SCIENTIFIC REPORTS
nature research

OPEN

A Super-Learner Model for Tumor Motion Prediction and Management in Radiation Therapy: Development and Feasibility Evaluation

Hui Lin, Wei Zou, Taoran Li, Steven J. Feigenberg, Boon-Keng K. Teo & Lei Dong*

In cancer radiation therapy, large tumor motion due to respiration can lead to uncertainties in tumor target delineation and treatment delivery, thus making active motion management an essential step in thoracic and abdominal tumor treatment. In current practice, patients with tumor motion may be required to receive two sets of CT scans – the initial free-breathing 4-dimensional CT (4DCT) scan for tumor motion estimation and a second CT scan under appropriate motion management such as breath-hold or abdominal compression. The aim of this study is to assess the feasibility of a predictive model for tumor motion estimation in three-dimensional space based on machine learning algorithms. The model was developed based on sixteen imaging features extracted from non-4D diagnostic CT images and eleven clinical features extracted from the Electronic Health Record (EHR) database of 150 patients to characterize the lung tumor motion. A super-learner model was trained to combine four base machine learning models including the Random Forest, Multi-Layer Perceptron, LightGBM and XGBoost, the hyper-parameters of which were also optimized to obtain the best performance. The outputs of the super-learner model consist of tumor motion predictions in the Superior-Inferior (SI), Anterior-Posterior (AP) and Left-Right (LR) directions, and were compared against tumor motions measured in the free-breathing 4DCT scans. The accuracy of predictions was evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) through ten rounds of independent tests. The MAE and RMSE of predictions in the SI direction were 1.23 mm and 1.70 mm; the MAE and RMSE of predictions in the AP direction were 0.81 mm and 1.19 mm, and the MAE and RMSE of predictions in the LR direction were 0.70 mm and 0.95 mm. In addition, the relative feature importance analysis demonstrated that the imaging features are of great importance in the tumor motion prediction compared to the clinical features. Our findings indicate that a super-learner model can accurately predict tumor motion ranges as measured in the 4DCT, and could provide a machine learning framework to assist radiation oncologists in determining the active motion management strategy for patients with large tumor motion.

Respiratory motion poses a great challenge in the treatment of lung cancer with radiation therapy[1–3]. Target and normal tissue motion can be quite complex and patient-dependent[4]. To address this issue in modern radiation therapy treatment planning, an internal margin is assigned based on the patient's 4-dimensional Computed Tomography (4DCT) to form an Internal Target Volume (ITV)[5,6], where the extent of the tumor motion is included. However, when the tumor motion extent is large, the ITV may possess prohibitively large volume that could lead to increased treatment toxicity. Therefore, patients with a large tumor motion are more suitable for using an active motion management strategy such as a breath-hold technique[7] or the use of compression belt[8] to reduce the magnitude of tumor motions due to diaphragmatic breathing[9]. These active motion management procedures, however, require extra steps in the simulation workflow since the need for an additional simulation

University of Pennsylvania, Department of Radiation Oncology, Pennsylvania, 19104, United States. *email: Lei.Dong@uphs.upenn.edu

CT scan with motion management is known only after the motion was assessed from an initial free-breathing 4DCT scan. The motivation of this work is to test the feasibility of predicting tumor motion without an initial free-breathing 4DCT scan such that these patients can be directly identified and undergo an active motion management strategy for simulation. If successful, the motion prediction can avoid extra radiation dose to the patient due to additional CT imaging. In addition, streamlined active motion management without the initial 4DCT motion assessment can be tolerated by most patients who might be too tired or too sick to perform two 4DCT procedures in the same day and increase the quality of the scan. This also potentially reduces staff and equipment resource needed for the simulation.

Previously, there were extensive studies for the characterization and prediction of lung tumor motion, such as the prediction of real-time tumor motion tracking during treatment[10–13], the characterization of tumor motions to estimate severity of 4DCT motion artifact[14,15], or the detection of inter-fraction motion pattern change[16]. The input of these studies was 4DCT images of patients or motion trace measured by fiducial markers. In comparison, the aim of this study is to develop a motion prediction system prior to any patient motion measurement. To achieve that, non-motion specific features such as those presented in the diagnostic CT and EHR data are used as the input. Due to potential complex relationship of these input parameters, a machine learning approach is proposed in this study to improve prediction accuracy. There were previous attempts trying to predict tumor motions based on clinical features. For example, Liu *et al.*[4] have shown the magnitude of tumor motion in the SI direction is correlated with the lobe location where the tumor resides; however, they failed to provide accurate predictions using this single-factor correlation approach. Considering the complexity of lung motion behavior and patient-specific factors, most of the previous studies only explored limited features. This work investigates the comprehensive clinical and imaging features to build a predictive model with a goal to identify those patients who may present a large range of tumor motion and require an active motion management strategy for their subsequent radiation therapy.

In this study, we first propose a machine learning approach to investigate the inherent correlations between input features (imaging and clinical features) and tumor motion ranges. In the second step, we developed a super-learner model that employed the input features to predict the three-dimensional lung tumor motion ranges. The development and validation of the proposed model were demonstrated in an extensive patient database of 150 lung patients.

Results

Feature selection. The selected features with top averaged F-scores after Recursive Feature Elimination (RFE) and collinearity removal in each direction are shown in Fig. 1. For the SI-direction, 12 features out of 27 features were selected; in the AP direction, 17 features out of 27 features were selected; and for the LR-direction, 17 features out of 27 features were selected. Overall, the imaging features demonstrated a higher degree of importance in motion prediction than the clinical features. Among imaging features, the tumor centroid or edge location relative to the boundaries of lungs (chest wall or the apex of the lung), lung dimensions, and the lung volume were the top selected features. Among clinical features, the patient's weight, age, and pack years of smoking were the most frequently selected features. During the ten rounds of independent tests, certain features consistently showed greater correlations with the tumor motion in each direction, and they are: Lung dimension (SI) and Tumor centroid distance to the lung apex (SI) for the SI direction; Distance of tumor edge to the chest wall (AP) and Contralateral lung volume for the AP direction; Distance of tumor centroid to the chest wall (LR) and GTV density for the LR direction.

