 (1998).
- ¹⁵J. C. M. Theuws *et al.*, "Prediction of over-all pulmonary function loss in relation to the 3-D dose distribution for patients with breast cancer and malignant lymphoma," *Radiother. Oncol.* **49**(3), 233–243 (1998).
- ¹⁶J. C. M. Theuws *et al.*, "Dose-effect relations for early local pulmonary injury after irradiation for malignant lymphoma and breast cancer," *Radiother. Oncol.* **48**(1), 33–43 (1998).
- ¹⁷Y. Seppenwoolde *et al.*, "Comparing different NTCP models that predict the incidence of radiation pneumonitis," *Int. J. Radiat. Oncol., Biol., Phys.* **55**(3), 724–735 (2003).
- ¹⁸L. B. Marks *et al.*, "Physical and biological predictors of changes in whole-lung function following thoracic irradiation," *Int. J. Radiat. Oncol., Biol., Phys.* **39**(3), 563–570 (1997).
- ¹⁹A. J. Hope *et al.*, "Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters," *Int. J. Radiat. Oncol., Biol., Phys.* **65**(1), 112–124 (2006).
- ²⁰S. McDonald *et al.*, "Injury to the lung from cancer-therapy-clinical syndromes, measurable end-points, and potential scoring systems," *Int. J. Radiat. Oncol., Biol., Phys.* **31**(5), 1187–1203 (1995).
- ²¹T. J. Robnett *et al.*, "Factors predicting severe radiation pneumonitis in patients receiving definitive chemoradiation for lung cancer," *Int. J. Radiat. Oncol., Biol., Phys.* **48**(1), 89–94 (2000).
- ²²P. V. Luijk *et al.*, "Sub-clinical heart damage enhances radiation-induced lung function loss," *Int. J. Radiat. Oncol., Biol., Phys.* **63**(2), S460 (2005).
- ²³L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(2), 281–286 (2002).
- ²⁴M. Taniguchi and V. Tresp, "Averaging regularized estimators," *Neural Comput.* **9**(5), 1163–1178 (1997).
- ²⁵J. Kittler *et al.*, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 226–239 (1998).
- ²⁶O. Melnik, Y. Vardi, and C. H. Zhang, "Mixed group ranks: Preference and confidence in classifier combination," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(8), 973–981 (2004).
- ²⁷Y. W. Liao, L. W. Nolte, and L. A. Collins, "Decision fusion of ground-penetrating radar and metal detector algorithms-A robust approach," *IEEE Trans. Geosci. Remote Sens.* **45**(2), 398–409 (2007).
- ²⁸R. A. Heckemann *et al.*, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *Neuroimage* **33**(1), 115–126 (2006).
- ²⁹M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Decision fusion for the classification of urban remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **44**(10), 2828–2838 (2006).
- ³⁰S. Tilie, I. Bloch, and L. Laborelli, "Fusion of complementary detectors for improving blotch detection in digitized films," *Pattern Recogn. Lett.* **28**(13), 1735–1746 (2007).
- ³¹G. Niu *et al.*, "Multi-agent decision fusion for motor fault diagnosis," *Mech. Syst. Signal Process.* **21**(3), 1285–1299 (2007).
- ³²S. Pal and D. J. Miller, "An extension of iterative scaling for decision and data aggregation in ensemble classification," *J. VLSI Signal Proc. Syst. Signal, Image, Video Technol.* **48**(1–2), 21–37 (2007).
- ³³S. Chen *et al.*, "A neural network model to predict lung radiation-induced pneumonitis," *Med. Phys.* **34**(9), 3420–3427 (2007).
- ³⁴S. Chen *et al.*, "Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis," *Med. Phys.* **34**(10), 3808–3814 (2007).
- ³⁵S. K. Das *et al.*, "Predicting lung radiotherapy-induced pneumonitis using a model combining parametric lyman probit with nonparametric decision trees," *Int. J. Radiat. Oncol., Biol., Phys.* **68**(4), 1212–1221 (2007).
