공학석사 학위논문

# 저선량 CT에서 딥러닝을 통한 폐기능 예측 연구

# Deep learning based approach to predict pulmonary function from low-dose chest CT

울산대학교대학원

의과학과

박현정

# 저선량 CT에서 딥러닝을 통한 폐기능 예측 연구

지도교수 김남국

이 논문을 공학석사 학위 논문으로 제출함

2022 년 08 월

울산대학교대학원

의과학과

박현정

박현정의 공학석사 학위 논문을 인준함

| | | |
|---|---|---|
| 심사위원 | 이세원 | (인) |
| 심사위원 | 이준구 | (인) |
| 심사위원 | 김남국 | (인) |

울산대학교 대학원

2022 년 8 월

# Abstract

In 2011 and 2018, two large randomized controlled trials demonstrating the superiority of low-dose chest CT over chest radiography in detecting early-stage cancer and reducing lung cancer mortality were reported. Several lung cancer-related societies issued guidelines for lung cancer screening in response to these results. In Korea, the nationwide lung cancer screening program was initiated in 2019. In this screening program, smokers aged 54 to 74 with more than 30 pack-years of smoking history are advised to undertake low-dose chest CT screening for lung cancer biannually. In this context, it is anticipated that the number of low-dose chest CTs will increase.

During lung cancer screening, many pulmonary or cardiovascular diseases other than lung cancer have been reported. Chronic obstructive pulmonary disease (COPD), the third leading cause of mortality worldwide in 2019, according to the World Health Organization, is frequently found in lung cancer screening CT. Spirometry is the diagnostic gold standard for COPD with a $FEV_1$/FVC ratio of less than 70%. However, evidence for large-scale spirometry screening is not reported.

There are studies to solve the link between the morphological characteristics featured in medical images and pulmonary functions or diseases. For example, various volumetric parameters, airway tree parameters, and texture-based or radiomic features are suggested to find the association with spirometry-measured pulmonary function. Though these studies have demonstrated a correlation between the corresponding indicators and lung function, they have limitations in that human intervention is required for developing models or extracting features from images. Recently, the convolutional neural network-based deep learning approach has gained popularity in the field of medical imaging since it employs automatically extracted information during the training process and outperforms previous algorithms. Though most deep learning studies focused on detecting structural abnormalities with deep learning, recent studies showed an interest in relating structural and functional indices. For example, one recently published study utilized a deep learning network to predict pulmonary function and classify COPD-risk groups based on chest radiographs and radiologic reports. Nevertheless, the potential of CT for this purpose has not yet been investigated. Therefore, we studied the ability of convolutional neural networks to predict pulmonary function from low-dose CT scans acquired from a single center's health screening participants.

Two models were separately trained using CNN-derived features from low-dose chest CT to regress the observed values of forced vital capacity (FVC) and forced expiratory volume in one second ($FEV_1$) in liters (L). Each CT image was resampled with 2.5 mm iso-cubic voxels, and the sequences of slices in coronal directions were used as input for the model. The deep learning predicted values were normalized to yield the percent of predicted values of FVC and $FEV_1$ (FVC% and $FEV_1$%), and $FEV_1$/FVC ratio.  Agreement performance was determined for all variables. Classification performance was evaluated using the clinically accepted cutoff values of FVC% < 80, $FEV_1$% < 80, and $FEV_1$/FEV < 70% to simulate the screening capability of the developed model,

Before getting the results with the whole dataset, preliminary experiments were undertaken to determine the operational parameters of the CNN model using data from a random subset. With results of these studies, an I3D network with a GoogleNet backbone and no pre-trained weights was chosen, as well as CT intensity with a 12-bits complete range.

The pulmonary function test parameters predicted by the trained deep-learning model were compared with the spirometry-measured values, showing a higher degree of agreement in FVC than $FEV_1$. FVC% and $FEV_1$% as well as $FEV_1$/FVC exhibited lower agreement performance than that of measured values. The area under the receiver-operator-characteristics curve was 0.90 for FVC, 0.86 for $FEV_1$, and 0.85 for $FEV_1$/FVC when clinically established cutoff values were utilized to predict risk on the temporally-independent testing dataset. Applying the same cutoff settings on the deep-learning-derived values to the same testing dataset, accuracy was 89.6 % for FVC%, 85.9% for $FEV_1$%, and 90.2% for $FEV_1$/FVC ratio. Sensitivity and specificity were 61.6% and 94.3% for FVC%, 46.9% and 94.3% for $FEV_1$%, and 36.1% and 95.7% for $FEV_1$/FVC ratio. Positive predictive value and negative predictive value were 64.5% and 93.6% for FVC%, 64.0 and 89.2% for $FEV_1$%, and 46.2% and 93.6% for $FEV_1$/FVC ratio. GradCAM analysis of FVC and $FEV_1$ indicated distinct regions. GradCAM focused the anterior right lung region along the anterior chest wall for the FVC-predicting model. The left lung's middle region was also slightly marked. In contrast, the GradCAM derived from $FEV_1$ model emphasized the central areas of both lungs, especially the right lung. Additionally, the anterior and posterior regions of both lower lungs were noted.

In conclusion, models based on deep learning that predict the measured value of FVC and $FEV_1$ were developed. In addition, preliminary experiments with subsets of data were conducted to determine the operational parameters of the network. This research is anticipated to serve as a baseline for future studies that employ a deep-learning approach to extract information regarding pulmonary function from CT scans.

**Keyword:** low-dose CT, pulmonary function, spirometry, convolution neural network, deep learning, GradCAM

## Abbreviations

| | |
|---|---|
| AMC | Asan Medical Center |
| AUROC | area under receiver operating characteristic curve |
| AUPRC | area under precision recall curve |
| CNN | convolutional neural network |
| COPD | chronic obstructive pulmonary disease |
| CT | computed tomography |
| DL | deep learning |
| $FEV_1$ | forced expiratory volume in one second |
| FVC | forced vital capacity |
| LDCT | low-dose chest computerized tomography |
| PFT | pulmonary function test |

# Table of Contents

# Tables

# Figures

# 1. Introduction

## Motivations

Whenever new imaging techniques evolve, they endeavor to find the optimal use of those modalities in various contexts, not to mention in the clinical context. Through those endeavors, every imaging modality gets its seat in the medical field and is considered as the “gold standard” to diagnose diseases.

It is the gold standard to detect abnormal structure from those structural images, in CT. But still there have been persistent efforts to get the hint of biomarkers associated with various phenomena in the human body. The area of these works is not only limited to disease-related biomarkers, but also includes the functions of each organ. It started with getting some visual parameters associated with the quantified function. With the development of CNN, which shows outstanding performance in image recognition, this technique was also applied to finding the biomarker for getting the functions of various organs, including IQ of the brain(1), etc.

Since the results of two large, randomized clinical trials, the National Lung Screening Trial (NLST) and Nederlands–Leuvens Longkanker Screenings Onderzoek (NELSON), which shows a meaningful effect of lung cancer screening with low-dose CT (LDCT) on decreasing lung cancer mortality among the heavy cigarette smokers had been published, lung cancer screening utilizing LDCT became a well-established practice (2, 3)**.** This data supports the recommendation for a routine LDCT for adults aged 50 to 80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years (4). Although the risk of radiation exposure and overdiagnosis remains(5), the use of screening chest CT in the general population is expected to increase (6).

On the contrary, the role of screening spirometry is still debatable (7, 8). Chronic obstructive pulmonary disease (COPD) is one of the leading causes of death worldwide (9, 10) and the prevalence of this disease among adults is relatively high. Spirometry is widely used to diagnose early cases of COPD, but the US Preventive Services Task Force found no net benefit to implement screening spirometry for the persons before they show symptoms (8). In light of this, the Global Initiative for Chronic Obstructive Lung Disease only “advocates” active case finding as opposed to “recommending” screening spirometry in individuals who have symptoms or risk factors (11). It is challenging to perform spirometry in a mass screening scenario because it requires quality control through calibration and interpretation of acceptability and reproducibility (12).

With these recommendations, a significant portion of smokers may receive routine CT scans without spirometry. Therefore, LDCT screening could be a potential solution for the underdiagnosis of patients with chronic respiratory disorders if screening chest CT can predict lung function and identify examinees who need spirometry (13, 14). The risk factors for lung cancer, such as smoking, are shared by chronic respiratory disorders  (15), indicating the potential cost-effectiveness of CT screening. Smoking is a primary indication for CT screening.

Thoracic imaging is a relatively new application for deep learning, a subfield of machine learning. In order to detect structural anomalies with annotations like nodules, pneumothorax, atelectasis, cardiomegaly, pleural effusion, and tuberculosis, a number of automated reporting methods have been created (16). In addition to such structural diagnostics, recent research has concentrated on functional issues including spirometry standardization (17) and COPD phenotyping (18). There is a correlation between pulmonary function indices and a variety of quantitative markers obtained from a chest CT (19-27). Therefore, we proposed that deep learning applied to LDCT can predict the outcomes of spirometry.

## Related study

Some studies demonstrated the relationship between various CT-derived parameters with pulmonary functions. Those features varied in the range, including volumetry features, densitometry, features from airway segments, texture or radiomic features, and other tissues such as muscle and fat tissues.

As preoperative research to get prognosis prediction on the results of lung surgery, some studies focusing on volumetry parameters predicting pulmonary functions were performed.

The study with twenty-one healthy subjects who were candidates for lung donation shows a typical example. Chen et al. (2011) showed a significant correlation between forced vital capacity (FVC) and the total lung volume calculated from lung volumetry data from CT using simple linear regression, obtaining an $R^2$ of 0.712. The correlation between total lung volume and spirometry-measured total lung capacity (TLC) was also examined with an $R^2$ of 0.622 (22).

Iwano et al. extracted some CT parameters with CT attenuation values with thresholding. They showed that the emphysematous lung capacity, the volume of voxel under -900 HU within the lung region, is negatively correlated with $FEV_1$ with $r = 0.56$ ($p < .001$) with a simple linear regression model in the study with 64 patients with pulmonary nodules. For this study, the surrounding soft tissues and the bronchi were captured and eliminated with imaging analysis techniques (19).

Pu et al. (2012) extracted airway tree measurements such as trachea length and non-normalized total airway volumes from 548 patients with COPD to show the relationship between those values and pulmonary function parameters. This study demonstrated that nearly all pulmonary function parameters are associated with airway parameters with a significance of $p < 0.01$. They categorized pulmonary function parameters. Additionally, they showed the correlation of the airway parameters with COPD severity (27). In another study showing airway structure and lung function, Chen et al. (2016) built a model from automatically extracted airway features, suggesting characteristics relating to $FEV_1$%, such as lumen space, within-segment homogeneity of each airway, and branch angles. Their work has meaning in that they obtained the data from a group of people with normal lung functions (24).

Koo et al. demonstrated the link between PFT variables and a combination of quantitative CT parameters, such as emphysema index, air-trapping index, airway parameters (Pi10), parenchyma attenuation parameter, and lung volume changes, derived from MDCT in both exhale and inhale stage in stratified GOLD severity groups. They used automated segmentation software to derive these parameters. Their findings suggest that the association varies depending on the severity of COPD according to the GOLD criterion. They showed that the more severe COPD (higher GOLD stages), the worse the parenchymal attenuation parameters. Using multiple linear regression analysis, they also construct a model to predict $FEV_1/FVC$ or $FEV_1$ using CT parameters for each subgroup. For the total dataset, the $R^2$ values for predicting $FEV_1/FVC$ were $R^2 = 0.38$, $p < 0.001$ and $R^2 = 0.28$, $p < 0.001$ for predicting $FEV_1$. The goodness-of-fit of the chosen model and chosen variables varied based on the GOLD stage of the disease (25).

Lafata et al. (2019) focused their view on finding the correlation between radiomic features, which includes features of intensity, morphology, fractal geometry, and higher-order features, as well as texture features, and pulmonary function parameters, using univariate statistical analysis and dynamic data clustering. They focused on $FEV_1$ and $D_{LCO}$, both the absolute measured values and the percent of predicted values (23).