Hyper-parameter tuning. All the machine learning base models utilized in this study contain several hyper-parameters that can affect performance significantly. Our experimental results allowed us to measure the extent to which hyper-parameter tuning via Bayesian optimization improved each machine learning base model's performance compared to its baseline settings. Figure 2 compared the MAE improvements of the tuned parameters to its default settings for each base model across the dataset in all three directions. The results demonstrated why it is unwise to use default ML algorithm hyper-parameters: hyper-parameter tuning improves the model's predictive MAE by ~2–11%.

Super-learner model. The MAE improvements of the super-learner model in five-fold cross-validation are illustrated in Fig. 2 by comparing to the MAE of each base model with optimized hyper-parameters. As expected, the super-learner model outperformed each individual base model. The results demonstrated that the use of the super-learner model leads to approximately 4–40% decrease in MAE.

Prediction performance of the super-learner model. The predicted tumor motion values in the SI, AP and LR direction were compared with the ground truth tumor motion ranges and illustrated by Fig. 3(a,c,e). The corresponding residual plots were shown in Fig. 3(b,d,f). A 2-mm margin, which is a typical CT slice resolution used in the patient simulation[17], was applied to evaluate the residual errors. The MAE and RMSE of predictions in the SI direction are 1.23 mm and 1.70 mm respectively; the MAE and RMSE of predictions in the AP direction are 0.81 mm and 1.19 mm respectively; the MAE and RMSE of predictions in the LR direction are 0.70 mm and 0.95 mm respectively. To quantify the outliers, 95th and 99th percentile errors of the predictions were also calculated: in the SI direction, the 95th percentile and 99th percentile error are 2.51 mm and 3.50 mm; in the AP direction, the 95th percentile and 99th percentile error are 2.10 mm and 3.81 mm; in the LR direction, the 95th percentile and 99th percentile error are 1.65 mm and 2.35 mm.

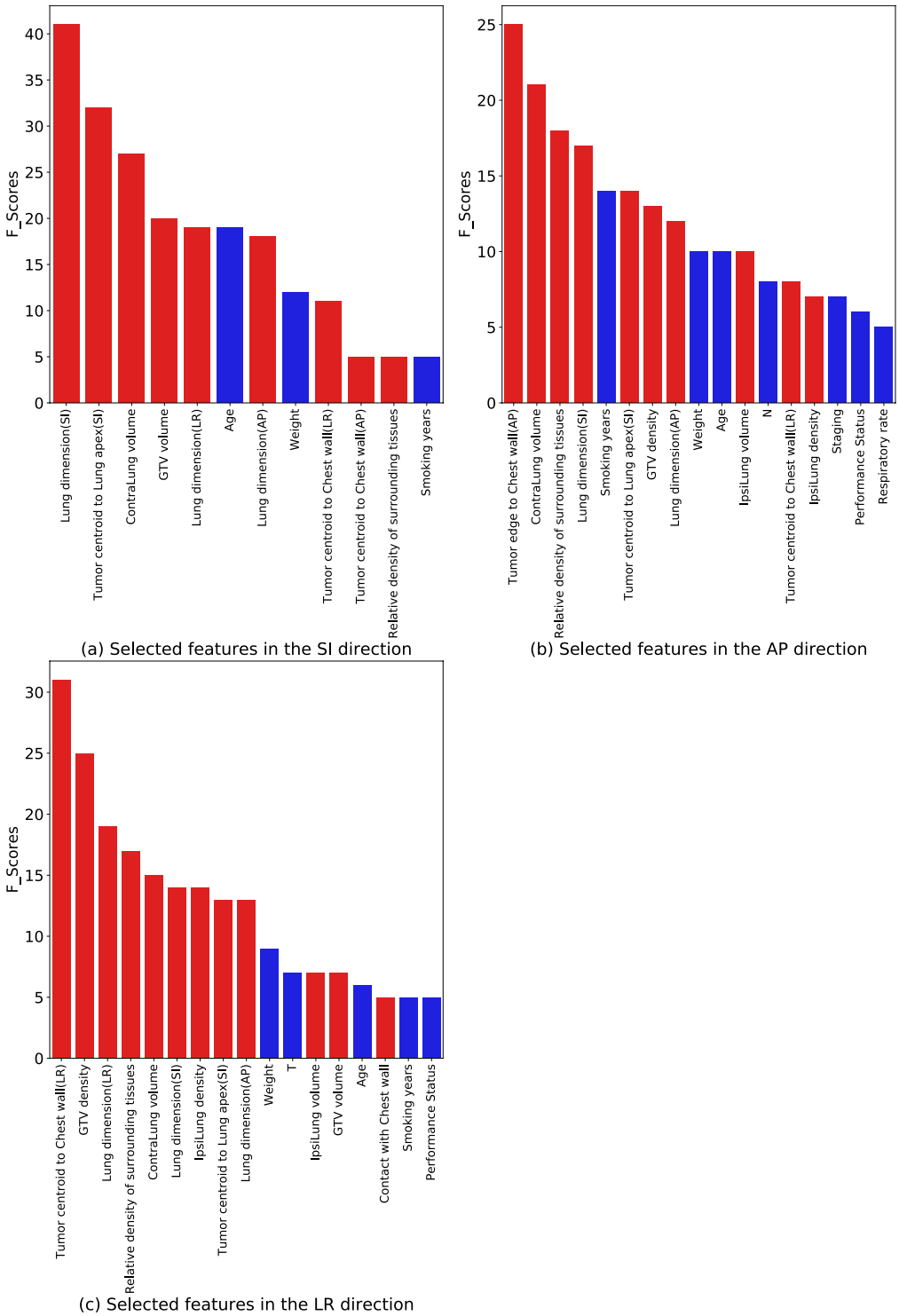

Figure 1. Feature importance ranks in the SI, AP and LR directions obtained by XGBoost RFE and col-linearity removal are shown in (**a**–**c**). Imaging features are plotted in red and clinical features are plotted in blue. The F-scores were averaged over ten rounds of independent tests.

