Machine Learning-based Prediction of Relative Regional Air Volume Change from Healthy Human Lung CTs

Eunchan Kim[1,3]†, YongHyun Lee[2]†, Jiwoong Choi[4,5]*, Byungjoon Yoo[1,3] , Kum Ju Chae[6]*, and Chang Hyun Lee[7]*

[1] Department of Intelligence and Information, Seoul National University
Seoul 08826, Republic of Korea
[2] Department of Computer Science and Engineering, Seoul National University
Seoul 08826, Republic of Korea
[3] Graduate School of Business, Seoul National University
Seoul 08826, Republic of Korea
[4] Department of Internal Medicine, School of Medicine, University of Kansas
Kansas City, Kansas 66160, United States of America
[5] Department of Bioengineering, University of Kansas
Lawrence, Kansas 66045, United States of America
[6] Department of Radiology, Research Institute of Clinical Medicine of Jeonbuk National University
Biomedical Research Institute of Jeonbuk National University Hospital
Jeonju 54907, Republic of Korea
[7] Department of Radiology, Seoul National University Hospital, Seoul National University, College of Medicine
Seoul 03080, Republic of Korea
[e-mail: eunchan@snu.ac.kr, leeyh@idb.snu.ac.kr, jchoi4@kumc.edu,
byoo@snu.ac.kr, para2727@gmail.com, changhyun.lee@snu.ac.kr]
*Corresponding authors: Jiwoong Choi, Kum Ju Chae, Chang Hyun Lee
†These authors contributed equally to this work

Received August 9, 2022; revised October 7, 2022; accepted October 27, 2022; published February 28, 2023

Abstract

Machine learning is widely used in various academic fields, and recently it has been actively applied in the medical research. In the medical field, machine learning is used in a variety of ways, such as speeding up diagnosis, discovering new biomarkers, or discovering latent traits of a disease. In the respiratory field, a relative regional air volume change (RRAVC) map based on quantitative inspiratory and expiratory computed tomography (CT) imaging can be used as a useful functional imaging biomarker for characterizing regional ventilation. In this study, we seek to predict RRAVC using various regular machine learning models such as extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and multi-layer perceptron (MLP). We experimentally show that MLP performs best, followed by XGBoost. We also propose several relative coordinate systems to minimize intersubjective variability. We confirm a significant experimental performance improvement when we apply a subject's relative proportion coordinates over conventional absolute coordinates.

Keywords: Biomedical machine learning, chronic obstructive pulmonary disease, deep learning, quantitative CT imaging, relative regional air volume change.

A preliminary version of this paper was presented at APIC-IST 2022, and was selected as a best paper.

1. Introduction

Machine learning and deep learning are used in many areas of modern human daily life [1]. It is used in various fields such as image recognition [2, 3], natural language processing [4–6], and anomaly detection [7, 8]. These areas have been significantly improved by using machine learning and deep learning techniques. Image recognition is used for security purposes, such as facial recognition, and is used for object detection, which is one of the key elements of autonomous driving. In the field of natural language processing, it is used for automatic translations, automatic response functions, and chatter robots (ChatBots). Anomaly detection can be used for network intrusion detection or abnormal credit card transaction detection. Machine learning is also used when Netflix recommends movies that you might potentially like. In these ways, machine learning and deep learning are common in our daily lives.

Integration of machine learning and deep learning in the field of medicine is also becoming more and more popular [9, 10]. Since medical data are in different formats such as reports, discharge summaries, images, and audio, the most appropriate model varies according to the applied fields and purposes. Models learned from medical data can be used to increase the accuracy of diagnosis or serve as an auxiliary aid to patient care.

Machine learning has been widely used in respiratory medicine, for example, for early prediction of childhood asthma, early prediction of asthma exacerbations, and characterization of asthma and chronic obstructive pulmonary disease (COPD) phenotypes [11–15]. COPD is the third leading cause of death and is expected to become the first leading cause of death [16, 17]. Pulmonary function test (PFT) results, such as forced expiratory volume in the first second (FEV1) and forced vital capacity (FVC), are used to determine the stages of COPD [18]. However, since the PFT reflects only the whole lung function, and not regional lung decline of functional features before the destruction of lung tissue [19, 20], early detection or self-recognition is difficult until the whole lung function is severely declined. Detection and management of early COPD have recently received a lot of attention [21]. Machine learning approaches with imaging have been capable of characterizing early stage progression of COPD [20]. Furthermore, machine learning has been widely used to investigate various aspects of COVID-19 [22–24]. Machine learning is intensively used for the early diagnosis of lung diseases, which is very important in treatment efficacy [14, 25, 26].

Recently, apart from machine learning, CT scans have been used to quantitatively characterize imaging-based regional lung function in lung diseases [27–33]. Chae et al. [33] introduced a standardized measure of CT-based local ventilatory capacity named relative regional air volume change (RRAVC) to differentiate COPD patients from normal subjects. They found that RRAVC can be an effective imaging biomarker for regional lung ventilation that shows how much air goes in and out of each lung region from CT scans, and that we can find how much air flow is limited as lung disease progresses. RRAVC also characterizes the effects of supine versus prone body positions in regional lung ventilation distribution of normal lungs [31].

From inspiratory and expiratory CT images of a human subject, segmentation and image registration are first conducted. Then, RRAVC values are calculated at local lung regions by mathematical definition. The authors show how the distribution of RRAVC differs between COPD patients and normal subjects [33]. However, the degrees of abnormality in RRAVC values in individual local lung regions are not evaluated solely using the defined calculation, because RRAVC is not uniform, even in normal lungs, and there is no reference value for each region. In this paper, we seek to predict RRAVC values at small-scale local lung regions using various machine learning modeling approaches.