McDonald et al. demonstrated a significant association between FVC and $FEV_1$ and the pectoral muscle area in COPD patients. Adjusting for body mass index, age, sex, height, pack-years of smoking, and current smoking status, the $FEV_1$ and FVC percentages improved by 0.67 and 0.4 ($p < 0.001$) for each 1 $cm^2$ increase in the pectoralis muscle area (26).

These studies endeavored to get the hint of features from morphological images by finding an associating relationship. Though some studies found the features in the normal groups, most of them obtained their data from diseased groups such as COPD. Also, the results were drawn from a relatively small cohort with handicraft process. Moreover, the goal of these studies, except a few studies, was to search for the correlation between each feature and the functional information, not directly predicting pulmonary function from the CT scans.

Recently, Schroeder et al. (2020) built a CNN model using a two-view chest radiograph with PFT data predicting COPD, defined as $FEV_1/FVC$ ratio < 0.7. Comparing the models based on the two-view radiographs and the radiologic report showed the importance of imaging features in predicting pulmonary functions. However, the efficacy of the 3D features from the CT scan over the 2D features from radiographs was not tested.

### Purpose

In this study, we aimed to explore the potential of a deep learning based approach for predicting pulmonary function from low-dose chest CT. Since the deep learning approach is quite a new method for this task, we also conducted preliminary studies for finding suitable input processing method to get the optimal outcome from this approach.

# 2. Predicting Pulmonary Function from Low-dose Chest CT

## 2.1. Preliminary study with subset data for finding input parameters

Because there is no known way to predict pulmonary function directly from a low-dose CT scan using deep learning, we conducted numerous preliminary investigations to determine the appropriate operational parameters for the deep learning network and experimental parameters. Two subsets of the obtained data were used.

The first subset data (later, subset 1) was a dataset initially designed to classify the two groups representing a different level of the $FEV_1$ (% of predicted value) (later, $FEV_1$%). The same number of CT scans are sampled from the two groups: $FEV_1$% under or equal to 70% and $FEV_1$% over or equal to 80%. To get a higher contrast between the severe and normal groups, the patients' data with borderline $FEV_1$% (70% < $FEV_1$% < 80%) were excluded.  Each group was divided into training, tuning, and validation dataset with a ratio of 8:1:1 based on patient ID (Table 1). In this dataset, several patients had multiple CT scans. The second version of subset data (later, subset 2) was randomly sampled to have 3000 patient IDs and paired CT scans from January 2015 to December 2017. Only the first data were included for the patients with multiple visits during this period in this subset.

The training was completed across 150 epochs, and the model was picked from among the 100-150 epochs having the lowest tuning loss. The additional information on the other training hyper-parameters was described in 2.2.

The input parameters with better agreement performance were considered as better parameters for the network training for this task. Root mean squared error (RMSE), mean absolute error (MAE) and concordance correlation coefficient (CCC) were used to evaluated agreements between spirometry parameters and the deep learning predicted values. Classification evaluation metric such as area under receiver operating characteristics (AUROC), area under precision-recall curve (AUPRC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are used to classify respiratory risky group: $FEV_1$% < 80 and FVC% < 80.

**Table 1 Constitution of subset data 1**

| Num. of CT images<br>(Num. of patient ID) | **Total** | **$FEV_1$ % ≤70** | **$FEV_1$ % ≥80** |
|---|---|---|---|
| | 2403 (2027) | 1201 (856) | 1202 (1171) |
| Training | 1928 (1622) | 968 (685) | 960 (937) |
| Tuning | 235 (203) | 115 (86) | 120 (117) |
| Validation | 240 (202) | 118 (85) | 122 (117) |

### 1) 2D vs 3D comparison

Two experiments were undertaken to evaluate the merit of predicting pulmonary function using 3D features than the model using 2D features.

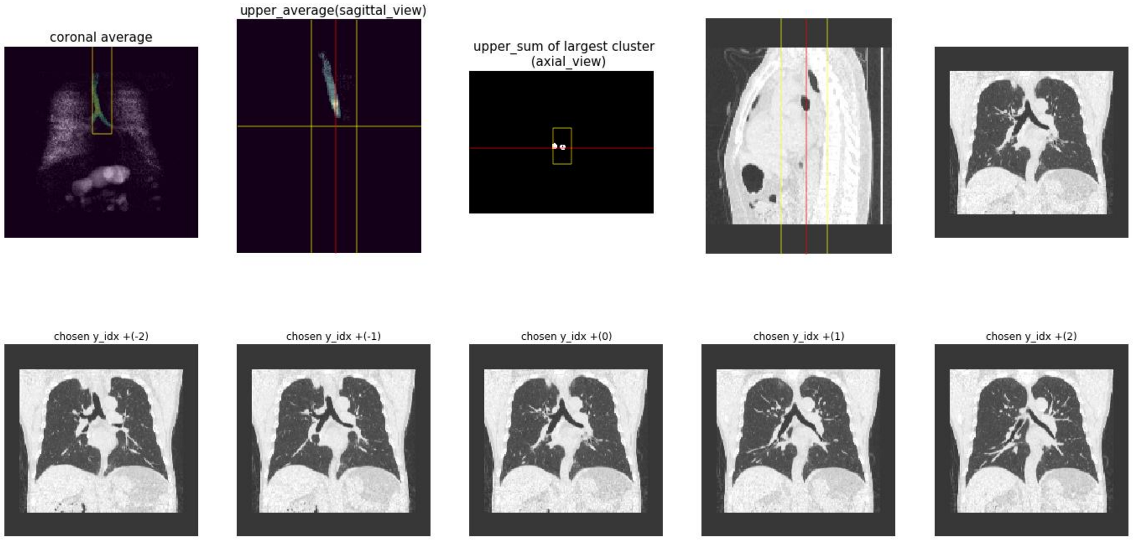


**Figure 1 Process of finding representative slice for 2D model.** The slice with carina of trachea was found by rules-based model with simple thresholding. The trachea was considered as the largest cluster in assigned region in the area.

**Materials and Methods:** Two experiments were undertaken to evaluate the merit of predicting pulmonary function using 3D features than the model using 2D features.

In the first experiment, we compared the two models employing 2D and 2.5D images, which were trained to classify risky groups over the non-risky group of $FEV_1$%: $FEV_1$% $\leq$70 as the risky group and $FEV_1$% $\geq$ 80 as the non-risky group. The 2D model was trained with a representative slice with a single channel, whereas the 2.5D model was trained using five slices, including two slices before and after the representative slice (Figure 1). Using a rule-based approach, a sample slice with the most distinct view of the carina of trachea was chosen. Both the two models utilized input images that normalized voxel values between -1450 and 50 into 0 and 1 (lung windowing, WL: -700, WW: 1500).

In the next experiments, the models trained with 2D and 3D images to regress the value of FVC (L) were compared. Two 3D models were examined: one with pre-trained weights inflated from the 2D model (3D, inflated weights in Table 3) and one without pre-trained weights (3D, scratch in Table 3). Additionally, we compared two network backbone architectures, namely GoogLeNet (Inception-v1) (28) and Inception-v3 (29). The input images were processed in the same manner in the previous experiment. Thirty-two frames, 16 frames each before and after the representative slice of carina index, were used in the 3D model. Because of the limit of GPU memory available at that time (24 GB), the training couldn’t have a fair comparison with the batch size of the model: the 2D model used a batch size of 10 while the 3D model, used 8. Augmentation of random rotation (±10 degrees) and random crop (random selection within ±20 slices) were applied.

**Results:** In the experiment comparing 2D and 2.5D, the model trained with five slices outperformed the 2D model in classification experiments, with an AUROC of 0.87 and an accuracy of 76.67 % [184 of 240], compared to 0.75 and 67.08 %, respectively (Table 2).

In trials comparing 2D and 3D, the 2D model yielded RMSE values of 0.515, 0.481, MAE values of 0.402 and 0.375, and R$^2$ values of 0.664 and 0.774 for the Inception-v1 and Inception-v3 backbones, respectively. RMSE of 0.371 and 0.388, MAE of 0.296 and 0.304, and R$^2$ of 0.826 and 0.809 were noted for the models trained with the 3D image with 32 frames utilizing inflated weights from the 2D model, respectively, with each backbone. With both backbones, the performance of the 3D model was much superior to that of the 2D model, with around 1 (L) less RMSE. Comparing the two 3D models with and without the inflated pre-trained weights from the 2D model, the model trained from scratch achieved a higher agreement with both backbone networks in experiments (Table 2 and Figure 2)

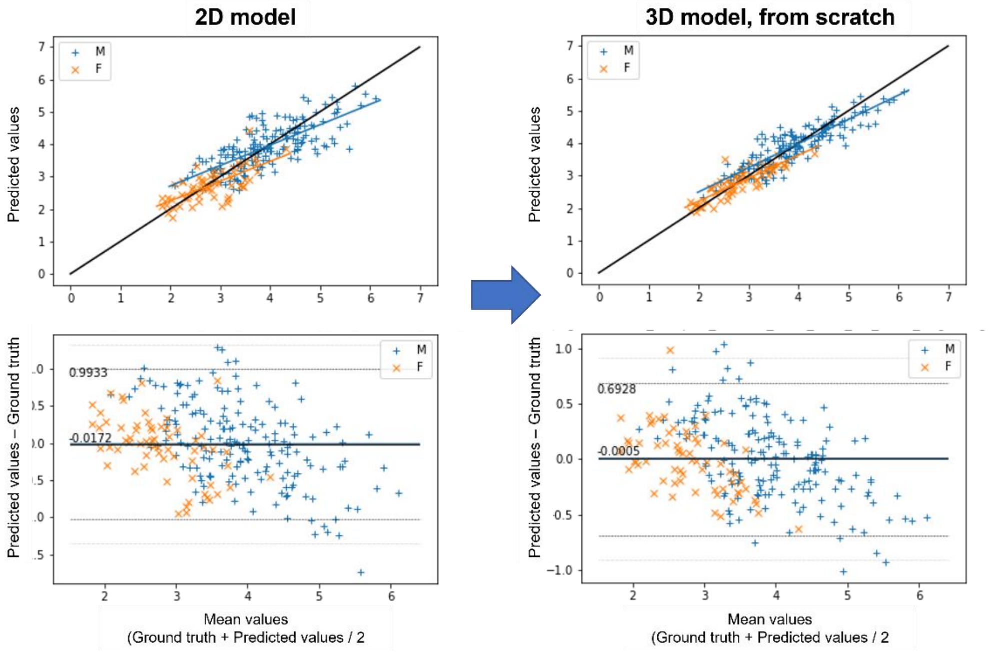


**Figure 2 Scatter and Bland-Altman plots of FVC prediction using 2D and 3D input** *(*Left) 2D model test results showing bias of -0.02 upper LOA 0.99 (L) and (Right) showing results of 3D, from scratch model with bias of -0.0005 and upper LOA 0.69 (L)

**Table 2 Result for classification based on FEV$_1$ %, 2D model, per image evaluation**

| | **With 1 slice** | **With 5 slices** |
|---|---|---|
| TP | 84 | 89 |
| FN | 34 | 29 |
| FP | 45 | 27 |
| TN | 77 | 95 |
| AUROC | 0.75 | 0.87 |
| Accuracy (%) | 67.08 | 76.67 |
| Sensitivity (%) | 71.19 | 75.42 |
| Specificity (%) | 63.11 | 77.87 |
| PPV (%) | 65.12 | 76.72 |
| NPV (%) | 69.37 | 76.61 |

**Table 3 FVC (L) prediction results of 2D and 3D models**

| **backbone network** | **evaluation metric** | **2D** | **3D, inflated weights** | **3D, scratch** |
|---|---|---|---|---|
| GoogLeNet (Inception-v1) | RMSE | 0.515 | 0.371 | **0.353** |
| | MAE | 0.402 | 0.296 | **0.282** |
| | $R^2$ | 0.664 | 0.826 | **0.842** |
| Inception-v3 | RMSE | 0.481 | 0.388 | 0.364 |
| | MAE | 0.375 | 0.304 | 0.281 |
| | $R^2$ | 0.707 | 0.809 | 0.832 |

### 2) Windowing Selection

**Materials and Method:** We compared the complete range of CT intensity from the original image (-1024, 3071), lung windowing (-1450, 50; WL: -700, WW: 1500), and mediastinum windowing (-200, 300; WL: 50, WW: 500) (Figure 3). Three models were trained to use a single input channel to predict measured value of $FEV_1$ and FVC (L), respectively. Each windowing's intensity was standardized to a range between 0 and 1.