Discussions and Conclusion

In this work we proposed and implemented a machine learning pipeline to investigate the relationship of extensive input features and lung tumor motion ranges, and developed a super-learner model to predict the tumor motion ranges in three dimensions based on the diagnostic CT images and EHR data of the patient. To the best of our knowledge, this is the first study that introduced super-learner models to the tumor motion estimation in radiotherapy, and the built model was validated and tested on one hundred and fifty clinical cases, which is the most extensive study to date.

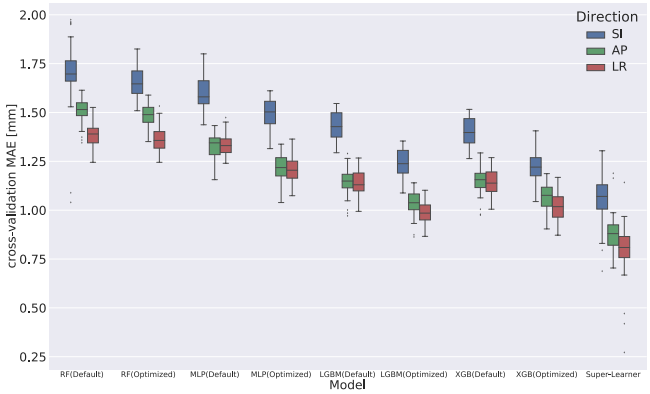

Figure 2. The five-fold cross-validated Mean Absolute Error (MAE) of each machine learning base model with default hyper-parameter settings, with optimized hyper-parameters and the MAE of the Super-Learner model. The benefit of hyper-parameter tuning can be demonstrated by comparing the MAE of four base models with default hyper-parameters and with optimized hyper-parameters. The base models include Random Forest (RF), Multi-Layer Perceptron (MLP) Networks, LightGBM (LGBM) and XGBoost (XGB). The power of building up super-learner models can be demonstrated by the MAE improvements between the super-learner model and each optimized base model.

The study findings indicated that imaging features were more discriminative than the clinical features for the task of predicting tumor motion ranges in the SI, AP and LR direction. Specifically, the tumor location relative to the boundaries of lungs (the chest wall or the apex of the lung) and the lung dimension in each direction are recognized to have great impacts on the tumor motions, which are new features not identified by previous studies. On the aspect of clinical features, the top selected features include the patient's weight, age and the pack years of smoking. This is consistent with clinical observations, as smoking can decrease the lung expansion and capacity[18], the ventilation of lungs can be altered by the patient weight[19], and the respiratory muscle strength may change along with age[20], subsequently affecting the respiratory pattern of a patient. The features selected by this study can potentially help clinicians to decide which features can be used to estimate the lung tumor motion. The predictive accuracy of the super-learner model suggested the feasibility of utilizing a super-learner model to estimate the tumor motion ranges according to the patient CT images and EHR data, which were available prior to the 4DCT simulation. If the findings in this study can be further extended on a larger scale and reproduced in prospective studies, the super-learner model described here can optimize the current radiotherapy simulation workflow for the lung cancer patients by providing individualized motion management strategy to each patient without the assistance of an initial free-breathing 4DCT scan.

In this study, we found that XGBoost-based Recursive Feature Selection helped reduce the redundancy of inputs by lowering the number of features, thus increased the ratio between the number of input data and the number of features. However, it is necessary to assess this supervised feature selection method with external validation data to avoid over-optimistic predictive performance due to the bias of feature selection. To alleviate the overfitting problem raised by feature selection bias, the XGBoost model used for feature selection used fewer feature sampling and a shallow depth of the tree than the XGBoost model used for prediction. More importantly, the testing data was kept independent of the feature selection process. A common mistake is to involve not only the training data but also the testing data for feature selection, which causes overfitting problem due to the information leakage from the testing data during the feature selection phase. The correct procedures are deriving a subset of features only within the training data, and then inferring the predictive model based upon the selected features. During our testing phase, the pre-trained model was evaluated using independent testing data, and therefore minimized the risk of bias that might be introduced using XGBoost for feature selection.

We note some limitations of the current study that highlight opportunities for future enhancement. The first limitation is the extraction of imaging features. The current approach for extraction of the imaging features was performed manually and is very time-consuming. As these imaging features were demonstrated to be critical to the model buildup and performance, such a manual approach can benefit from the development of an automatic imaging feature extraction tool. In addition, our selection of the imaging features may be subjective and miss some important information. Some studies have indicated the solution to this problem is to introduce deep learning-based automatic feature extractor[21–24]. In the future, we plan to explore the incorporation of a Convolutional Neural Network feature extractor. The addition of automatically extracted imaging features can lead to a better representation of the patient features to obtain a more reliable prediction model.