There are three key contributions in this research. 1) This study is the first attempt to predict normal subjects' RRAVC values using machine learning and deep learning without direct calculation by its definition. We provide a machine learning baseline model as the first step to measure the degrees of functional abnormality in local lung regions from quantitative CT models. 2) The x, y, and z coordinates of the lung are entered into the RRAVC calculation at local lung regions. Since each patient's height and body type are all different, the x, y, z coordinate values of each patient's lungs are standardized to reflect this. We try to maximize the performance of machine learning models by introducing a total of three relative coordinate systems. 3) We explore the optimal number of layers for multi-layer perceptron (MLP), which is one of the deep learning models. In addition, we experimentally show that this optimized MLP model performs better than existing regression models.

2. Related Work

RRAVC is defined as the ratio between normalized local air volume change and normalized entire lung air volume change. **Fig. 1** illustrates the process of how RRAVC values are created [33].

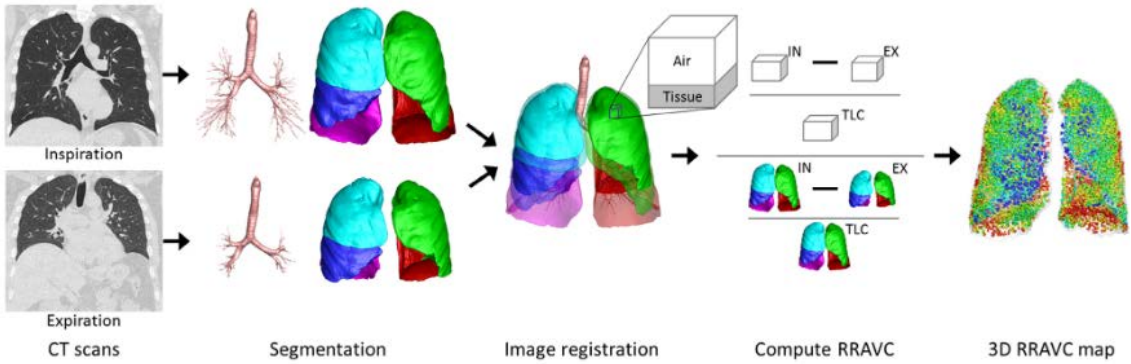

Fig. 1. A schematic of the process from segmentation of inspiration and expiration CTs to CT image matching by image registration to the final color-coded 3D RRAVC map.

We use two CT images acquired from a patient: one at full inspiration and the other at full expiration. From those CT images, an image processing called "segmentation" identifies separate regions of airways, lungs, lobes, and blood vessels using VIDA Vision software (Coralville, IA, USA). A human lung has five lobes, as you see in five different colors in the segmentation part of **Fig. 1**. This software is a tool to analyze airway and segment lung CT images. After segmentation, a process called image registration matches the anatomically same lung regions from the two CT images and maps the expiratory CT image onto the inspiratory CT image in the same coordinate system. This means that we cannot only compare local lung regions from the two images but also compute various variables, such as air ventilation. We compute relative regional air volume change, i.e. RRAVC, at small lung units as shown here, to quantitatively characterize the three-dimensional map of regional air ventilation distribution characteristics. RRAVC is calculated as follows.

$$\mathrm{RRAVC} = \frac{\left(v_{air}^{insp} - v_{air}^{exp}\right)/v_{air}^{TLC}}{\left(v_{air}^{insp} - v_{air}^{exp}\right)/TLC} \quad (1)$$

Fig. 2 shows how the generated RRAVC values differ between about three-thousand regions of normal lungs and COPD lungs [33]. The left image shows the normal lung of a 53-year-old man. More ventilation happens in the red regions than blue region. Ventilation in the healthy lung increases toward the bottom and back of the lung. The right image is from a 60-year-old COPD patient who showed significant heterogeneous ventilation.

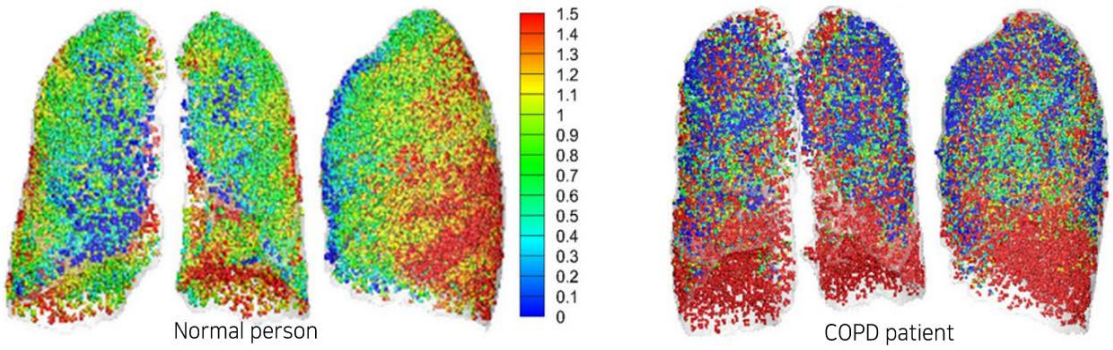

Fig. 2. Representative images of 3D RRAVC maps of (a) a 53-year-old normal subject, and (b) a 60-year-old man with emphysema.

In the case of a normal subject, the RRAVC values gradually change from near zero to 1.5 over the lung regions. However, in a COPD patient, more extreme RRAVC values are predominant. And from the side view, the RRAVC values do not change smoothly. We can clearly see the difference in the distribution of RRAVC values between normal and COPD patients. Therefore, interpretation of the RRAVC distribution is considered a very useful imaging biomarker for understanding regional decline of air ventilatory function in COPD and other diseased lungs. We expect a machine learning-based prediction of RRAVC values will help better understand lung ventilation decline in COPD.

3. Machine Learning Models

Recently, various machine learning techniques for pulmonary nodule automatic detection have been applied in the respiratory field [36, 37]. Similarly, we would like to use machine learning techniques to predict RRAVC.

Our predictive target variable, RRAVC, is a continuous value. The regression algorithm that predicts these continuous variable values has been actively researched before the recent popularity of machine learning. The simplest multivariate linear regression among various linear regression models is noted in (2).

$$\underset{\alpha \in R,\, \beta \in R^p}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{2}$$

There are more advanced linear regression models, for example, the ridge regression model and the LASSO regression model. These two are very famous regularization techniques. They reduce the weight of independent variables with low explanatory power by limiting the size of the regression coefficients when calculating the regression coefficients. Ridge regression is the addition of L2-norm regularization and LASSO is the addition of L1-norm regularization.