**Results:** The results from single experiments for each setting were compared. The model trained with the image normalized with mediastinum windowing showed the worst performance with AUROC with 0,80 and 0.85 for $FEV_1$ and FVC, respectively (Table 4-5). For $FEV_1$ prediction, the agreement performance between the spirometry and deep learning predicted values were comparable for the models using the range of 12bit full intensity range and the lung windowing. With similar AUROC of 0.85 and 0.84 for 12bit and lung windowing, other metrics were all higher for the model using 12bit input (Table 4). For FVC prediction, the agreement metric for measured values (FVC (L)) and for standardized value (FVC%) were all slightly higher in the 12-bit model, but AUROC, AUPRC, and accuracy were better in the lung windowing model (Table 5).

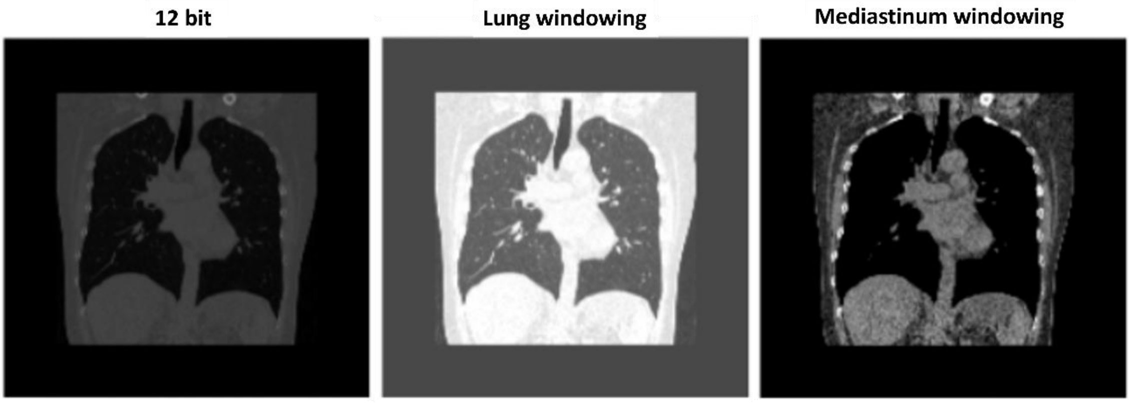


**Figure 3 Examples of various windowing selections**

**Table 4 FEV$_1$ prediction outcomes for comparing various windowing selections (anti-aliasing, 140 frame, aligned by back)**

| | 12bit range | lung | mediastinum |
|---|---|---|---|
| FEV$_1$ (L) | | | |
| RMSE | **0.302** | 0.314 | 0.323 |
| MAE | **0.236** | 0.243 | 0.249 |
| CCC | **0.884** | | 0.867 |
| FEV$_1$ %pred (%) | | | |
| RMSE | **9.06** | 9.34 | 9.81 |
| MAE | **7.14** | 7.34 | 7.57 |
| CCC | **0.665** | 0.654 | 0.579 |
| AUROC | **0.85** | 0.84 | 0.80 |
| AUPRC | **0.53** | 0.48 | 0.42 |
| Accuracy (%) | **86.1** | 84.8 | 85.0 |
| Sensitivity (%) | **47.3** | 44.0 | 38.5 |
| Specificity (%) | **93.1** | 92.1 | 93.3 |
| PPV (%) | **55.1** | 0.50 | 50.7 |
| NPV (%) | **90.8** | 90.2 | 89.4 |

**Table 5 The same as Table 4, but for FVC prediction**

| | 12bit range | lung | mediastinum |
|---|---|---|---|
| FVC (L) | | | |
| RMSE | **0.297** | 0.318 | 0.319 |
| MAE | **0.237** | 0.252 | 0.249 |
| CCC | **0.930** | 0.921 | 0.919 |
| FVC % (%) | | | |
| RMSE | **7.17** | 7.60 | 7.53 |
| MAE | **5.69** | 6.01 | 5.94 |
| CCC | **0.75** | 0.736 | 0.715 |
| AUROC | 0.88 | **0.89** | 0.85 |
| AUPRC | 0.56 | **0.60** | 0.54 |
| Accuracy (%) | 87.6 | **88.3** | 87.1 |
| Sensitivity (%) | 52.4 | 39.3 | 47.6 |
| Specificity (%) | 93.4 | 96.3 | 93.6 |
| PPV (%) | 56.4 | 63.5 | 54.8 |
| NPV (%) | 92.3 | 90.7 | 91.6 |

### 3) Number of channels

**Materials and Methods:** We compared performance of the model trained with 3 channel (3ch) or 1 channel (1ch) images. The GPU memory usage and the time consumption were also compared. The 3ch model utilized 12-bit full range, lung windowing, and mediastinum windowing and the 1ch model, 12-bit full range. 12-bit model was chosen since it has best performances among the three-windowing selection, though the difference was not tested as statistically significant. The input of each channel was rescaled to have float between the 0 to 1. The second subset data was used to derive the results.

**Results:** The experimental results for predicting $FEV_1$ (L) and FVC (L) with 100 frames were summarized in Table 6. The 3ch model shows slightly better performance with lower MAE and RMSE, and higher $R^2$. Noting that they are the results from the single experiment for each, the differences are not seemed to be statistically different considering the stochastic nature of the deep learning network. The 3-channel model needs GPU of more memory even with the limit of the input size. In Table 7, we compared the performance of each 3 and 1 channel model along with the resources they needed for training. The 3-channel model need GPU with bigger memory and more time for training[①], while the performance does not show merit with statistical significance.

**Table 6 Effects of the number of channels**

$FEV_1$ prediction trained with 100 frames

| | 3ch | 1ch |
|---|---|---|
| MAE | 0.2228 | 0.2299 |
| RMSE | 0.2891 | 0.2942 |
| $R^2$ | 0.8074 | 0.8005 |

FVC prediction trained with 100 frames

| | 3ch | 1ch |
|---|---|---|
| MAE | 0.2307 | 0.2338 |
| RMSE | 0.3000 | 0.3027 |
| $R^2$ | 0.8666 | 0.8641 |

[①] Time consumption varies and can be decreased depending on the efficiency of the scripts, but still will be proportional to written among of the time. The one channel model with 140 frames can be occasionally run on the 24GB GPU, but failed sometimes with OOM error.

**Table 7 Memory and time efficiency of 3- and 1-channel models.**

| Number of input channels | Number of frames | RMSE | MAE | $R^2$ | GPU memory used | Time consumption |
|---|---|---|---|---|---|---|
| 3ch | 100 | 0.289 | 0.223 | 0.807 | 48GB | 167 epoch / 7 days |
| 1ch | 120 | 0.290 | 0.226 | 0.806 | 24GB | 166 epoch / 1.5 days |
| 1ch | 140 | 0.293 | 0.226 | 0.802 | 32GB / 24GB | 200 epoch / 4 days |

### 4) Alignment of the body

**Materials and Methods:** The two models trained to predict $FEV_1$ with different body positions were evaluated to examine the effect of the input images' body position on the performance of CNN. The body segments were either located at the center or the back line of the coronal axis, respectively. The experiment was performed with subset data 2, with 140 frames of single channel normalized to the range of (0, 1) from the 12bit full range (-1024, 3071). RMSE were used for the comparison.

**Results**: RMSE from the model using body-centered image achieved 0.293 L), while the others with back-positioned image achieved 0.3060 (Table 8).

**Table 8 Experimental results by the location of the body segment**

| Body alignment y-axis | Num. of frames | Num. of channels | RMSE (L) |
|---|---|---|---|
| Centered | 140 | 1 | 0.2929 |
| Back-aligned | 140 | 1 | 0.3060 |

### 5) Number of frames

**Materials and Methods:** The models trained with various number of frames (image size) to predict $FEV_1$ were compared. In the first experiment, the body segments in each image were located in the center of the y axis and the frames are chosen symmetrically for the location of chosen representative slice used in experiment 1) 2D vs 3D comparison. For the experiment of 120 frames, we eliminated 10 frames each at the beginning and the end position. In the second experiment, the image with back-positioned body were used. Input image was sequentially chosen from the back position depending on the number of frames. Randomly chosen toy dataset with number of 3000 patient data used in these experiments (subset data 2).

**Results:** The performances among the model using the 100 – 140 frames sampled from the central region didn't show significant differences in predicting both $R^2$ of 0.865 and 0.864 for predicting FVC, 0.801 to 0.806 for predicting $FEV_1$ with (Table 9 – 10). It showed better agreement and classification performance with 140 frames with the image where body segments aligned with the back side, though the statistical significance still is not guaranteed (Table 11 – 12).

**Table 9 FVC (L) regression results with body-centered images over different number of frames.**

| | 120 frames | 100 frames | 64 frames |
|---|---|---|---|
| RMSE | 0.302 | 0.303 | 0.317 |
| MAE | 0.236 | 0.234 | 0.239 |
| $R^2$ | 0.865 | 0.864 | 0.852 |

**Table 10 The same as Table 9, but for $FEV_1$ (L).**

| | 140 frames | 120 frames | 100 frames | 64 frames |
|---|---|---|---|---|
| RMSE | 0.293 | 0.290 | 0.294 | 0.316 |
| MAE | 0.226 | 0.225 | 0.230 | 0.247 |
| $R^2$ | 0.802 | 0.806 | 0.801 | 0.770 |

**Table 11 FVC (L) regression results with back-aligned images over different number of frames.**

| | 140 frames | 120 frames | 100 frames |
|---|---|---|---|
| FVC (L) | | | |
| RMSE | 0.297 | 0.338 | 0.306 |
| MAE | 0.237 | 0.266 | 0.237 |
| CCC | 0.930 | 0.912 | 0.928 |
| | | | |
| FVC %pred (%) | | | |
| RMSE | 7.17 | 8.15 | 7.29 |
| MAE | 5.69 | 6.37 | 5.67 |
| CCC | 0.75 | 0.70 | 0.75 |
| AUROC | **0.88** | **0.88** | **0.89** |
| AUPRC | **0.56** | **0.56** | **0.59** |
| Accuracy (%) | 87.6 | 88.0 | 88.48 |
| Sensitivity (%) | 52.4 | 32.1 | 48.8 |
| Specificity (%) | 93.4 | 97.1 | 85.0 |
| PPV (%) | 56.4 | 64.3 | 61.2 |
| NPV (%) | 92.3 | 89.8 | 91.9 |

**Table 12 The same as Table 11, but for FEV$_1$ (L).**

| | 140 frames | 120 frames | 100 frames |
|---|---|---|---|
| FEV$_1$ (L) | | | |
| RMSE | 0.302 | 0.310 | 0.320 |
| MAE | 0.236 | 0.243 | 0.244 |
| CCC | 0.884 | 0.880 | 0.870 |
| | | | |
| FEV$_1$ %pred (%) | | | |
| RMSE | 9.06 | 9.36 | 9.62 |
| MAE | 7.14 | 7.36 | 7.39 |
| CCC | 0.665 | 0.648 | 0.619 |
| AUROC | **0.85** | **0.81** | **0.81** |
| AUPRC | **0.53** | **0.42** | **0.45** |
| Accuracy (%) | 86.1 | 82.8 | 85.0 |
| Sensitivity (%) | 47.3 | 38.5 | 41.8 |
| Specificity (%) | 93.1 | 90.7 | 92.7 |
| PPV (%) | 55.1 | 42.7 | 50.7 |
| NPV (%) | 90.8 | 89.2 | 90.0 |

## Discussion

In our preliminary experiments, we found that models with 2.5D or 3D features predicted spirometry parameters substantially better than 2D models. However, the effect of pre-trained models with inflated weights seems unclear. Comparing the windowing options, the model that normalized the complete 12bit range of the input array to 0 to 1 had the highest agreement score CCC of 0.88 and 0.93 for $FEV_1$ and FVC prediction, respectively, followed by lung and mediastinum windowing. However, the classification scores did not always have same order across the windowing selection.