Another limitation is the handling of tumor motion close to the motion management threshold. Tumor motion predicted around the motion management threshold needs to be taken special care, as any small error in this region can deviate the motion management decision. For example, a tumor motion of 8 mm in the SI direction was predicted to be 7.5 mm. If the motion threshold of using active motion management is also 8 mm, although the absolute predictive error is only 0.5 mm, the prediction will lead to a decision on no motion management as opposed to an active motion management strategy is needed. In future work, we plan to investigate

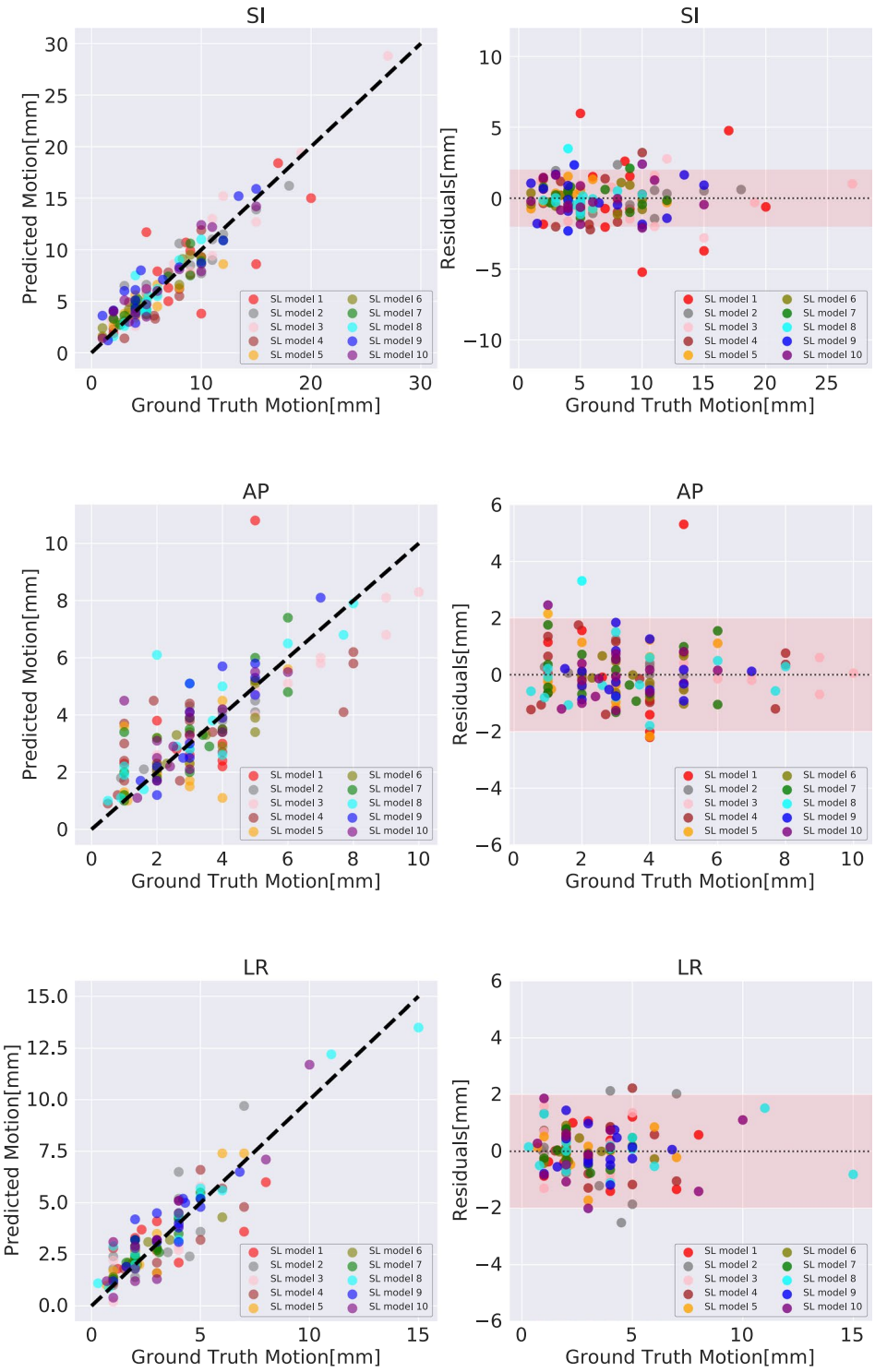

Figure 3. Predicted values of the super-learner models versus ground truth values in the SI, AP and LR directions and the corresponding residual plots. A 2 mm error region is highlighted in each residual plot. The independent ten test set results of each super-learner model are plotted in different colors.

the effects of adding adaptive margins at the threshold. We also plan to quantify and associate an uncertainty level with each tumor motion prediction to further assist the clinical motion management decisions.

Methods and Materials

This section presents the machine learning approach and describes each step of the pipeline implemented to build and evaluate a super-learner model for tumor motion range prediction.

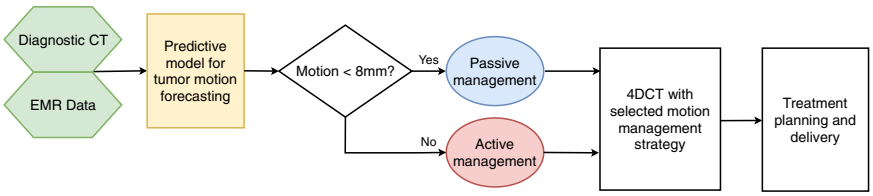

Figure 4. Proposed workflow of performing the automatic selection of motion management strategy prior to the patient simulation.

Imaging features	Clinical features
Tumor centroid and edge locations to the apex of the lung (SI)	Age [yrs]
Tumor centroid location relative to the chest wall (AP, LR)	Weight [Lbs]
Tumor edge location relative to the chest wall (AP, LR)	Respiratory rate
Lung dimension in SI, AP and LR directions [cm]	Smoking history [pack yrs]
Tumor contact with the chest wall (2-classes)	Staging
Target lung volume [cm^3]	Primary tumor (T)
Contralateral lung volume [cm^3]	Regional lymph nodes (N)
Volume of Gross Tumor Volume (GTV) [cm^3]	Distant metastasis (M)
GTV density [HU]	Tumor location: lung-wise (Left/Right)
Density of surrounding tissues around GTV relative to the lung	Tumor location: lobe-wise (Upper/Middle/Lower)
Target lung density [HU]	Performance status

Table 1. Characteristics of input features. The number of features is determined before data pre-processing.