Ridge regression can be expressed as (3) as shown below, and LASSO regression is expressed as (4).

$$\underset{\alpha\in R,\,\beta\in R^p}{\operatorname{argmin}} \sum_{i=1}^{N}\left(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|^2 \tag{3}$$

$$\underset{\alpha\in R,\,\beta\in R^p}{\operatorname{argmin}} \sum_{i=1}^{N}\left(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{4}$$

ElasticNet is a combination of ridge regression and LASSO regression, and it is a technique that includes both the L1 norm and the L2 norm [38].

$$\underset{\alpha\in R,\,\beta\in R^p}{\operatorname{argmin}} \sum_{i=1}^{N}\left(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda_1\sum_{j=1}^{p}\left\|\beta_j\right\|_1+\lambda_2\sum_{j=1}^{p}\left\|\beta_j\right\|_2^2 \tag{5}$$

Support vector regression is a method that learns to fit as much data as possible within the margin (tube). We used both linear and non-linear kernel functions [39]. We also used two famous ensemble techniques. Extreme gradient boosting (XGBoost) [40] and light gradient boosting machine (LightGBM) [41] are ensemble models that use multiple decision trees. These two are known for their good performance in various fields. XGBoost has level-wise tree growth whereas LightGBM has leaf-wise tree growth.

MLP is an artificial neural network in which perceptrons are stacked using two or more hidden layers. We use the Adam optimizer as the MLP setting in this paper and the dropout probability is 0.2. Rectified linear unit (ReLU) is used as our nonlinear activation function. In addition to the input layer and output layer, 3 hidden layers are used and the structure information of the number of nodes per layer is as follows. The input layer is 46 or 52 depending on which coordinate is used. Experiments with more diverse MLP structures are conducted in section 5.3.

Table 1. MLP Architecture for RRAVC Prediction

Layers	# of nodes
Input layer	46 or 52
Hidden layer 1	30
Hidden layer 2	20
Hidden layer 3	10
Output layer	1

In MLP, a decision of the number of layers (i.e., how many hidden layers to stack) is a very important process of hyper-parameter tuning. Therefore, several experiments are conducted to find the optimal model while changing the number of these layers. Of course, as the number of hidden layer changes, the number of nodes in each hidden layer also changes.

After segmentation and image registration from the CT images, we obtain features for each data point of the lung. At this time, the x, y, and z coordinates of each lung were also obtained. These are absolute coordinates. However, these absolute coordinates may not reflect the patient's characteristics properly. Because humans have different height sizes and different

chest thicknesses, the x, y, and z coordinate values are different even if the specific data point is exactly the center coordinates of each human. To solve this problem, we try to maximize the performance of our model by introducing the following three relative coordinate systems: 1) relative coordinates, 2) relative proportional coordinates, and 3) potential relative coordinates.

4. Relative Coordinates for Standardization

4.1 Relative Coordinates

Relative coordinates indicate how far each data point coordinate is from the minimum position. The formula is as follows.

$$\left(x_{rpc_i}, y_{rpc_i}, z_{rpc_i} \right) = \left(x_i - \min(x) , y_i - \min(y) , z_i - \min(z) \right) \quad (6)$$

For example, suppose a simple lung modeling of a subject S1 is represented in two-dimensional coordinates. In this case, we have x and y coordinates. Let us assume that the lower-left data point (the data point with the smallest x and y coordinates) is (5, 10) and the upper-right data point (the data point with the largest both the x and y coordinates) is (95, 100). In this case, min(x) = 5, min(y) = 10, max(x) = 95, max(y) = 100. Assume we have a data point A (50, 55) above that lung. In this case, the absolute coordinate of point A is (50, 55), but the relative coordinates are (50-5, 55-10) = (45, 45). The absolute coordinate for data point A is not above the y=x line. However, it is above the y=x line for the relative coordinates system.

4.2 Relative Proportional Coordinates

This indicates how far each coordinate is between the minimum and maximum values. In the S1 subject example above, the relative coordinates value of point A is (45, 45). And the top right (95, 100) is also converted to (95-5, 100-10) = (90, 90) in the relative coordinates. Therefore, it can be seen that point A is located in the center of the lung modeling of that subject.

Suppose there is another subject S2, and S2 is taller than S1. This means that coordinates values of the upper-right data point of S2 is larger than that of the previous subject S1. Let's assume that the lower-left corner and point A are kept as S1 example, and only the coordinates of the upper-right corner are changed to (125, 130). In this case, the relative coordinates of S2 are (45, 45), which is the same as that of S1. However, we cannot see that point A is located in the center of the lungs of S2 like S1. If we transform the upper-right coordinate of S2 to relative coordinates, (125-5, 130-10) = (120, 120). Therefore, point A of the S2 subject is only located about 1/3 of the lower left corner of the overall coordinate system. To solve this problem, we propose a relative proportional coordinates that normalizes each relative coordinate as follows.

$$\left(x_{rppc_i}, y_{rppc_i}, z_{rppc_i}\right) = \left(\frac{x_i - \min(x)}{\max(x) - \min(x)}, \frac{y_i - \min(y)}{\max(y) - \min(y)}, \frac{z_i - \min(z)}{\max(z) - \min(z)} \right) \quad (7)$$

Using the above relative proportional coordinates, point A in S1 is transformed to $\left(\frac{50-5}{95-5} , \frac{55-10}{100-10} \right) = (0.5, 0.5)$ and point A in S2 to $\left(\frac{50-5}{125-5} , \frac{55-10}{130-10} \right) = (0.375, 0.375)$ respectively. We can see that the problem mentioned above is solved.

4.3 Potential Relative Coordinates

This method does not compute the relative coordinates, relative proportion coordinates, or other new features of each data point. However, we add the max and min values of each coordinate of the subject as new features to every data point. The max and min values of coordinates are different for each subject. Therefore, we experiment to see if these new feature columns can potentially represent relative coordinates.