Comparing the results of the 3ch model and the 1ch model with various number of frames, the RMSE, MAE, and $R^2$ showed superior performance in 3ch models; however, the difference between the RMSE and MAE values for the model with three channels and the model with a single channel did not exceed 0.015 (L) for any of the results. In addition, the single-channel model with more frames demonstrated equivalent or superior performance while consuming less GPU memory and time consumption.

For models employing input images from the central region, the number of frames had no effect on the outcomes. In contrast, the model that utilized the input sequentially from the back had marginally superior performance, although statistical significance cannot be guaranteed. $FEV_1$ prediction models appear to have a larger difference. In a previous study, Kwack et al. demonstrated a negative correlation between CT-measured thoracic fat volume and $FEV_1$% and $FEV_1$/FVC percent in a cohort of 18-80-year-old health screening participants (21). McDonald et al. demonstrated a substantial association between FVC and $FEV_1$ and the area of the pectoral muscle in COPD patients (26). Neither work addresses the measured values of FVC and $FEV_1$, but we can guess that there is substantial correlation between muscle or fat composition measured in CT and pulmonary function.  Since the body segment is aligned with the back, the effect of removing the frontal portion of the body could not be the same for all patients, with larger groups being affected more than smaller groups. Muscle and fat from the omitted slices may have a stronger correlation with pulmonary function, and additional research will be required to confirm this hypothesis.

For these preliminary studies, we conducted only one experiment in a single environment using a randomly selected portion of data. Due to the stochastic nature of finding the optimal minimum for the neural network, caution is required when comparing the results from single trials.

## 2.2 Prediction of Pulmonary function with chosen parameters

### Material and Methods

#### 1) Dataset

We retrospectively obtained the patients' data from the health screening center in Asan Medical Center, Seoul, the Republic of Korea (later, AMC), from January 2015 to December 2018. The examinees who underwent low-dose chest CT and spirometry on the same day were chosen. Only the first records were included for examinees who had taken the examination multiple times during the study period. The data acquired from January to December 2018 were set aside as temporally-independent testing dataset (n = 2720), and the data from January 2015 to December 2017 were used as a development dataset (n = 13,428). The development dataset was split into training (n = 9,394; 70%), tuning (n = 1,343; 10%), and validation (n = 2,687; 20%, later, internal validation) dataset (Figure 4).

The Institutional review board of AMC approved this retrospective study and waived the requirement for informed written consent (IRB no. 2019-0061).

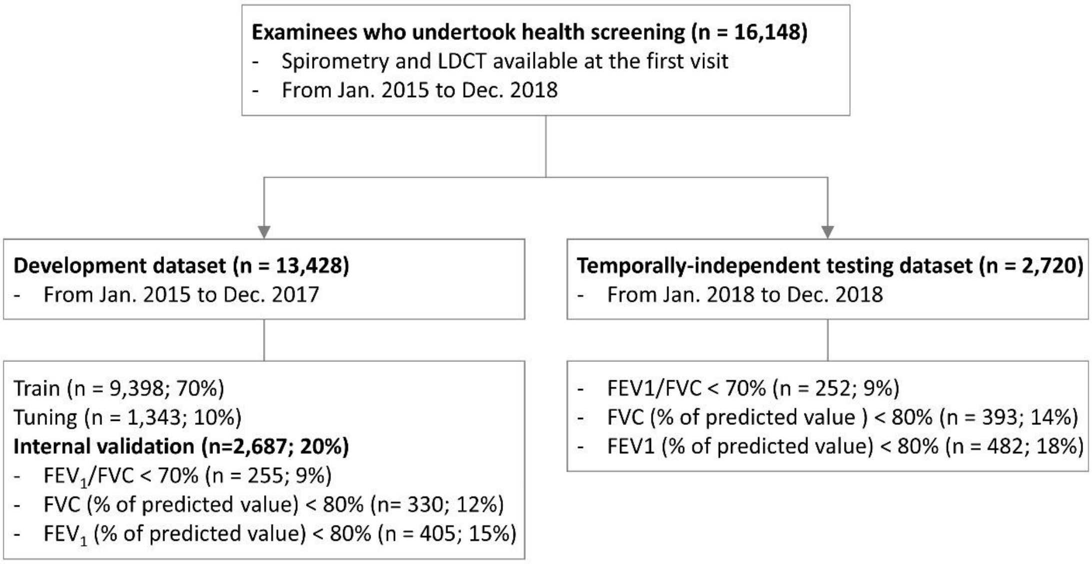

Examinees who undertook health screening (n = 16,148)
- Spirometry and LDCT available at the first visit
- From Jan. 2015 to Dec. 2018
Development dataset (n = 13,428)
- From Jan. 2015 to Dec. 2017
Temporally-independent testing dataset (n = 2,720)
- From Jan. 2018 to Dec. 2018
Train (n = 9,398; 70%)
Tuning (n = 1,343; 10%)
Internal validation (n=2,687; 20%)
- FEV$_1$/FVC < 70% (n = 255; 9%)
- FVC (% of predicted value) < 80% (n= 330; 12%)
- FEV$_1$ (% of predicted value) < 80% (n = 405; 15%)
- FEV1/FVC < 70% (n = 252; 9%)
- FVC (% of predicted value ) < 80% (n = 393; 14%)
- FEV1 (% of predicted value) < 80% (n = 482; 18%)

**Figure 4 Data split of collected dataset.**

### 2) Low-dose CT scan

The examinees underwent chest CT using SOMATOM Definition Flash CT system (Siemens Healthineers, Forchheim, Germany), LightSpeed VCT or Discovery CT750HD (GE Healthcare Technologies, Milwaukee, WI). We employed 2.5-mm thickness axial CT images of the entire thorax with full inspiration (512 x 512 matrix; 64 x 0.6mm or 64 x 0.625mm collimation; 20 or 25 mAs at 120 kV; 1 pitch; I50f or CHST kernel).

### 3) Spirometry

An experienced laboratory technician performed spirometry using a vMax 20 spirometer (Viasys, San Diego, CA, USA) in accordance with ATS/ERS guidelines (12). FVC and $FEV_1$ were measured by spirometry in liters (FVC, $FEV_1$ in Figure 5). The percentage of predicted values (FVC% and $FEV_1$% in Figure 5) were calculated using the predicted values representing the normal population, which were determined using reference equations derived from representative samples of the Korean population (30). Normal spirometry results were determined by a pre-bronchodilator $FEV_1$/FVC ratio over 70 % and FVC% and $FEV_1$% over 80% of the predicted values. Airflow limitation was determined as a $FEV_1$/FVC ratio under 70%.

### 4) Training deep learning model

The deep morphology models predicting PFT outcomes from LDCT scan are based on the inflated 3D ConvNet (I3D) (31) having global average pooling and one fully connected layer of 512 nodes instead of the final fully connected layer which was used in original I3D network (Figure 6). Two distinct models were trained to predict each values of FVC and $FEV_1$ upon their paired low-dose CT via a linear output layer by minimizing the mean squared error loss (Figure 5).

Due to the restricted processing capabilities of existing graphics processing units (GPU), we were unable to exploit CT scans with the original resolution, therefore we resampled them to have 2.5mm iso-voxel. After resizing each CT scan, the images were adjusted using the detected body area to be aligned by their posterior, and then cropped or padded to have 180x140x178 size. Our deep morphology models utilized the coronal image sequences of resampled CT scans as input pictures. NVIDIA Tesla V100 was utilized to train the deep morphology models (32GB). The initial learning rate for the Adam optimizer was 0.001 and then was reduced with a factor of 0.7 when a metric has stopped improving for 10 epochs. A batch size of 10 was used.

Using the predicted FVC and $FEV_1$ by deep learning models ($FVC_{DL}$ and $FEV_{1DL}$), we simulated the respiratory high-risk group's screening test with $FVC\%_{DL}$, $FEV_1\%_{DL}$, and $FEV_1/FVC_{DL}$ ratio. $FVC\%_{DL}$ and $FEV_1\%_{DL}$ were computed using Korean reference equations in the same manner as the gold standard (30). Classification of the respiratory high-risk group was based on normal spirometry findings defined in the section above.

### 5) Interpretation of the convolution neural network with GradCAM

To interpret how the proposed deep morphology models made the predictions, we employed Grad-CAM which visualizes the saliency area through extracting features in a layer of CNN. For capturing the importance, we computed the gradient for the linear output layer with respect to feature map of the concatenate layer after the 5th inception module.

GradCAM was suggested to understand and visualize the inner logic of deep learning models using convolutional layers. It highlights where a convolution layer's assigned feature map highly influences output of the network. Since the output value of our model was not the likelihood for potential class but the value linearly proportional to the abstraction of the feature map, the interpretation of the GradCAM was not straightforward. We interpreted the changes in the region in a feature map has a positive correlation in predicting the higher value of each output variable.

### 6) Evaluation metrics for the model

The performance of the proposed deep learning method was assessed on two distinct datasets: an internal validation dataset and a temporally-independent testing dataset. The mean absolute error (MAE), root mean squared error (RMSE), and concordance correlation coefficient (CCC) were computed to determine agreements between ground truth and anticipated PFT values. Also evaluated were Bland-Altman plots with bias estimates and 95% limits of agreement (LOA). Area under the receiver-operating-characteristics curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were utilized to evaluate the classification performance based on the clinically established cut-off values mentioned in section *3) Spirometry*. Unless otherwise indicated, values are presented as the mean ± standard deviation.