Ethics, consent and permissions. The Institutional Review Board (IRB) of the Hospital of the University of Pennsylvania approved this retrospective patient study (IRB# 831407). All methods used in this study were conducted in accordance with the relevant guidelines and regulations. Considering that this is a retrospective study involving minimal risk to the privacy of the study subjects, our IRB waived the need for obtaining written informed consent from the participants.

Proposed workflow. Figure 4 shows the high-level conceptual framework for the deployment of a super-learner predictive model in the patient simulation workflow. Compared to the current clinical practice, which requires an initial free-breathing 4DCT for tumor motion evaluation, the super-learner model can forecast tumor motion extent by utilizing imaging features extracted from the patient's prior diagnostic CT images and the clinical features extracted from the Electronic Health Record (EHR) data. Based upon the estimated tumor motion extent, the appropriate motion management strategy can be pre-assigned to the patient for the 4DCT simulation to avoid the need for an extra 4DCT scan.

Patient dataset collection and processing. A cohort of one hundred and fifty consecutive lung cancer patients who received proton therapy from 2014 to 2018 was retrospectively identified from the Hospital of the University of Pennsylvania. Two main categories of input features were collected: imaging features extracted from the pre-simulation diagnostic CT images (free-breathing acquisition with 3 mm slice thickness) and clinical features extracted from the EHR data. There are in total twenty-seven input features extracted from the patient data based on clinical observations and literatures[4,25], which are summarized in Table 1. The collected output data consist of tumor motion ranges in the superior/inferior (SI), anterior/posterior (AP) and left/right (LR) directions. The tumor motion ranges were extracted from the in-house 4DCT motion evaluation records and were crosschecked by two medical physicists.

Machine learning pipeline. The aim of developing a super-learner model was to learn the potential correlation of clinical and imaging features with tumor motion extent. With such a model, the tumor motion ranges in three dimensions can be predicted for a new patient and a proper motion management technique can be chosen for the patient. The study design is depicted in Fig. 5. The upper block demonstrates the model development process. The entire dataset was first divided into the training and test sets, where the test set was 10% of the dataset size and was kept independent of the entire training and validation process (step 1 in Fig. 5). The training and test sets were pre-processed independently as described in the previous section, and a subset of input features was selected from the training set (step 2). The training set was further partitioned in a five-fold cross-validation fashion, in which the entire training set was split into five sub-samples (step 3), each of which acts once as a validation set and four times as a part of a training set. Four machine learning models were selected as the base models and the optimal hyper-parameters of each model were tuned using Bayesian optimization on the training set (step 4). Once the base models were trained, a new dataset consisting of the predicted outputs of each base model on the five-fold cross-validation data and the ground-truth outputs of the cross-validation data was formed. This newly

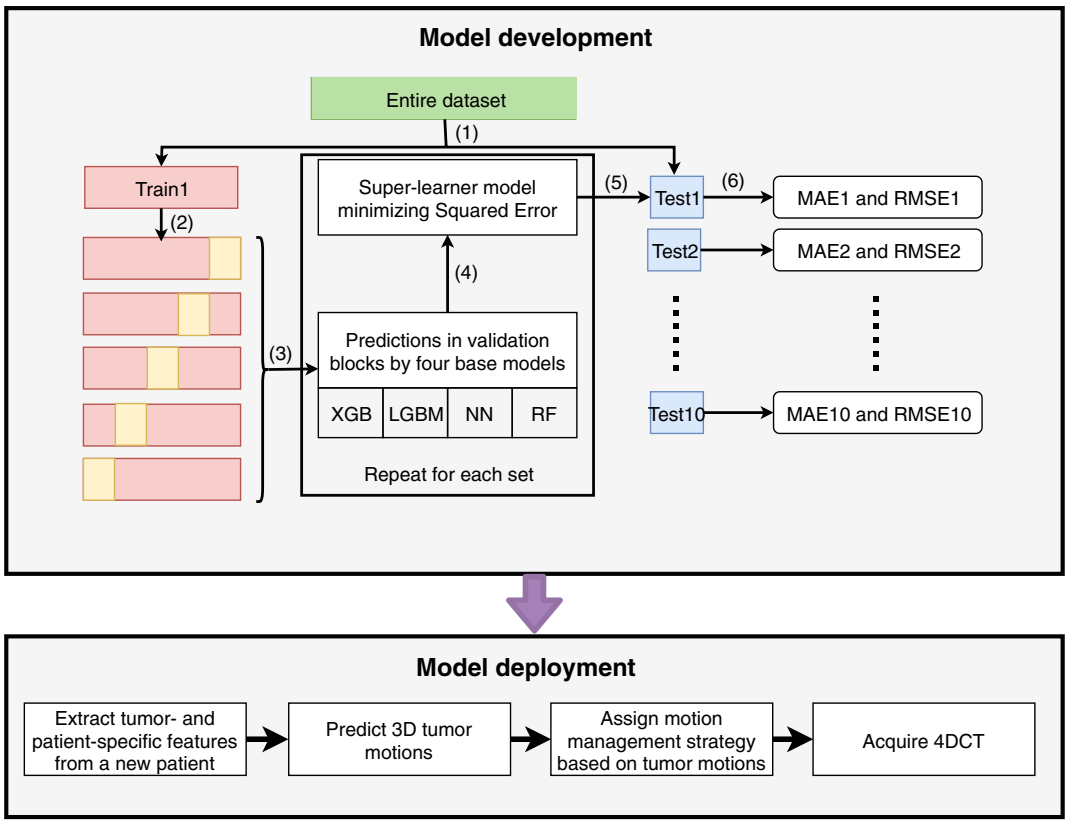

Figure 5. Experimental design of the Super-Learner model. The entire dataset is first divided into the training and independent test groups, in which the training group is further divided using five-fold cross-validation method (The validation set is shown in yellow).

formed dataset is the basis to train the super-learner model, in which all convex combinations of base models will be evaluated and the optimal coefficients of the super-learner model will be determined by minimizing the cross-validation errors (step 5). More mathematical details of the buildup process of the super-learner model were covered in the Super-Learner Model section. The trained super-learner model was applied to the independent test set (step 6) and the Mean Absolute Errors (MAE) and Root Mean Square Errors (RMSE) of predicted values in three directions were computed. The aforementioned model building processes (step 2 to step 6) were repeated ten times to obtain MAE and RMSE using the training and test pairs produced in step 1. The model performance (MAE and RMSE) of each test set was recorded to demonstrate the true generalizability of the super-learner model. Such a super learner model can then be used to predict a new patient's motion range in a clinical setting, which is shown in the lower block: when a new patient comes in, his/her imaging and clinical features would be extracted from the patient's diagnostic CT images and EHR data, and were used as the inputs of the pre-trained super-learner model to predict the tumor motion extents.