5. Experimental Results

5.1 Dataset and Computing Resources

The data set was approved by the institutional review board of Jeonbuk National University Hospital, and informed consent was received for the expiratory CT scan in the original study. We use full inspiratory CT and full expiratory CT images. The data set was generated from 292 subjects, and each subject has about 60,100 data points (rows). Each row represents a data point of an acinar scale local lung region (a sphere in **Fig. 1**). We used only 8,772,704 data points, which have nonzero RRAVC values among all 17,543,764 overall data points.

Each row consists of 37 columns, including the x, y, and z coordinates of its data point, changes in x, y, and z coordinates of each data point, length, diameter, lobe region, air volume, tissue volume, Horsfield ordering, determinant of Jacobian matrix (J), anisotropic deformation index (ADI), slab-rob index (SRI), reference index of bronchi, displacement, normalized displacement, and angle. Of the 292 people total, we use 80% as training data and 20% as test data. As the performance metrics, we utilize the R2 score, mean squared error (MSE), and mean absolute error (MAE). The experiments were conducted using Nvidia Quadro RTX 5000 GPU, Intel Xeon Gold 5220R CPU, and 187 GB of RAM.

5.2 Performance Result

5.2.1 RRAVC Prediction Using Machine Learning

We experiment with a total of 5 scaling changes (i.e., non-scaling, standard, robust, minmax, maxabs) for the robustness of determining which model performs best. Since each scaling has different characteristics, this is to compare more general performance. Based on these experiments we can find the best performing model, which can be a baseline model for future indicators of performance evaluation. Grid experiments are conducted on a total of 40 settings (5 scaling changes and 8 models mentioned in section 3). The 4 scaling formulas excluding non-scaling are as follows.

$$x_{standard} = \frac{x - \mu}{\sigma} \tag{8}$$

$$x_{robust} = \frac{x - x_{median}}{x_{Q_3} - x_{Q_1}} \tag{9}$$

$$x_{minmax} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{10}$$

$$x_{maxabs} = \frac{x}{|max(x)|} \tag{11}$$

Table 2 shows the results of comparing the R2 scores of each model for different scaling techniques. MSE of all models are shown in **Table 3**. MAE of all models are listed in the **Table 4**. Deep learning MLP models are additionally tested in section 5.2.3 so they are not included in **Table 2**. Among the above 8 models, the trained linear support vector regressor model does not converge. Therefore, it is excluded from the experimental results of this paper. Nonlinear support vector regressor takes more than 24 hours to train, which is not considered as realistic. Nonlinear support vector regressor is also not included in the experimental results.

Table 2. R2 scores of RRAVC prediction

	none	standard	minmax	robust	maxabs
Linear	0.0622803	0.0622803	0.0622803	0.0622803	0.0622803
Ridge	0.0622806	0.0622808	0.0623275	0.0622812	0.0623058
LASSO	0.0298870	0.0003844	-0.0002219	0.0004708	-0.0002219
ElasticNet	0.0431111	0.0004880	-0.0002219	0.0003826	-0.0002219
XGBoost	**0.3683279**	**0.3775327**	**0.3207365**	**0.3608702**	**0.3369287**
LGBM	0.2975713	0.3328849	0.3100740	0.3340164	0.2975726

Table 3. Mean squared error of RRAVC prediction

	none	standard	minmax	robust	maxabs
Linear	1.0818174	1.1550367	0.0001345	6.5694631	0.0003202
Ridge	1.0818170	1.1550361	0.0001345	6.5694563	0.0003202
LASSO	1.1191885	1.2312770	0.0001434	7.0024867	0.0003415
ElasticNet	1.1039323	1.2311495	0.0001434	7.0031050	0.0003415
XGBoost	**0.7287400**	**0.7667244**	**0.0000974**	**4.4776060**	**0.0002264**
LGBM	0.8103700	0.8217194	0.0000989	4.6657379	0.0002398

Table 4. Mean absolute error of RRAVC prediction

	none	standard	minmax	robust	maxabs
Linear	0.329931465	0.340913857	0.0036784	0.81303959	0.00567611
Ridge	0.329931405	0.340913626	0.00367823	0.81303862	0.00567592
LASSO	0.309842234	0.346390702	0.00373153	0.82971405	0.0057581
ElasticNet	0.307328835	0.347636997	0.00373153	0.83158253	0.0057581
XGBoost	**0.230503337**	**0.235101718**	**0.00270653**	**0.572267**	**0.0041632**
LGBM	0.253868842	0.25894121	0.00285972	0.62120382	0.00436755

The above tables show that XGBoost has the best performance for all scaling. When we look at the parameters of the trained XGBoost model to predict RRAVC, we found that *J*, a local volume expansion ratio, plays the most important role among 37 columns. This may reflect that *J* as a geometrical deformation index is important for accurate prediction of air ventilation distribution, even though *J* is not explicitly used in the formula for calculating RRAVC. *J* was also used in a deep learning patten cluster-based detection of regional lung features in a COPD population [26].

5.2.2 Effect of Relative Coordinates for Standardization

Fig. 3 shows the experimental results of comparing the absolute coordinate and three types of relative coordinates proposed in this paper for XGBoost. Similarly, the experimental results for LightGBM and MLP are shown in **Fig. 4** and **Fig. 5**, respectively.

In all cases of XGBoost, LightGBM, and MLP, the relative proportional coordinates record the highest adjusted r2 score. This coordinate system shows better performance than the

conventional absolute coordinate. XGBoost, LightGBM, and MLP also show the lowest MSE when we use relative proportional coordinates among a total of four coordinates. The relative coordinate of XGBoost and LightGBM show similar or slightly worse performance to the conventional absolute coordinate. From this result, we can deduce that 'in what proportion between the min and max' is more important than 'how far from the min'.