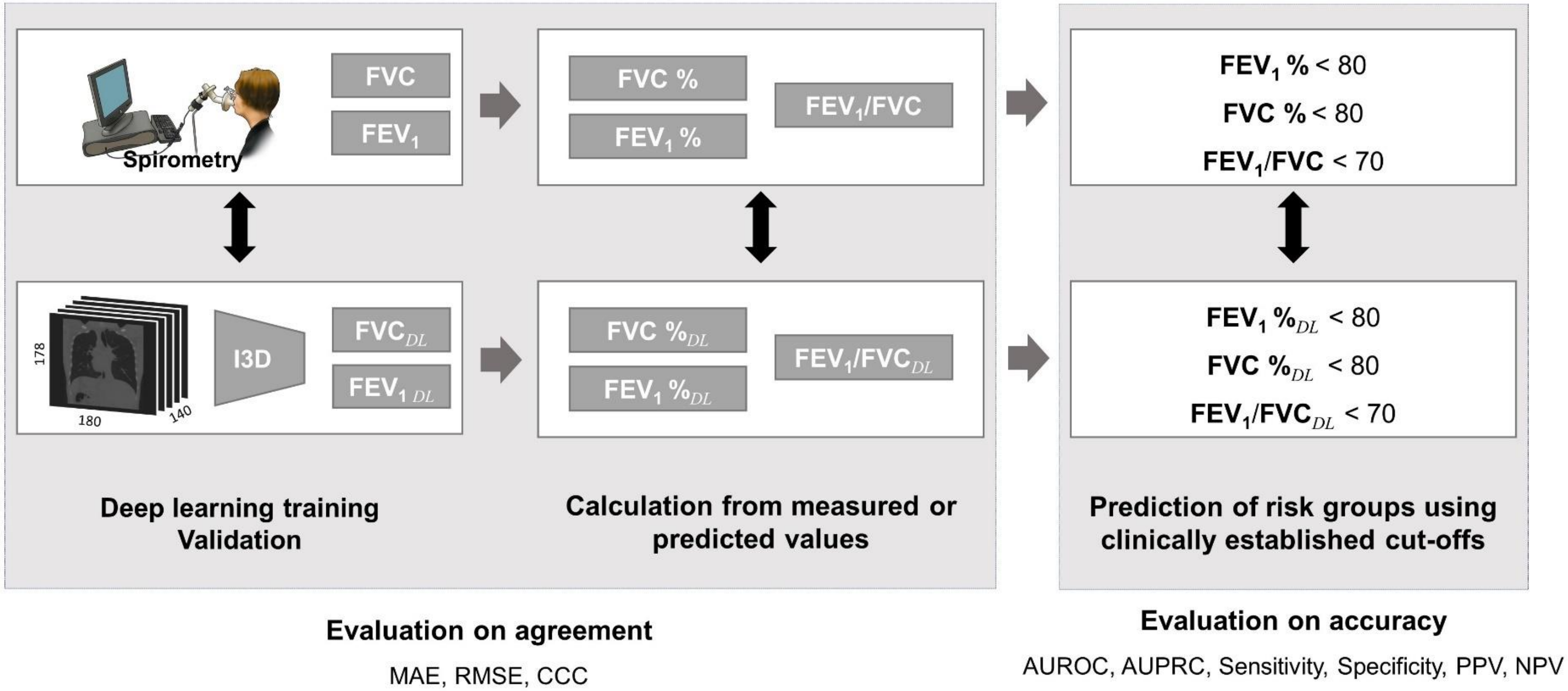


**Figure 5 Design of deep learning system for predicting pulmonary function tests from chest CT.**

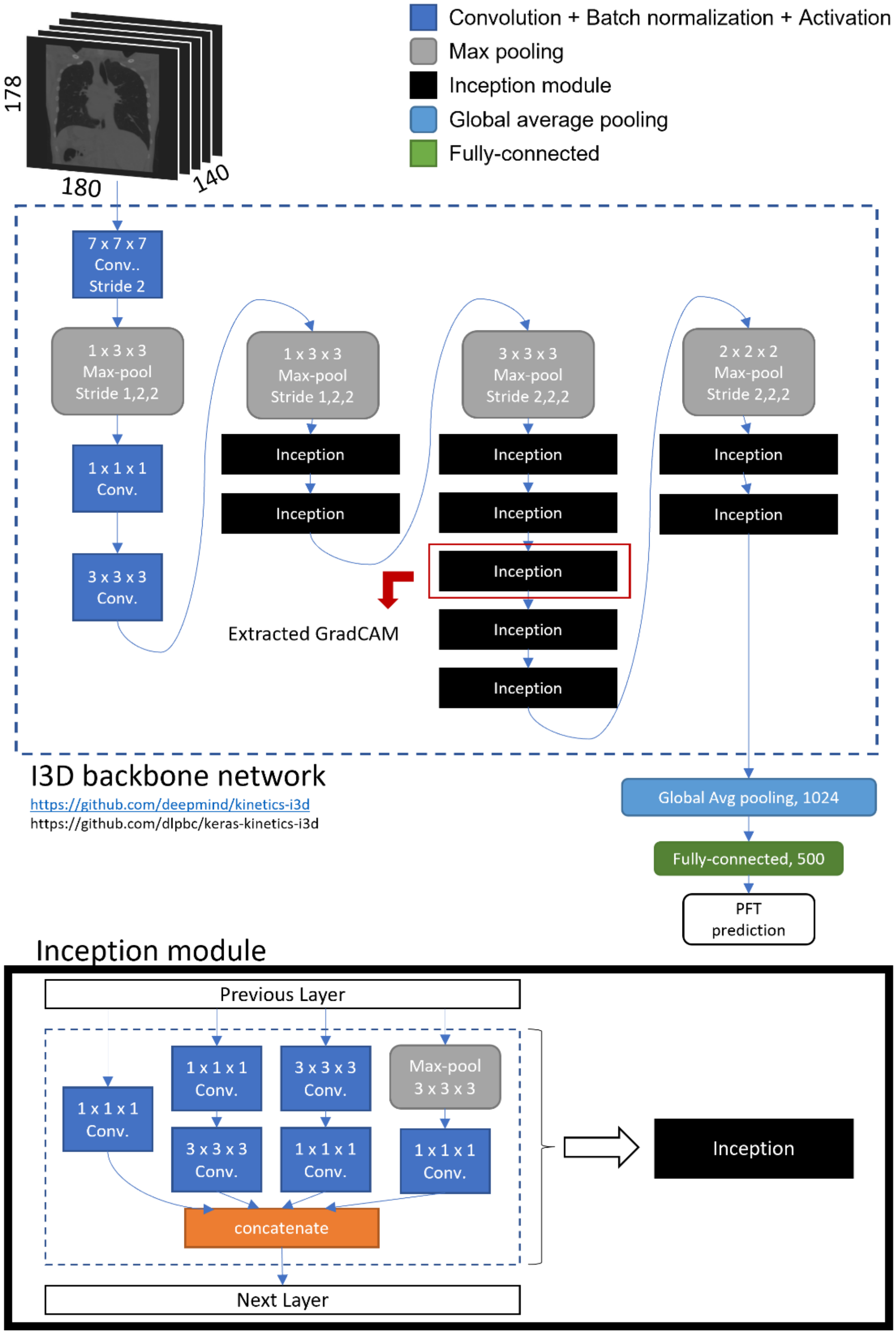


**Figure 6 I3D network architecture.**

## Results

### 1) Clinical characteristics of the study population

A total of 16148 examinees were split into two cohorts based on the time of their first visit: a development dataset of 13428 examinees (January 2015 to December 2017) and a temporally independent testing dataset of 2720 examinees (January 2018 to December 2018; Figure 4). The mean ages were 54.6 ± 9.7 and 54.1 ± 10.9 years in the development and testing datasets, respectively. There were 9238 (68.8%) male subjects and 5848 (58.5%) ever smokers in development dataset, while 1742 (54.0%) males and 1480 (54.4%) ever smokers in testing dataset. 1%–6% of the participants had a history of a chronic respiratory condition such as COPD, asthma, or tuberculosis. $FEV_1$, FVC, and $FEV_1$/FVC values in the development set were 3.04 ± 0.67 L, 3.92 ± 0.83 L, and 77.8 ± 6.8 percent, respectively. In the testing set, these values were 3.85 ± 0.84 L, 2.99 ± 0.67 L, and 77.8 ± 6.7 percent, respectively. (Table 13)

### 2) Prediction performance of the deep learning model

The prediction performance measures of our proposed models on the development and temporally independent testing dataset are summarized in Table 14 and Figures 7 and 8. The MAE, RMSE, and CCC were 0.223, 0.286, and 0.935, respectively, for the agreement between FVC and $FVC_{DL}$, and 0.216, 0.276, and 0.902, for the agreement between $FEV_1$ and $FEV_{1DL}$. This prediction was fairly robust in the temporally-independent testing dataset, with MAE, RMSE, and CCC of 0.220, 0.286, and 0.940 for FVC versus $FVC_{DL}$, and 0.218, 0.277, and 0.907 for $FEV_1$ versus $FEV_{1DL}$.

In the internal validation, the MAE, RMSE, and CCC were 5.251, 6.716, and 0.783, respectively, for agreements between FVC% and $FVC\%_{DL}$, 6.501, 8.282, and 0.708 for agreements between $FEV_1$% and $FEV_1\%_{DL}$, and 4.703, 6.084, and 0.580 for agreements between $FEV_1$/FVC and $FEV_1/FVC_{DL}$. Applying the same for the temporally-independent testing dataset; the MAE, RMSE, and CCC were 5.239, 6.798, and 0.787, respectively, for FVC% versus $FVC\%_{DL}$, 6.658, 8.484, and 0.688 for $FEV_1$% versus $FEV_1\%_{DL}$, and 4.768, 6.101, and 0.578 for $FEV_1$/FVC versus $FEV_1/FVC_{DL}$.

**Table 13 Baseline characteristics of the study population.**

| | Development dataset (n = 13 428) | Temporally-independent testing dataset n = 2720) | p value |
|---|---|---|---|
| Age, years | 54.6 ± 9.7 | 54.1 ± 10.9 | <.05 |
| Sex, male | 9238 (68.8%) | 1742 (64.0%) | <.01 |
| Body mass index (kg/m$^2$) | 23.9 ± 3.1 | 23.9 ± 3.1 | .39 |
| Smoking history | | | <.01 |
| Current smoker | 3380 (25.2%) | 644 (23.7%) | |
| Ex-smoker | 4468 (33.3%) | 836 (30.7%) | |
| Never smoker | 5555 (41.4%) | 1239 (45.6%) | |
| Smoking amount (smokers only), pack years | 28.8 ± 17.5 | 33.5 ± 17.5 | |
| Respiratory disease history | | | |
| Tuberculosis | 715 (5.3%) | 149 (5.5%) | .75 |
| Asthma | 312 (2.3%) | 70 (2.6%) | .43 |
| COPD | 158 (1.2%) | 30 (1.1%) | .74 |
| Lung function | | | |
| FEV$_1$ (L) | 3.04 ± 0.67 | 2.99 ± 0.67 | <.01 |
| FEV$_1$ (% of pred.) | 90.4 ± 11.7 | 89.6 ± 11.5 | <.01 |
| FVC (L) | 3.92 ± 0.83 | 3.85 ± 0.84 | <.01 |
| FVC (% of pred.) | 91.0 ± 10.8 | 90.4 ± 11.2 | <.05 |
| FEV$_1$/FVC (%) | 77.8 ± 6.8 | 77.8 ± 6.7 | .97 |

Note. All data are presented as mean ± standard deviation or number (%), unless otherwise indicated.
Data were compared using Pearson’s Chi-squared test and Welch’s *t*-test.
FEV$_1$, forced expiratory volume in one second; FVC, forced vital capacity.

### 3) Risk prediction using the deep learning model

The classification performance for risk groups according to FVC%$_{DL}$, FEV$_1$%$_{DL}$, and FEV$_1$/FVC$_{DL}$ are summarized in Table 15 and confusion matrices, ROC curves, and precision-recall curves for those classification are shown in Figure 9-11.

In the internal validation, the classification of the respiratory high-risk group achieved AUROC and AUPRC of 0.91 and 0.62 for FVC%, 0.87 and 0.59 for FEV$_1$%, and 0.84 and 0.45 for FEV$_1$/FVC. The same cutoff settings were applied on the deep-learning-derived values to derive following results. Accuracy was 88.9 % [2388 of 2687] for FVC%, 87.4% [2348 of 2687] for FEV$_1$%, and 90.5% [2431 of 2687] for FEV$_1$/FVC ratio. Sensitivity and specificity were 64.2% [212 of 330] and 92.3% [2176 of 2357] for FVC%, 49.9% [202 of 405] and 94.0% [2146 of 2282] for FEV$_1$%, and 40.8% [104 of 255] and 95.7% [2327 of 2432] for FEV$_1$/FVC ratio.  Positive predictive value and negative predictive value were 53.9% [212 of 393] and 94.9% [2176 of 2294] for FVC%, 59.8% [202 of 338] and 91.4% [2146 of 2349] for FEV$_1$%, and 49.8% [104 of 209] and 93.9% [2327 of 2478] for FEV$_1$/FVC ratio.

For the temporally-independent testing dataset, the same classification scheme was applied and achieved robust results. AUROC and AUPRC were achieved as 0.90 and 0.68 for FVC%, 0.86 and 0.61 for FEV$_1$%, and 0.85 and 0.40 for FEV$_1$/FVC ratio. Applying the same cutoff settings on the deep-learning-derived values to the same testing dataset, accuracy was 89.6 [2436 of 2720] % for FVC%, 85.9% [2337 of 2720] for FEV$_1$%, and 90.2% [2453 of 2720] for FEV$_1$/FVC ratio. Sensitivity and specificity were 61.6% [242 of 393] and 94.3% [2194 of 2327] for FVC%, 46.9% [226 of 482] and 94.3% [2111 of 2238] for FEV$_1$%, and 36.1% [91 of 252] and 95.7% [2362 of 2468] for FEV$_1$/FVC ratio.  Positive predictive value and negative predictive value were 64.5% [242 of 375] and 93.6% [2194 of 2345] for FVC%, 64.0% [226 of 353] and 89.2% [2111 of 2367] for FEV$_1$%, and 46.2% [91 of 197] and 93.6% [2362 of 2523] for FEV$_1$/FVC ratio.