Feature selection. Feature selection methods involved in this study include Model-based Recursive Feature Elimination (RFE) and Feature Collinearity Removal.

Model-based RFE is a greedy algorithm based on the feature ranking technique. According to a specific feature ranking standard, RFE starts from a complete set and then eliminates the least relevant feature one by one to select the most important features. In this study, XGBoost was employed as the feature selection model, and the processes of XGBoost–RFE, a feature selection method that combines XGBoost and RFE, were shown as follow.

(1) Input: The training set was denoted as $\boldsymbol{X_0} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_K}]^T$, in which each $\boldsymbol{x_i}$, $i = 1, 2, \ldots, K$ covered a set of m features $\boldsymbol{P} = [p_1, p_2, \ldots, p_m]$. The output values of tumor motions were denoted as $\boldsymbol{y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_K}]^T$, where K is the number of training patients.
(2) Output:The feature rank $\boldsymbol{R}$, which was initialized as null at the initial point.
(3) Repeat the following steps m times:

Train the XGBoost model and rank the features in the input feature set $\boldsymbol{P}$ by minimizing the mean absolute error;

Find out the least important features f;

Update the list of feature rank $\boldsymbol{R}$;

Exclude the feature with minimum criterion: $\boldsymbol{P} = \boldsymbol{P} - \boldsymbol{P}(f)$.

Model	Characteristics	Parameters
Random Forest[39]	A large number of decision trees based on random subsampling	n_estimators, max_depth, max_features, min_samples_split, min_samples_leaf
Multi-Layer Perceptron (MLP) Networks[40]	Auxiliary features are generated by each layer; a high number of tunable weights	layer compositions, number of hidden units, dropouts, learning rate, number of epochs
XGBoost[41]	A variation of boosting; generalizes weak learners by allowing optimization of the differentiable loss function	max_depth, min_child_weight, subsample, colsample_bytree, learning rate, num_boost_round
LightGBM[42]	Gradient boost based on the decision tree algorithm	num_leaves, feature_fraction, lambdas, max_depth, min_child_samples

Table 2. Summary of the base machine learning models used in this study.

The RFE selection is basically a recursive process that ranks features according to the measure of importance. The recursion is needed because for certain measures the relative importance of each feature can change substantially when evaluated over a different subset of features during the step-wise elimination process (in particular for highly correlated features). After the optimal subset of features was selected, the Pearson correlation heatmap was employed to discover the col-linearity among the features, in which features with high col-linearities were identified by the eigenvalues of the heatmap and were removed, as the existence of highly correlated features may reduce the predictive power of machine learning models. The feature importance was evaluated by the F-score[26,27], which is based on the frequency of a feature being selected for the tree splitting, scaled by the squared improvement to the model and averaged over all trees. The larger the F-score, the more discriminative the feature is. The RFE and collinearity removal process was conducted in each round of independent tests, and the relative importance of each feature was measured by the F-score.

Super-learner model. Super-learner model, also known as the model ensemble, refers to a loss-based learning method that has been proposed and analyzed by van der Laan *et al.*[28]. It addressed a common task in data mining, which is the estimator selection for prediction. In the tumor motion prediction task, it is feasible to create a set of machine learning models estimating the tumor motions in the SI, AP and LR directions but with function varieties. Specifically, each machine learning model is an estimator that mapped the input feature dataset of K patients $(\boldsymbol{x_k}, \boldsymbol{y_k}), k = 1, 2, \ldots, K$ into a prediction function $F(\boldsymbol{y}|\boldsymbol{x})$ that can be used to map an input $\boldsymbol{x}$ into a predictive value $\boldsymbol{y}$. The model selection was not limited to select only a single model. Recent studies have demonstrated that the predictive accuracy of an ensemble of multiple models can outperform a single model. One of the ensemble techniques is the super-learner modeling, which was related to the stacking algorithm introduced in neural networks by Wolpert[29] and adapted to the regression model by Breiman[30]. The stacking algorithm was evaluated in LeBlanc and Tibshirani[31] and the relationship to the mixed model method of Stone[32] was discussed.

Our development of a super-learner model involved two major processes. The first step was to select and train a collection of base machine learning models. The base models may range from a simple regression model to a multi-step model involving feature screening and hyper-parameters optimization. The second step was to build up the super-learner model upon the trained base models and minimize the cross-validation risks. Specifications of each step were illustrated in the following sub-sections.

Base machine learning models. Four classical machine learning models were selected to work as the base models in this study. Table 2 summarized each model's characteristics. All of the base models were implemented using Python 3.6 with the scikit-learn package[33] and Keras API[34].

Super-learner model. The procedure of building up and training the super-learner model can be illustrated as follows. The library of base models was consisted of aforementioned four base models. Each base model is trained in a five-fold cross-validation fashion, where the validation samples of each fold are denoted as $\boldsymbol{V}(\boldsymbol{v})$ and the training samples of each fold are denoted as $\boldsymbol{L}(\boldsymbol{v})$ ($v = 1, 2, \ldots, 5$). For the v–th fold, each base model was fit on $\boldsymbol{L}(\boldsymbol{v})$ and the predictions on the corresponding validation data are denoted as $\hat{y}_{n,\boldsymbol{L}(\boldsymbol{v})}(\boldsymbol{X_i}), n = 1, 2, .., 4, \boldsymbol{X_i} \in \boldsymbol{V}(\boldsymbol{v})$.