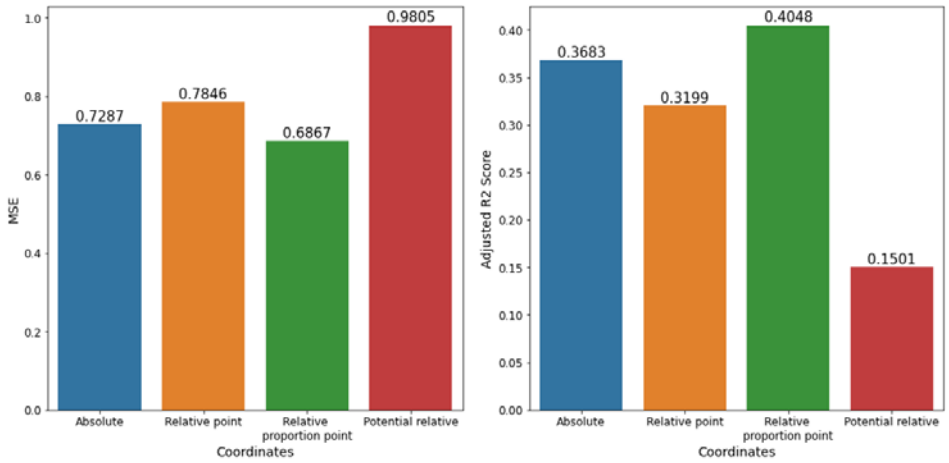

Fig. 3. XGBoost performance according to different coordinates; MSE (left) and adjusted R2 score (right).

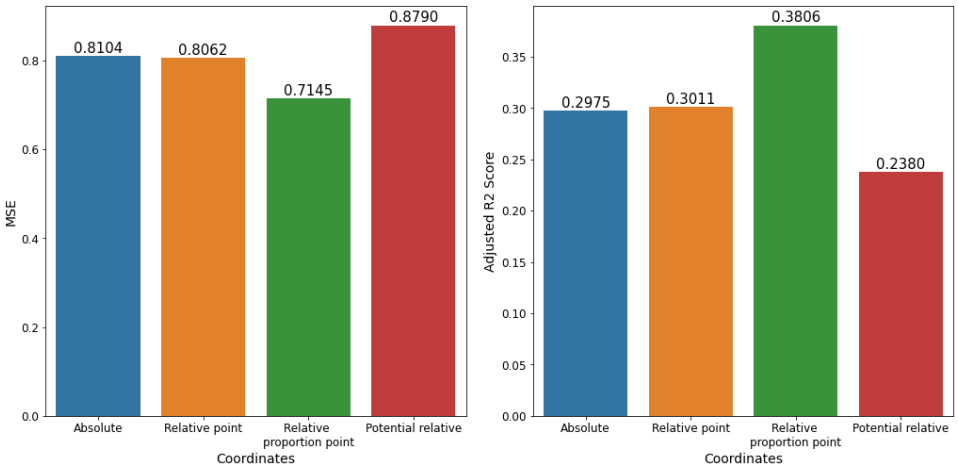

Fig. 4. LightGBM performance according to different coordinates; MSE (left) and adjusted R2 score (right).

In XGBoost and LightGBM, relative coordinates perform worse than absolute coordinates. However, for MLP, other interesting results are obtained. In MLP the relative proportional coordinates show the best performance but the relative coordinates show the second best performance. And the performance gap is not that big. This is probably the effect of deep-stacking the neural network. MLP also defeats other regression algorithms, as well as XGBoost and LightGBM. MLP with absolute coordinate performs even better than XGBoost or LightGBM with relative proportional coordinates. Therefore, we can say that MLP has the best performance among the various machine learning models we experiment with.

In case of potential relative coordinates, the performance is not good in all 3 models of XGBoost, LightGBM, and MLP. MSE increases and the adjusted R2 score decreases compared to the absolute coordinates. From this, we can see that the method of adding the max and min values of each coordinate as a feature is inefficient.

We conclude that among the conventional absolute coordinates and the three relative coordinates, the relative proportional coordinates show the best performance. Therefore, in the next subsection, when experimenting with MLPs of various structures, we use the relative proportional coordinates to find the best performing model.

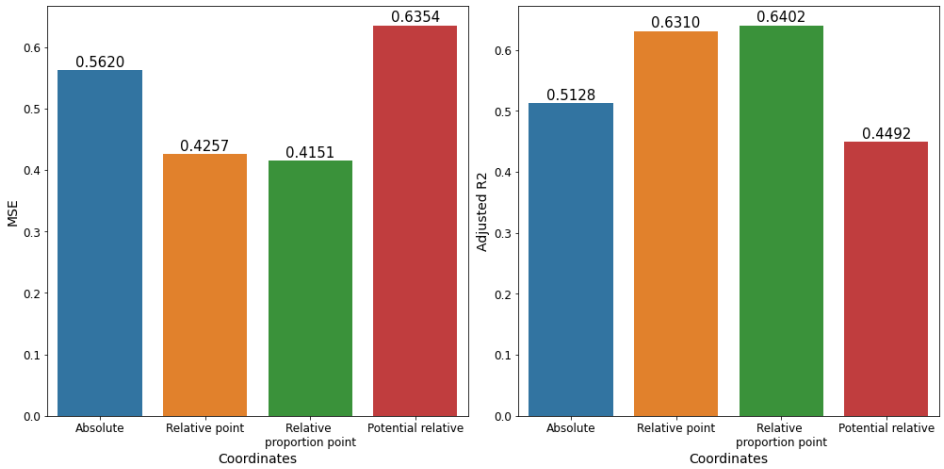

Fig. 5. MLP performance according to different coordinates; MSE (left) and adjusted R2 score (right).

5.2.3 Performance Evaluation According to the Number of MLP Layers

MLP has relatively many hyper-parameters compared to other machine learning algorithms. For example, dropout probability, learning rate, optimizer, number of hidden layers, number of nodes in each hidden layer, and number of epochs. In this paper, to find the optimal MLP model, we experiment by changing the number of hidden layers and the number of nodes in each hidden layer. The number of hidden layers is changed from 1 to 6 and the number of nodes in each hidden layer is set arbitrarily according to the number of hidden layers. For the reproducibility of the experimental results, the number of nodes in each layer is listed in **Table 5**. We set dropout probability=0.2, learning rate=10^{-4}, Adam optimizer, and epoch=10. The number of input layers is 46, and our prediction is regression, so the number of output layers is 1.