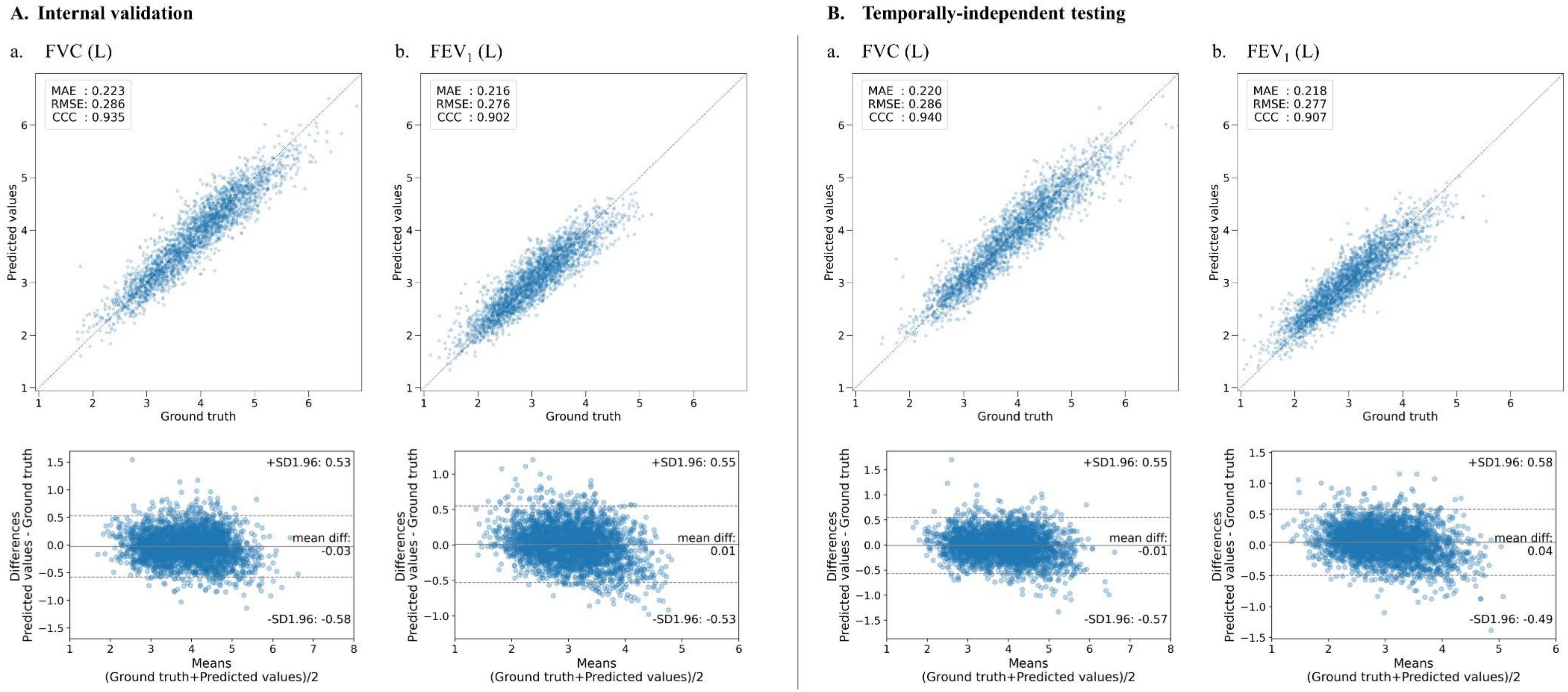


**Figure 7 Agreement between the deep learning-predicted FVC (L) and FEV$_1$ (L) and spirometry-measured values** for internal validation (A) and temporally-independent testing (B). Upper panels show ground truth values versus predicted values of, where the diagonal lines of scatter plots represent the ideal lines for perfect prediction. Lower panels show Bland-Altman plots.

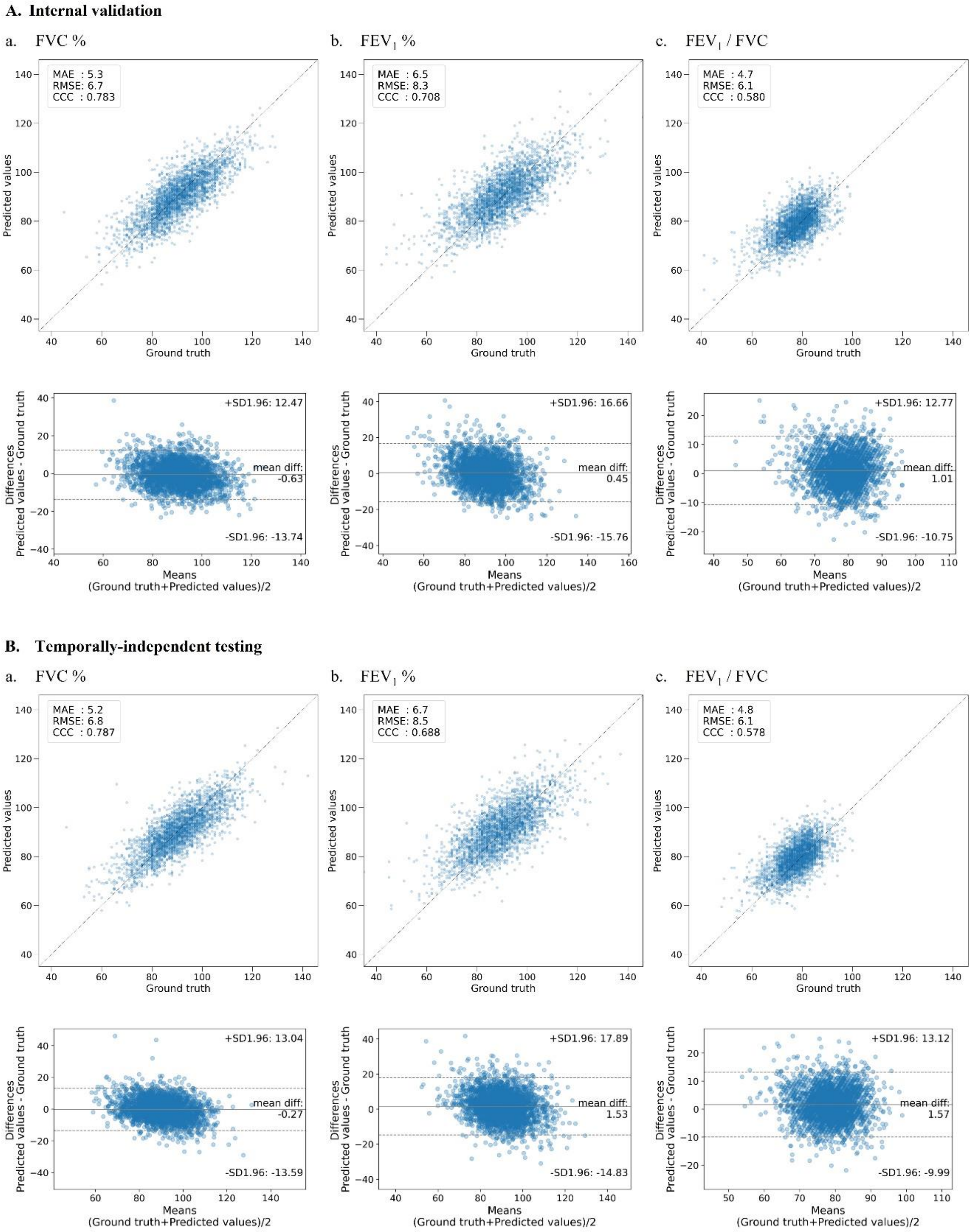


**Figure 8 The same as Figure 7, but for FVC%, FEV$_1$%, and FEV$_1$/FVC.**

**Table 14 Performance of the deep learning model for predicting PFT results**

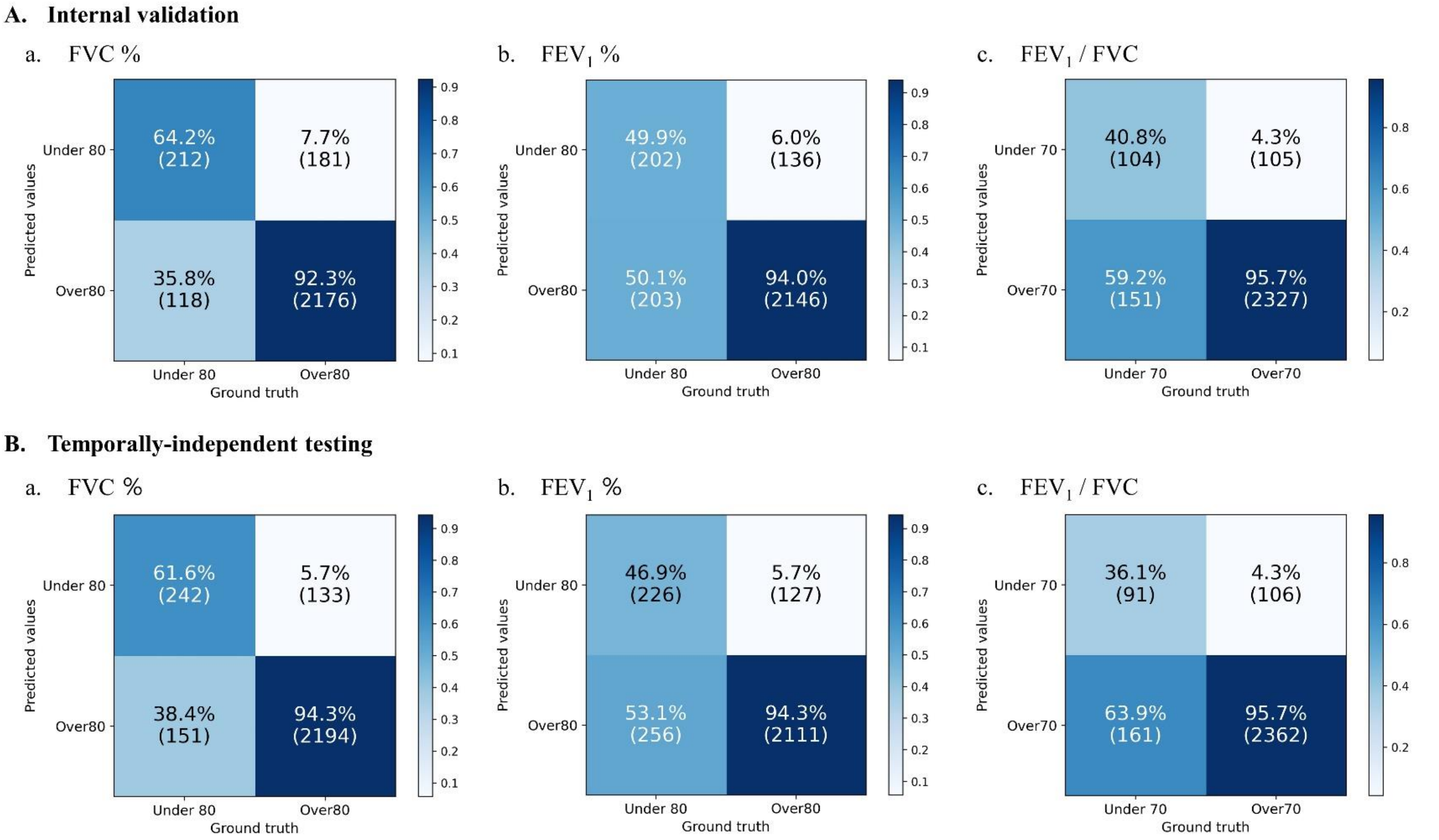
| | Internal validation | | | | | Temporally-independent testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | CCC | Bias | LOA | MAE | RMSE | CCC | Bias | LOA |
| FVC (L) | 0.223 | 0.286 | 0.935 | −0.03 | (−0.58, 0.53) | 0.220 | 0.286 | 0.940 | −0.01 | (−0.57, 0.55) |
| $FEV_1$ (L) | 0.216 | 0.276 | 0.902 | 0.01 | (−0.53, 0.55) | 0.218 | 0.277 | 0.907 | 0.04 | (−0.49, 0.58) |
| FVC (% or pred.) | 5.3 | 6.7 | 0.783 | −0.63 | (−13.74, 12.47) | 5.2 | 6.8 | 0.787 | −0.27 | (−13.59, 13.04) |
| $FEV_1$ (% of pred.) | 6.5 | 8.3 | 0.708 | 0.45 | (−15.76, 16.66) | 6.7 | 8.5 | 0.688 | 1.53 | (−14.83, 17.89) |
| $FEV_1$/FVC (%) | 4.7 | 6.1 | 0.580 | 1.01 | (−10.75, 12.77) | 4.8 | 6.1 | 0.578 | 1.57 | (−9.99, 13.12) |

Note. $FEV_1$, forced expiratory volume in one second; FVC, forced vital capacity; MAE, mean absolute error; RMSE, root mean square error; CCC, concordance correlation coefficient. Bias and LOA (limits of agreement) are from the Bland-Altman plot.