The second step is to stack the predictions from each model to create a prediction matrix $z = \{y_{n,\boldsymbol{L}(\boldsymbol{v})}(\boldsymbol{X}_{V(v)}), n = 1, 2, \ldots, 4, v = 1, 2, \ldots, 5\}$, where we used the notation $\boldsymbol{X}_{V(v)}$ for the input feature vectors of the validation sample. A cohort of weighted combinations $\boldsymbol{m}(\boldsymbol{z}|\boldsymbol{\beta})$ of all candidate models are proposed and indexed by a weight vector $\boldsymbol{\beta}$, where

$$\boldsymbol{m}(\boldsymbol{z}|\boldsymbol{\beta}) = \sum_{n=1}^{4} \boldsymbol{\beta}_n \boldsymbol{y}_{n,\boldsymbol{L}(\boldsymbol{v})}(\boldsymbol{X}_{V(v)}), \sum_{n=1}^{4} \boldsymbol{\beta}_n = 1 \tag{1}$$

The third step is to determine $\boldsymbol{\beta}$ that minimizes the cross-validated errors by calculating the squared difference between the weighted vector combinations and the ground truth output $\boldsymbol{Y_k}$ through

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{k=1}^{K} (\boldsymbol{Y_k} - \boldsymbol{m}(\boldsymbol{z_k}|\boldsymbol{\beta}))^2 \tag{2}$$

Finally, the optimal weight vector $\hat{\boldsymbol{\beta}}$ was combined with $\boldsymbol{y}_n(\boldsymbol{X})$ based on the weights $\boldsymbol{m}(\boldsymbol{z}_k|\boldsymbol{\beta})$ to create the predictions obtained by the final super-learner $\boldsymbol{y}_{SL}(\boldsymbol{X})$, where

$$\boldsymbol{y}_{SL}(\boldsymbol{X}) = \sum_{n=1}^{4} \hat{\boldsymbol{\beta}}_n \boldsymbol{y}_n(\boldsymbol{X}) \tag{3}$$

Bayesian optimization for hyper-parameter tuning. The performance of machine learning methods depends crucially on hyper-parameter settings and thus on the method used to select hyper-parameters. Recently, Bayesian optimization methods[35] have been shown to outperform established methods for this problem[36]. In this study, we utilized Bayesian optimization to construct a probabilistic model to select subsequent hyper-parameter configurations. In order to select its next hyper-parameter configuration using the probability model, Bayesian optimization used an acquisition function that relied on the predictive distribution of the probability model at arbitrary hyper-parameter configurations to quantify how useful knowledge about the hyper-parameter configuration would be. The acquisition function used in this study is the expected improvement[37] over the best previously observed function value attainable at a hyper-parameter configuration[38]. Among existing Bayesian optimization algorithms, the major difference is the model classes being used. In this paper, we empirically chose Tree Parzen Estimator (TPE)[36].

Metrics of predictive performance. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used to evaluate the predictive performance of the super-learner model. The evaluation criteria indicated the model's prediction ability by comparing the tumor motion ranges in the testing set with their corresponding predicted values generated by the super-learner model.

MAE is a measure of the magnitude of errors, and can be calculated by

$$MAE = \frac{1}{K}\sum_{k=1}^{K} |\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k| \tag{4}$$

where $\boldsymbol{y}_k$ is the ground truth tumor motion extents of k-th patient, and $\hat{\boldsymbol{y}}_k$ is the predicted tumor motion extents, and K is the number of investigated patients.

RMSE is calculated by taking the square root of the mean of the square of all the errors. RMSE represents the standard deviation of the differences between predicted values and ground truth values. The effect of each error on RMSE is proportional to the size of the squared error. Consequently, RMSE is more sensitive to outliers. RMSE can be expressed by

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K} (\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k)^2} \tag{5}$$

Data availability

The patient datasets used for the model building and evaluation in the current study are available from the corresponding author on reasonable request and subject to Institutional Review Board (IRB) approval.

Received: 22 February 2019; Accepted: 25 September 2019;

Published online: 16 October 2019

References

1. Negoro, Y. *et al.* The effectiveness of an immobilization device in conformal radiotherapy for lung tumor: reduction of respiratory tumor movement and evaluation of the daily setup accuracy. *International Journal of Radiation Oncology Biology Physics* **50**, 889–898 (2001).
2. Ozhasoglu, C. & Murphy, M. J. Issues in respiratory motion compensation during external-beam radiotherapy. *International Journal of Radiation Oncology Biology Physics* **52**, 1389–1399 (2002).
3. Shimizu, S. *et al.* Impact of respiratory movement on the computed tomographic images of small lung tumors in three-dimensional (3d) radiotherapy. *International Journal of Radiation Oncology Biology Physics* **46**, 1127–1133 (2000).
4. Liu, H. H. *et al.* Assessing respiration-induced tumor motion and internal target volume using four-dimensional computed tomography for radiotherapy of lung cancer. *International Journal of Radiation Oncology Biology Physics* **68**, 531–540 (2007).
5. Underberg, R. W., Lagerwaard, F. J., Slotman, B. J., Cuijpers, J. P. & Senan, S. Use of maximum intensity projections (mip) for target volume generation in 4dct scans for lung cancer. *International Journal of Radiation Oncology Biology Physics* **63**, 253–260 (2005).
6. Xi, M. *et al.* Defining internal target volume (itv) for hepatocellular carcinoma using four-dimensional ct. *Radiotherapy and Oncology* **84**, 272–278 (2007).
7. Remouchamps, V. M. *et al.* Significant reductions in heart and lung doses using deep inspiration breath hold with active breathing control and intensity-modulated radiation therapy for patients treated with locoregional breast irradiation. *International Journal of Radiation Oncology Biology Physics* **55**, 392–406 (2003).
8. Lin, L. *et al.* Evaluation of motion mitigation using abdominal compression in the clinical implementation of pencil beam scanning proton therapy of liver tumors. *Medical physics* **44**, 703–712 (2017).
9. Keall, P. J. *et al.* The management of respiratory motion in radiation oncology report of aapm task group 76a. *Medical physics* **33**, 3874–3900 (2006).
10. Sharp, G. C., Jiang, S. B., Shimizu, S. & Shirato, H. Prediction of respiratory tumour motion for real-time image-guided radiotherapy. *Physics in Medicine & Biology* **49**, 425 (2004).