Table 5. Various MLP architectures

Number of hidden layers	Nodes in each layers
1	{ 46, 23, 1 }
2	{ 46, 23, 12, 1 }
3	{ 46, 30, 20, 10, 1 }
4	{ 46, 32, 20, 13, 6, 1 }
5	{ 46, 36, 26, 18, 10, 4, 1 }
6	{ 46, 32, 20, 14, 10, 7, 4, 1 }

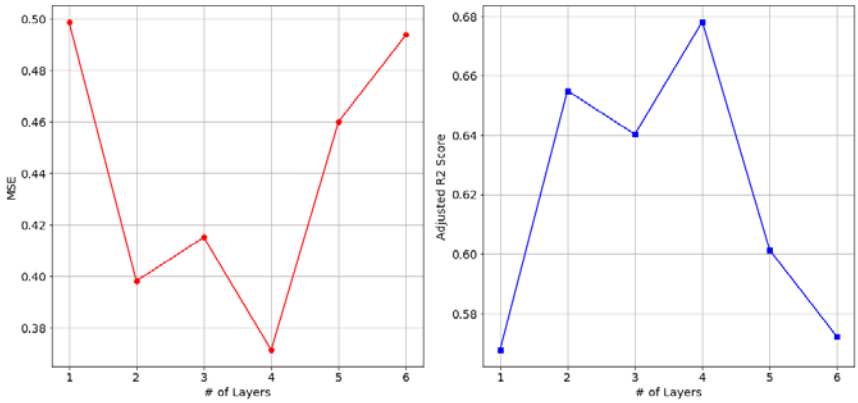

Fig. 6. MLP performance according to different number of hidden layers; MSE (left) and adjusted R2 score (right).

Fig. 6 shows the MSE and R2 scores for the evaluation. It shows the best performance when the number of hidden layers is 4; the MSE is lowest and the R2 score is highest. 2 and 3 hidden layers follow in performance. As the number of hidden layers increases, the MSE increases again. Therefore, we think it is good to use MLP models with 2-4 hidden layers for RRAVC prediction.

6. Conclusion

In this paper, we predict RRAVC, a quantitative CT imaging biomarker useful for distinguishing COPD patients from normal subjects, using machine learning rather than a defined formula. Among the various regression models, XGBoost shows the largest r2 score and the smallest MAE and MSE. We have proposed three relative coordinate methodologies and experimentally confirmed a better performance than conventional naive coordinates. In particular, among the three coordinates, relative proportional coordinates, which reflect the relative position of each point in the subject's lung, show the best performance. We also try to predict RRAVC using MLP, which is one of the basic deep learning models. We find the best MLP model architecture for RRAVC prediction by varying the number of MLP layers and structure of MLP models and show their improved performance compared to conventional machine learning models. The XGBoost or MLP model tested in this paper can be used as a baseline model for RRAVC prediction using ML in the future.

For future research, we are planning to apply more sophisticated machine learning models and deep learning models to predict other functional variables, such as *J* and functional small airway disease (fSAD). The lung bronchi can also be viewed as a graph structure, so we will try to apply several graph neural networks. Also, in this study, we used data extracted from two images: inspiratory and expiratory CT images. However, we will also try to predict the RRAVC using either an inspiratory or expiratory CT image only for efficiency.

Acknowledgement

Eunchan Kim and YongHyun Lee contributed equally. This work was supported by the Biomedical Research Institute, Jeonbuk National University Hospital Grant CUH2016-0009, the National Research Foundation of Korea (NRF) Grant 2021R1C1C1009818, and Korea Environmental Industry & Technology Institute (KEITI) Grant 2018001360001. This Study was supported by the Institute of Management Research at Seoul National University.

References

[1] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350-361, 2017. Article (CrossRef Link)

[2] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 8697-8710, Jun. 2018. Article (CrossRef Link)

[3] E. Kim, J. Lee, H. Jo, K. Na, E. Moon, G. Gweon, B. Yoo, and Y. Kyung, "SHOMY: Detection of Small Hazardous Objects using the You Only Look Once Algorithm," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 8, pp. 2688-2703, 2022. Article (CrossRef Link)

[4] B. Choi, Y. Lee, Y. Kyung, and E. Kim, "Albert with knowledge graph encoder utilizing semantic similarity for commonsense question answering," *Intelligent Automation & Soft Computing*, vol. 36, no.1, pp. 71–82, 2023. Article (CrossRef Link)

[5] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, pp. 5753-5763, Dec. 2019. Article (CrossRef Link)

[6] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. of the Empiricial Methods in Natural Language Processing*, pp. 1724-1734, 2014. Article (CrossRef Link)

[7] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," *IEEE Access*, vol. 9, pp. 78658-78700, 2021. Article (CrossRef Link)

[8] S. P. Maniraj, A. Saini, S. Ahmed, and S. Sarkar, "Credit card fraud detection using machine learning and data science," *International Journal of Engineering Research*, vol. 8, no. 9, pp. 110-115, 2019. Article (CrossRef Link)

[9] S. Siddique and J. C. Chow, "Machine learning in healthcare communication," *Encyclopedia*, vol. 1, no.1, pp. 220-239, 2021. Article (CrossRef Link)

[10] K. J. W. Tang, C. K. E. Ang, T. Constantinides, V. Rajinikanth, U. R. Acharya, and K. H. Cheong, "Artificial intelligence and machine learning in emergency medicine," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 1, pp. 156-172, 2021. Article (CrossRef Link)

[11] R. Howard, M. Rattray, M. Prosperi, and A. Custovic, "Distinguishing asthma phenotypes using machine learning approaches," *Current Allergy and Asthma Reports*, vol. 15, no. 7, pp. 1-10, 2015. Article (CrossRef Link)