**Table 15 Risk prediction performance of the deep learning model**

| | Internal validation | | | | | | | Temporally-independent testing | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUROC | AUPRC | Acc. | Sen. | Spec. | PPV | NPV | AUROC | AUPRC | Acc. | Sen. | Spec. | PPV | NPV |
| FVC (% of pred.) | 0.91 | 0.62 | 88.9 | 64.2 | 92.3 | 53.9 | 94.9 | 0.90 | 0.68 | 89.6 | 61.6 | 94.3 | 64.5 | 93.6 |
| $FEV_1$ (% of pred.) | 0.87 | 0.59 | 87.4 | 49.9 | 94.0 | 59.8 | 91.4 | 0.86 | 0.61 | 85.9 | 46.9 | 94.3 | 64.0 | 89.2 |
| $FEV_1$/FVC ratio | 0.84 | 0.45 | 90.5 | 40.8 | 95.7 | 49.8 | 93.9 | 0.85 | 0.40 | 90.2 | 36.1 | 95.7 | 46.2 | 93.6 |

Note. $FEV_1$, forced expiratory volume in one second; FVC, forced vital capacity; AUROC, area under the receiver operating characteristics curve; AUPRC, area under the precision-recall curve; Acc., accuracy; Sen., sensitivity; Spec., specificity; PPV, positive predictive value; NPV, negative predictive value.



A. Internal validation

a. FVC %

b. $FEV_1$ %

c. $FEV_1$ / FVC

B. Temporally-independent testing

a. FVC %

b. $FEV_1$ %

c. $FEV_1$ / FVC

**Figure 9 Confusion matrices for risk prediction on FVC%, $FEV_1$%, and $FEV_1$/FVC** for internal validation (A) and temporally-independent testing (B). Risky groups are categorized using deep learning model predictions and compared with ground-truth spirometry results using the same clinical cut-off values. Each cell shows sensitivity, false positive ratio, false negative ratio, and specificity. True positive, false positive, false negative, true negative, in that order.

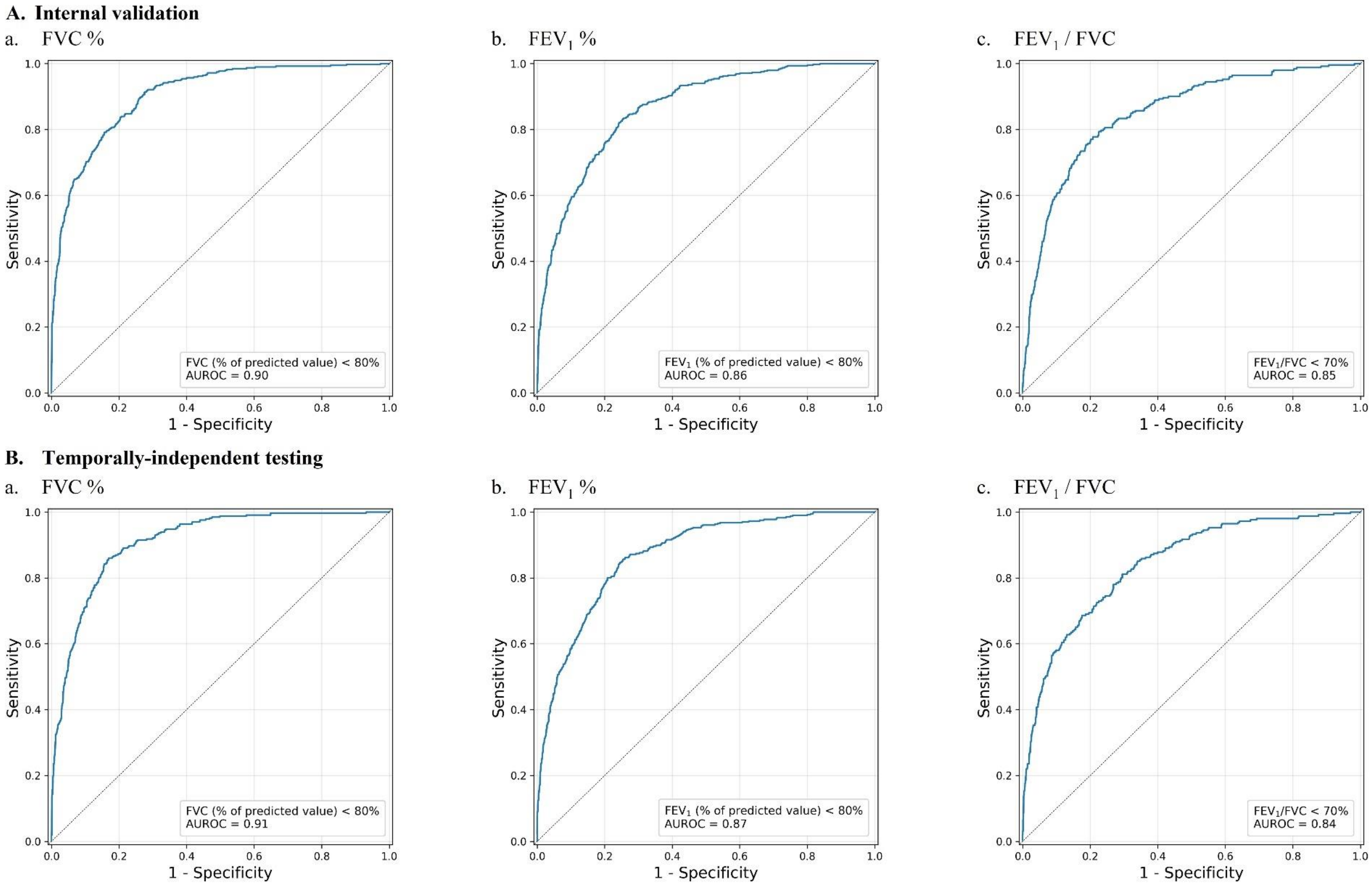


**Figure 10 Receiver operating characteristic curves for risk prediction on FVC%, FEV$_1$%, and FEV$_1$/FVC for internal validation (A) and temporally-independent testing (B).**

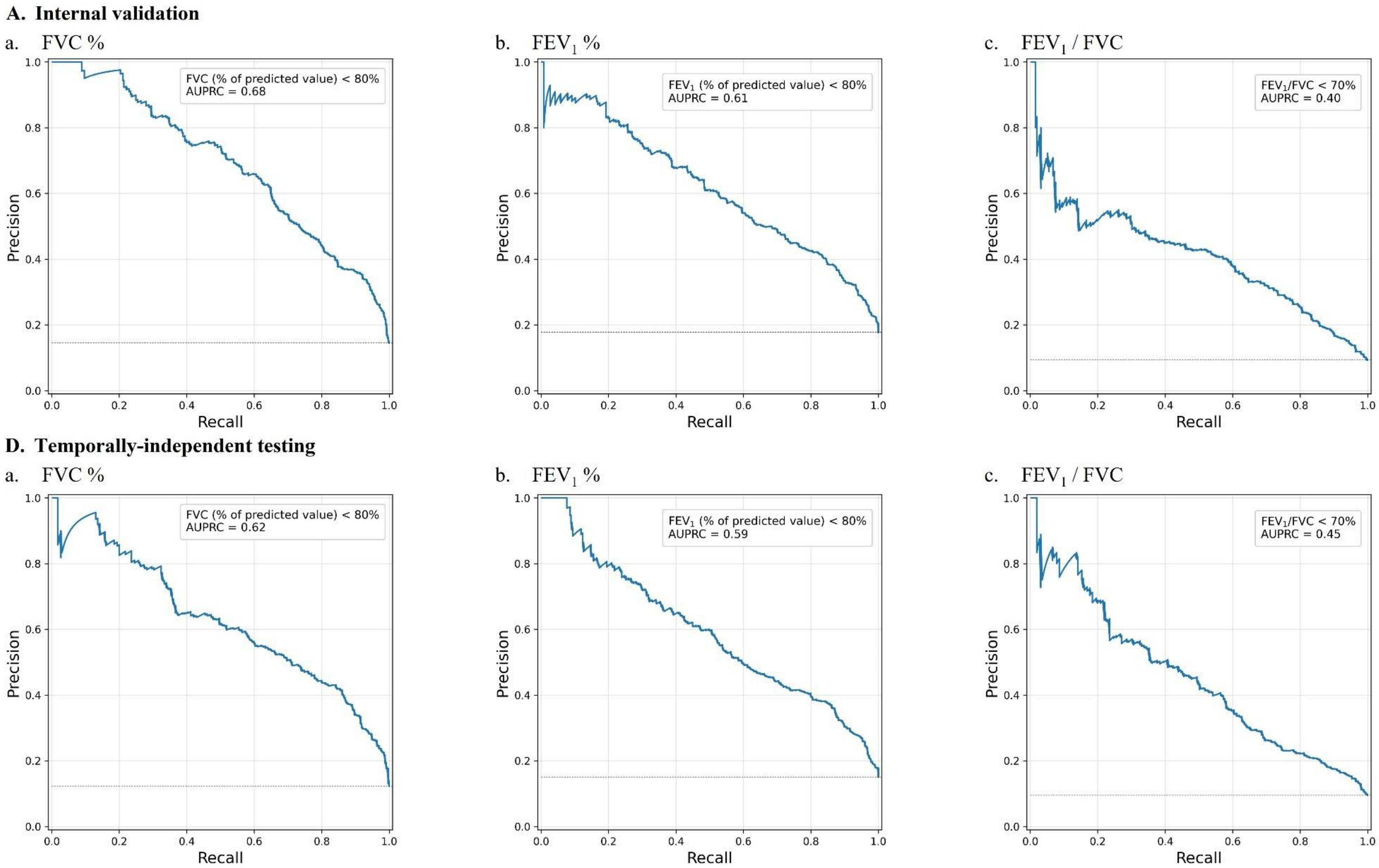


**Figure 11 Precision-recall curves for risk prediction on FVC%, FEV$_1$%, and FEV$_1$/FVC for internal validation (A) and temporally-independent testing (B).**

### 4) Interpretation of the deep learning model using saliency mapping

Figure 12 depicts typical images of saliency maps generated by the deep learning algorithm for predicting FVC and $FEV_1$. To generalize the highlighted region from GradCAM, the average intensities of 200 GradCAM samples were superimposed on a randomly chosen background chest CT background (in short, averageCAM). The averageCAM for $FVC_{DL}$ and $FEV_{1DL}$ demonstrates distinct regions in both lungs. The averageCAM for $FVC_{DL}$ emphasizes the anterior right lung region along the anterior chest wall. The central area of the left lung is also weakly highlighted. In contrast, the averageCAM from the $FEV_{1DL}$ model emphasizes the center regions of both lungs, particularly the right lung. Additionally, it illustrates the anterior and posterior regions of both lower lungs.