11. Bukovsky, I. *et al*. A fast neural network approach to predict lung tumor motion during respiration for radiation therapy applications. *BioMed research international* **2015** (2015).
12. Teo, T. P. *et al*. Feasibility of predicting tumor motion using online data acquired during treatment and a generalized neural network optimized with offline patient tumor trajectories. *Medical physics* **45**, 830–845 (2018).
13. Balasubramanian, A., Shamsuddin, R., Prabhakaran, B. & Sawant, A. Predictive modeling of respiratory tumor motion for real-time prediction of baseline shifts. *Physics in Medicine & Biology* **62**, 1791 (2017).
14. Li, G. *et al*. Rapid estimation of 4dct motion-artifact severity based on 1d breathing-surrogate periodicity. *Medical physics* **41**, 111717 (2014).
15. Ruan, D., Fessler, J. A., Balter, J. M. & Sonke, J.-J. Exploring breathing pattern irregularity with projection-based method. *Medical physics* **33**, 2491–2499 (2006).
16. Jones, B. L., Schefter, T. & Miften, M. Adaptive motion mapping in pancreatic sbrt patients using fourier transforms. *Radiotherapy and Oncology* **115**, 217–222 (2015).
17. Stephenson, J. A. & Wiley, A. L. Jr. Current techniques in three-dimensional ct simulation and radiation treatment planning. *Oncology-Huntington* **9**, 1225–1234 (1995).
18. Tantisuwat, A. & Thaveeratitham, P. Effects of smoking on chest expansion, lung function, and respiratory muscle strength of youths. *Journal of physical therapy science* **26**, 167–170 (2014).
19. Parameswaran, K., Todd, D. C. & Soth, M. Altered respiratory physiology in obesity. *Canadian respiratory journal* **13**, 203–210 (2006).
20. Chen, H. & Kuo, C.-S. Relationship between respiratory muscle function and age, sex, and other factors. *Journal of Applied Physiology* **66**, 943–948 (1989).
21. Lin, H. *et al*. Su-g-brc-13: Model based classification for optimal position selection for left-sided breast radiotherapy: Free breathing, dibh, or prone. *Medical physics* **43**, 3629–3630 (2016).
22. Lin, H. *et al*. Feasibility study of individualized optimal positioning selection for left-sided whole breast radiotherapy: Dibh or prone. *Journal of applied clinical medical physics* **19**, 218–229 (2018).
23. Thomaz, R. L., Carneiro, P. C. & Patrocinio, A. C. Feature extraction using convolutional neural network for classifying breast density in mammographic images. In *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 101342M (International Society for Optics and Photonics, 2017).
24. Jiang, Y. *et al*. Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? In *International Conference on Artificial Intelligence in Education*, 198–211 (Springer, 2018).
25. Yu, Z. H., Lin, S. H., Balter, P., Zhang, L. & Dong, L. A comparison of tumor motion characteristics between early stage and locally advanced stage lung cancers. *Radiotherapy and Oncology* **104**, 33–38 (2012).
26. Chen, Y.-W. & Lin, C.-J. Combining svms with various feature selection strategies. In *Feature extraction*, 315–324 (Springer, 2006).
27. Xie, J. & Wang, C. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications* **38**, 5809–5815 (2011).
28. Van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Statistical applications in genetics and molecular biology* **6** (2007).
29. Wolpert, D. H. Stacked generalization. *Neural networks* **5**, 241–259 (1992).
30. Breiman, L. Bagging predictors. *Machine learning* **24**, 123–140 (1996).
31. LeBlanc, M. & Tibshirani, R. Combining estimates in regression and classification. *Journal of the American Statistical Association* **91**, 1641–1650 (1996).
32. Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* **36**, 111–133 (1974).
33. Pedregosa, F. *et al*. Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).
34. Chollet, F. *et al*. Keras (2015).
35. Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, 2951–2959 (2012).
36. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13**, 281–305 (2012).
37. Schonlau, M., Welch, W. J. & Jones, D. R. Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series* 11–25 (1998).
38. Eggensperger, K. *et al*. Towards an empirical foundation for assessing bayesian optimization of hyperparameters. In *NIPS workshop on Bayesian Optimization in Theory and Practice*, vol. 10, 3 (2013).
39. Liaw, A. *et al*. Classification and regression by randomforest. *R news* **2**, 18–22 (2002).
40. Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E. & Suter, B. W. The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Transactions on Neural Networks* **1**, 296–298 (1990).
41. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (ACM, 2016).
42. Ke, G. *et al*. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 3146–3154 (2017).

Author contributions

H.L. collected the data, developed the pipeline and carried out the analyses. B.K.T. and T.L. validated the data, B.K.T., W.Z., T.L., S.J.F. and L.D. discussed the problem and analyzed the results. L.D. conceived the study, supervised and supported the research. All the authors edited and reviewed the manuscript.

Competing interests

L.D. is on the Speakers Bureau for Varian Medical Systems. The remaining authors declare no conflict of interest as defined by Nature Research, or other financial and non-financial interests that might be perceived to influence the results and/or discussion reported here.

Additional information

Correspondence and requests for materials should be addressed to L.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.