[12] P. J. Castaldi, A. Boueiz, J. Yun, R. S. J. Estepar, J. C. Ross, G. Washko, and F. Banaei-Kashani, "Machine learning characterization of COPD subtypes: insights from the COPDGene study," *Chest*, vol. 157, no. 5, pp. 1147-1157, 2020. Article (CrossRef Link)

[13] J. Finkelstein and I. C. Jeong, "Machine learning approaches to personalize early prediction of asthma exacerbations," *Annals of the New York Academy of Sciences*, vol. 1387, no. 1, pp. 153-165, 2017. Article (CrossRef Link)

[14] D. Patel, G. L. Hall, D. Broadhurst, A. Smith, A. Schultz, and R. E. Foong, "Does machine learning have a role in the prediction of asthma in children?," *Paediatric Respiratory Reviews*, vol. 41, pp. 51-60, 2022. Article (CrossRef Link)

[15] B. Haghighi, S. Choi, J. Choi, E. A. Hoffman, A. P. Comellas, J. D. Newell Jr, C. H. Lee, R. G. Barr, E. Bleecker, C. B. Cooper, D. Couper, M. L. Han, N. N. Hansel, R. E. Kanner, E. A.

Kazerooni, E. A. C. Kleerup, F. J. Martinez, W. O'Neal, R. Paine III, S. I. Rennard, B. M. Smith, P. G. Woodruff, and C.-L. Lin, "Imaging-based clusters in former smokers of the COPD cohort associate with clinical characteristics: the SubPopulations and intermediate outcome measures in COPD study," *Respiratory Research*, vol. 20, no. 1, pp. 1-14, 2019. Article (CrossRef Link)

[16] World Health Organization (WHO), "The top 10 causes of death," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[17] OECD, "Main Causes of Mortality Among Women and Men in EU Countries, 2015," in *Health at a Glance: Europe 2018: State of Health in the EU Cycle*, Paris: OECD Publishing, 2018. Article (CrossRef Link).

[18] Global Initiative for Chronic Obstructive Lung Disease (GOLD), "Global Strategy for Prevention, Diagnosis and Management of Chronic Obstructive Pulmonary Disease," 2020. [Online]. Available: https://goldcopd.org/gold-reports/

[19] C. J. Galbán, M. K. Han, J. L. Boes, K. A. Chughtai, C. R. Meyer, T. D. Johnson, S. Galbán, A. Rehemtulla, E. A. Kazerooni, F. J. Martinez, and B. D. Ross, "Computed tomography–based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression," *Nature medicine*, vol. 18, no. 11, pp. 1711-1715, 2012. Article (CrossRef Link)

[20] A. P. Comellas, J. D. Newell Jr, C. H. Lee, R. G. Barr, E. Bleecker, C. B. Cooper, D. Couper, M. Han, N. N. Hansel, R. E. Kanner, E. A. Kazerooni, E. C. Kleerup, F. J. Martinez, W. O'Neal, R. Paine III, S. I. Rennard, B. M. Smith, P. G. Woodruff, E. A. Hoffman, and C.-L. Lin, "Longitudinal imaging-based clusters in former smokers of the COPD cohort associate with clinical characteristics: the subpopulations and intermediate outcome measures in COPD study (SPIROMICS)," *International journal of chronic obstructive pulmonary disease*, vol. 16, pp. 1477-1476, 2021. Article (CrossRef Link)

[21] M. E. Laucho-Contreras and M. Cohen-Todd, "Early diagnosis of COPD: myth or a true perspective," *European Respiratory Review*, vol. 29, no. 158, 2020. Article (CrossRef Link)

[22] A. S. Kwekha-Rashid, H. N. Abduljabbar, and B. Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications," *Applied Nanoscience*, pp. 1-13, 2021. Article (CrossRef Link)

[23] J. Rasheed, A. A. Hameed, C. Djeddi, A. Jamil, and F. Al-Turjman, "A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 1, pp. 103-117, 2021. Article (CrossRef Link)

[24] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight COVID-19," *Physiological Genomics*, vol. 52, no. 4, pp. 200-202, 2020. Article (CrossRef Link)

[25] T. Kadir and F. Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques," *Translational Lung Cancer Research*, vol. 7, no. 3, pp. 304-312, 2018. Article (CrossRef Link)

[26] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. M. Ngai, and J. Shao, "Classification of lung cancer using ensemble-based feature selection and machine learning methods," *Molecular BioSystems*, vol. 11, no. 3, pp. 791-800, 2015. Article (CrossRef Link)

[27] J. Choi, E. A. Hoffman, C.-L. Lin, M. M. Milhem, J. Tessier, and J. D. Newell Jr, "Quantitative computed tomography determined regional lung mechanics in normal nonsmokers, normal smokers and metastatic sarcoma subjects," *PLoS One*, vol. 12, no. 7, e0179812, 2017. Article (CrossRef Link)

[28] R. Amelon, K. Cao, K. Ding, G. E. Christensen, J. M. Reinhardt, and M. L. Raghavan, "Three-dimensional characterization of regional lung deformation," *Journal of Biomechanics*, vol. 44, no. 13, pp. 2489-2495, 2011. Article (CrossRef Link)

[29] S. Choi, E. A. Hoffman, S. E. Wenzel, M. H. Tawhai, Y. Yin, M. Castro, and C.-L. Lin, "Registration-based assessment of regional lung function via volumetric CT images of normal subjects vs. severe asthmatics," *Journal of Applied Physiology*, vol. 115, no. 5, pp. 730-742, 2013. Article (CrossRef Link)