- ³⁶J. F. Mao *et al.*, "The impact of induction chemotherapy and the associated tumor response on subsequent radiation-related changes in lung function and tumor response," *Int. J. Radiat. Oncol., Biol., Phys.* **67**(5), 1360–1369 (2007).
- ³⁷Z. Kocak *et al.*, "Prospective assessment of dosimetric/physiologic-based models for predicting radiation pneumonitis," *Int. J. Radiat. Oncol., Biol., Phys.* **67**(1), 178–186 (2007).
- ³⁸Z. Kocak *et al.*, "Challenges in defining radiation pneumonitis in patients with lung cancer," *Int. J. Radiat. Oncol., Biol., Phys.* **62**(3), 635–638 (2005).
- ³⁹S. F. Chen *et al.*, "Using patient data similarities to predict radiation pneumonitis via a self-organizing map," *Phys. Med. Biol.* **53**(1), 203–216 (2008).
- ⁴⁰T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical learning: Data Mining, Inference, and Prediction* (Springer-Verlag, New York, 2002).
- ⁴¹S. L. Wang *et al.*, "Investigation of clinical and dosimetric factors associated with postoperative pulmonary complications in esophageal cancer patients treated with concurrent chemoradiotherapy followed by surgery," *Int. J. Radiat. Oncol., Biol., Phys.* **64**(3), 692–699 (2006).
- ⁴²Y. Seppenwoolde, K. De Jaeger, and J. V. Lebesque, "In regard to Tsujino *et al.*: Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer," *IJROBP* **2003**;55: 110–115," *Int. J. Radiat. Oncol., Biol., Phys.* **56**(4), 1208–1209 (2003).
- ⁴³S. L. Wang *et al.*, "Association between systemic chemotherapy before chemoradiation and increased risk of treatment-related pneumonitis in esophageal cancer patients treated with definitive chemoradiotherapy," *J. Thorac. Oncol.* **3**(3), 277–282 (2008).
- ⁴⁴W. F. Hartsell *et al.*, "Can serum markers be used to predict acute and late toxicity in patients with lung cancer? Analysis of RTOG 91-03," *Cancer Clin. Trials* **30**(4), 368–376 (2007).
- ⁴⁵J. Zeng *et al.*, "Impact of incidental cardiac irradiation on the development of shortness of breath and changes in pulmonary function tests in patients receiving radiation for lung cancer," *Int. J. Radiat. Oncol., Biol., Phys.* **66**(3), S157–S158 (2006).
- ⁴⁶D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature (London)* **323**(6088), 533–536 (1986).
- ⁴⁷T. P. Vogl *et al.*, "Accelerating the convergence of the back-propagation method," *Biol. Cybern.* **59**(4–5), 257–263 (1988).
- ⁴⁸R. A. Jacobs, "Increased rates of convergence through learning rate adaptation," *Neural Networks* **1**(4), 295–307 (1988).
- ⁴⁹Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.* **55**(1), 119–139 (1997).
- ⁵⁰J. T. Lyman, "Complication probability as assessed from dose volume histograms," *Radiat. Res.* **104**(2), S13–S19 (1985).
- ⁵¹V. Vapnik, *The Nature of Statistical Learning Theory* (Springer-Verlag, New York, 1995).
- ⁵²C. J. C. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Data Min. Knowl. Discov.* **2**(2), 121–167 (1998).
- ⁵³S. S. Keerthi and C. J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.* **15**(7), 1667–1689 (2003).
- ⁵⁴T. Kohonen, *Self-Organizing Maps* (Springer-Verlag, New York, 1995).
- ⁵⁵Y. Seppenwoolde *et al.*, "Regional differences in lung radiosensitivity after radiotherapy for non-small-cell lung cancer," *Int. J. Radiat. Oncol., Biol., Phys.* **60**(3), 748–758 (2004).