[30] S. Bodduluri, J. D. Newell Jr, E. A. Hoffman, and J. M. Reinhardt, "Registration-based lung mechanical analysis of chronic obstructive pulmonary disease (COPD) using a supervised machine learning framework," *Academic Radiology*, vol. 20, no. 5, pp. 527-536, 2013. Article (CrossRef Link)
[31] K. M. Shin, J. Choi, K. J. Chae, G. Y. Jin, A. Eskandari, E. A. Hoffman, C. Hall, M. Castro, and C. H. Lee, "Quantitative CT-based image registration metrics provide different ventilation and lung motion patterns in prone and supine positions in healthy subjects," *Respiratory Research*, vol. 21, no. 1, pp. 1-9, 2020. Article(CrossRef Link)
[32] F. Li, J. Choi, C. Zou, J. D. Newell, A. P. Comellas, C. H, Lee, H. Ko, R. G. Barr, E. R. Bleecker, C. B. Cooper, F. Abtin, I. Barjaktarevic, D. Couper, M. Han, N. N. Hansel, R. E. Kanner, R. Paine III, E. A. Kazerooni, F. J. Martinez, W. O'Neal, S. I. Rennard, B. M. Smith, P. G. Woodruff, E. A. Hoffman, and C.-L. Lin, "Latent traits of lung tissue patterns in former smokers derived by dual channel deep learning in computed tomography images," *Scientific Reports*, vol. 11 no. 1, pp. 1-15, 2021. Article (CrossRef Link)
[33] K. J. Chae, J. Choi, G. Y. Jin, E. A. Hoffman, A. T. Laroia, M. Park, and C. H, Lee, "Relative regional air volume change maps at the Acinar scale reflect variable ventilation in low lung attenuation of COPD patients," *Academic Radiology*, vol. 27, no. 11, pp. 1540-1548, 2020. Article (CrossRef Link)
[34] J. Choi, K. J. Chae, C. H, Lee, G. Y. Jin, M. Park, C.-L. Lin, and E. A. Hoffman, "Relative regional air volume change distributions in normal subjects," in *Proc. of 8th International Workshop on Pulmonary Functional Imaging*, Seoul, Korea, March 2017.
[35] C.-L. Lin, S. Choi, B. Haghighi, J. Choi, and E. A. Hoffman, "Cluster-guided multiscale lung modeling via machine learning," *Handbook of Materials Modeling: Applications: Current and Emerging Materials*, pp. 2699-2718, 2020. Article (CrossRef Link)
[36] C. Liu, S. C. Hu, C. Wang, K. Lafata, and F. F. Yin, "Automatic detection of pulmonary nodules on CT images with YOLOv3: development and evaluation using simulated and patient data," *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 10, pp. 1917-1929, 2020. Article (CrossRef Link)
[37] S. Mei, H. Jiang, and L. Ma, "YOLO-lung: A Practical Detector Based on Improved YOLOv4 for Pulmonary Nodule Detection," in *Proc. of 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pp. 1-6, Oct. 2021. Article (CrossRef Link)
[38] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301-320, 2005. Article (CrossRef Link)
[39] F. Zhang and L. J. O'Donnell, "Support vector regression," in *Machine Learning*, Academic Press, 2020, pp. 123-140.
[40] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, Aug. 2016. Article (CrossRef Link)
[41] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, 2017.

Eunchan Kim received the B.A. degree in economics from the University of Minnesota, Twin Cities in 2012 and the M.S. degree in management information systems from Seoul National University in 2017. He is currently pursuing his Ph.D. degree at the Department of Intelligence and Information, Seoul National University in parallel with his career as a Senior Researcher with Hanwha Group, Republic of Korea. He is also working as a visiting researcher at both Seoul National University Hospital and Jeonbuk National University Hospital. His research interests include information systems, artificial intelligence (AI), and applications of AI in academic fields such as medical science and financial studies.

YongHyun Lee received the B.A. degree in computer engineering from the Sungkyunkwan University in 2015. He is currently pursuing his Ph.D. degree at the Department of Computer Science and Engineering, Seoul National University. He is working as a researcher at both Seoul National University Hospital and Jeonbuk National University Hospital. His research interests include artificial intelligence (AI), applications of AI, data science, graph neural network, graph classification, medical data mining and financial data mining.

Jiwoong Choi received the B.S. and M.S. degrees in Mechanical Engineering from Seoul National University in 2000 and 2005, respectively, and the Ph.D. degree in Mechanical Engineering from the University of Iowa in 2011. He is currently working as a Research Assistant Professor at the Division of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, the University of Kansas School of Medicine, Affiliate Assistant Professor at the Department of Bioengineering, the University of Kansas, and Adjunct Professor at the Department of Mechanical Engineering, the University of Iowa. His research interests focus on quantitative lung CT imaging, and computational lung modeling, using image registration and computational fluid dynamics (CFD) simulations. He uses machine learning and deep learning approaches for physiological and pathophysiological interpretation of multiscale lung structural and functional features in lung health and diseases.

Byungjoon Yoo is a Professor with the College of Business Administration, Seoul National University. Prior to joining Seoul National University, he worked at Korea University and Hong Kong University of Science and Technology. His research interests include B2B e-commerce, online auctions, and pricing strategies of digital goods such as software products and online games. He has published on these topics in journals such as Management Science, Journal of Management Information Systems, Journal of Marketing, and Decision Support Systems. He has consulting experience with Korea Stock Exchange, Korea Association of Game Industry, and other companies in which he measured the impact of information systems and online transactions, and recommended ways to use information systems strategically.

Kum Ju Chae received the B.A. and Ph.D. degree in medical science (Radiology) from the Jeonbuk National University in 2013 and 2019, respectively. She is currently working as an Assistant Professor at Department of Radiology in Jeonbuk National University Hospital. Her research interests include quantitative lung imaging, and imaging of obstructive lung diseases include chronic obstructive pulmonary disease and asthma. Her publications are directed towards studying quantitative CT features for objective structural evaluation of the lung.

Chang Hyun Lee received the B.A. and M.D. from Kyungpook National University and the Ph.D. from Seoul National University, College of Medicine. He is currently working as a professor and a chief of chest radiology at the Department of Radiology, Seoul National University Hospital. His research interests are applying quantitative and physiologic imaging and computer applications in diagnostic imaging including contrast agent usage in CT, functional and physiologic CT and MR imaging, and radiation dose monitoring and modulation techniques. His recent research includes functional CT imaging for lung structures and functions such as air trapping, emphysema, and ventilation imaging using noble gas, including xenon. Also applying image registration using inspiration and expiration CT images, air volume change, and lung motion analysis during respiration in patients with asthma, COPD, and interstitial lung diseases have been his topics.