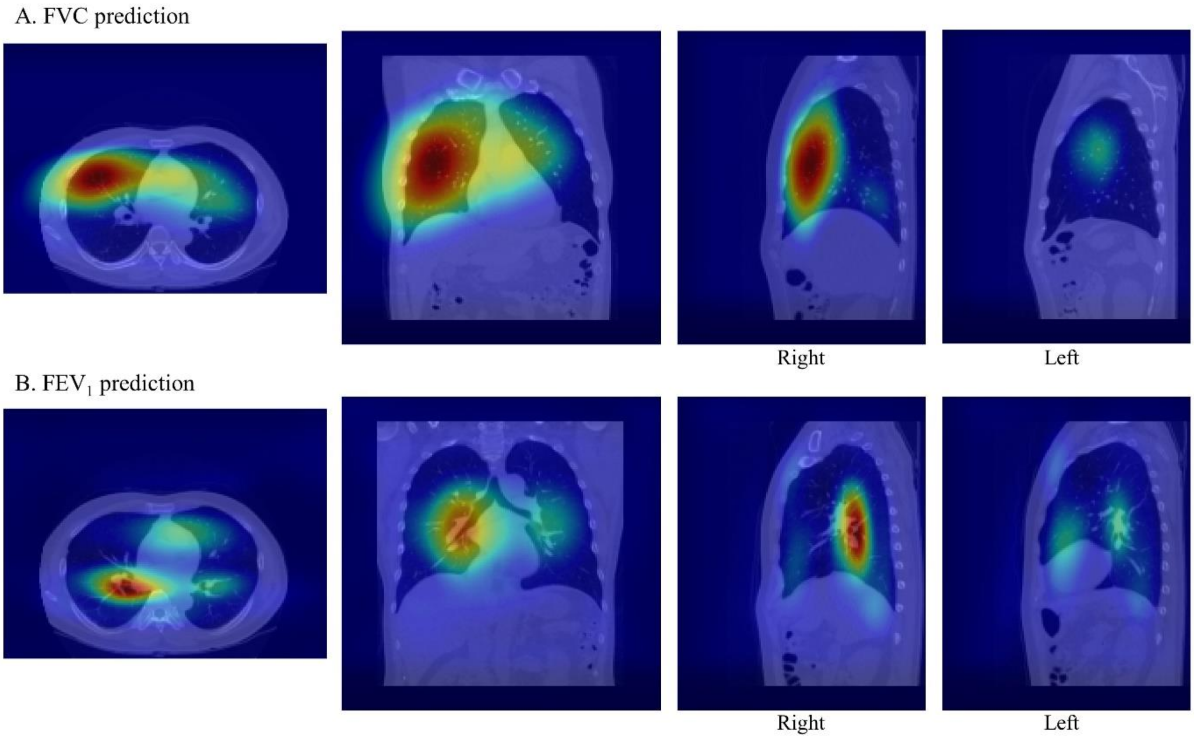


**Figure 12 AverageCAM analysis of the model for FVC prediction (A) and $FEV_1$ prediction (B)**

## Discussion

In our study, two models which were separately trained to predict the measured value of FVC and $FEV_1$ with 3D features directly extracted from low-dose CT scan by using convolutional neural networks achieved nice agreement performance with the high concordance correlation coefficient of 0.940 and 0.907 for FVC and $FEV_1$, respectively. While the clinically used parameters, the percent of predicted FVC (FVC %), the percent of predicted $FEV_1$ ($FEV_1$ %), and $FEV_1$/FVC, which are calculated from the deep learning predicted values, showed worse performance than the measured values, the AUROC, a representative measure for classification performance, still showed 0.91, 0.87, and 0.84 for FVC%, $FEV_1$%, $FEV_1$/FVC, respectively.

In our study, the agreement between ground truth and predicted values was better for the measured values than for the predicted values. The classification performance from which we can get a sense of simulating screening test didn't show reliable results yet with low sensitivity and positive predictive value.  In clinical practice, percentages of predicted values are used to identify disease severity and risk categories, although they have limits due to the reference equations. Age, height, and sex are employed to create the reference equations, although these variables have complex interactions when attempting to equate the expected lung function in healthy people (32, 33). Therefore, normal lung function demonstrates a broad range of variability for the majority of pulmonary function measures, and this variability varies with age (34). This constraint in predicting normal reference values may reduce the predictive power of $FEV_1\%_{DL}$ and FVC $\%_{DL}$, despite our use of a reference equation developed from a large local population (30).

Several CT parameters measuring small airway disease (20, 35), emphysema (35) were identified as being related with pulmonary function deterioration and we may expect those features to be incorporated into the automated deep learning application. However, the most highlighted region in the averageCAM predicting $FEV_1$ were quite apart from the peripheral area. This may be due in part to the down-sampling methods we adapted. In this exploration study, we attempted to use the entire chest CT for training to identify hints of regions related with predicting pulmonary function in CT scans, and not just the lung region. Due to the constraint of GPU capacity, we resampled the CT scan, which initially had pixel spacing of 0.69 (0.05) x 0.69 (0.05) x 2.51 (0.07) $mm^3$ in the development dataset, into identical 2.5 mm iso-cubic voxels. As each voxel in the CT scan becomes larger because of our preprocessing procedure, the contrast between the small cell sections may be diminished, resulting in the loss of their physio-morphological information, hence diminishing the importance of the region in calculating the output value.

Several limitations exist in our investigation.

Our model was trained and validated using data from a single center, therefore its performance may vary when applied to data from another center. To address this issue with generalizability, we have divided the data along the time axis into the development dataset and a temporally independent testing dataset. The FVC and $FEV_1$ exhibited remarkably good concordance. However, this model must be interpreted and applied with caution to the data received from other institutions.

Due to the radiation exposure risk associated with CT scans, low-dose chest CT screening for lung cancer is recommended for individuals with risk factors. Our dataset is derived from a health check-up center, where the visitors are expected to have relatively healthy, normal condition to the expected LDCT scan group who may be older male who have a longer pack-year history. Our dataset is derived from a health check-up center, where the visitors are expected to have relatively healthy, normal condition to the expected LDCT scan group who may be older male who have a longer pack-year history. In a previous study, the relationship between CT parameters and spirometry results are shown to be not linear but changed depending on COPD severity (25). The imaging features associated with predicting pulmonary function generated from the convolution might vary in other datasets with differing characteristics. Therefore, further studies involving different characteristics are needed to gain a solid understanding of this issue.

In this study, we employed the I3D model, a convolution network first designed for action recognition in video frames, with a sequence of coronal CT scan slices as input. This was chosen because developing this network had an advantage over the 3D convolution network given the limited GPU resources. While we explored some preprocessing methods that might influence the experimental results and chose some settings to implement, our experiments were conducted using only a single network architecture, without a thorough examination of the various network designs that would be optimal for completing our task. Given that the stride and kernel size were designed for the unique objective of the natural video dataset (31), it is anticipated that there exists a combination of network parameters suitable for resolving our issue. Later, the search for networks that match the properties of images and structures should be conducted. We believe this is work is anticipated to serve as a benchmark for similar future research.

# 3. Conclusion

In this retrospective study, we investigated the potential of low-dose CT as a screening tool to find risky groups of respiratory diseases with convolution neural networks (CNN) while finding the optimal parameters for input using CT and PFT values obtained from a single health screening center. With the CNN-derived 3D features trained and tuned with 9398 and 1343 subjects, respectively, our model achieved a concordance correlation coefficient of 0.940 and 0.907 for FVC and $FEV_1$ prediction, respectively, and an AUROC of 0.85 for classifying the risky group ($FEV_1$/FVC ratio < 70%) using an independent testing dataset of 2720 subjects. Our study has the meaning in that we utilized the automatically derived features from whole volume low-dose chest CT to predict the measured values of spirometry, obtaining excellent agreement. Also, we searched the proper setting for input at the same time.

In our study, we compared input parameters with only one network architecture, which has the limitation of not being able to explore the network design suitable for our purpose. The dataset was gathered from a single site with relatively normal participants, preventing us from obtaining accurate characteristics of diseased groups. If a subsequent study with a dataset containing a diseased group is conducted with network architecture better suited for this purpose, the clinical utility and importance of this study will increase. In any case, the significance of our findings was that they revealed the baseline score for predicting spirometry-measured parameters from a normal population using a deep learning approach and evaluated the optimal task setup.

# Bibliography

1. Wang L, Wee CY, Suk HI, Tang X, Shen D. MRI-based intelligence quotient (IQ) estimation with sparse learning. *PLoS One* 2015; 10: e0117295.
2. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011; 365: 395-409.
3. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, Lammers JJ, Weenink C, Yousaf-Khan U, Horeweg N, van 't Westeinde S, Prokop M, Mali WP, Mohamed Hoesein FAA, van Ooijen PMA, Aerts J, den Bakker MA, Thunnissen E, Verschakelen J, Vliegenthart R, Walter JE, Ten Haaf K, Groen HJM, Oudkerk M. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med* 2020; 382: 503-513.
4. Force USPST, Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, Caughey AB, Davis EM, Donahue KE, Doubeni CA, Kubik M, Landefeld CS, Li L, Ogedegbe G, Owens DK, Pbert L, Silverstein M, Stevermer J, Tseng CW, Wong JB. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* 2021; 325: 962-970.

5. Meza R, Jeon J, Toumazis I, Ten Haaf K, Cao P, Bastani M, Han SS, Blom EF, Jonas DE, Feuer EJ, Plevritis SK, de Koning HJ, Kong CY. Evaluation of the Benefits and Harms of Lung Cancer Screening With Low-Dose Computed Tomography: Modeling Study for the US Preventive Services Task Force. *JAMA* 2021; 325: 988-997.

6. Nawa T, Fukui K, Nakayama T, Sagawa M, Nakagawa T, Ichimura H, Mizoue T. A population-based cohort study to evaluate the effectiveness of lung cancer screening using low-dose CT in Hitachi city, Japan. *Jpn J Clin Oncol* 2019; 49: 130-136.

7. Qaseem A, Snow V, Shekelle P, Sherif K, Wilt TJ, Weinberger S, Owens DK, Clinical Efficacy Assessment Subcommittee of the American College of P. Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline from the American College of Physicians. *Ann Intern Med* 2007; 147: 633-638.

8. Force USPST, Siu AL, Bibbins-Domingo K, Grossman DC, Davidson KW, Epling JW, Jr., Garcia FA, Gillman M, Kemper AR, Krist AH, Kurth AE, Landefeld CS, Mangione CM, Harper DM, Phillips WR, Phipps MG, Pignone MP. Screening for Chronic Obstructive Pulmonary Disease: US Preventive Services Task Force Recommendation Statement. *JAMA* 2016; 315: 1372-1377.

9. Lopez Varela MV, Montes de Oca M, Rey A, Casas A, Stirbulov R, Di Boscio V, Team P. Development of a simple screening tool for opportunistic COPD case finding in primary care in Latin America: The PUMA study. *Respirology* 2016; 21: 1227-1234.

10. Hill K, Goldstein RS, Guyatt GH, Blouin M, Tan WC, Davis LL, Heels-Ansdell DM, Erak M, Bragaglia PJ, Tamari IE, Hodder R, Stanbrook MB. Prevalence and underdiagnosis of chronic obstructive pulmonary disease among patients at risk in primary care. *CMAJ* 2010; 182: 673-678.

11. GOLD : Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2020. . 2021.

12. Graham BL, Steenbruggen I, Miller MR, Barjaktarevic IZ, Cooper BG, Hall GL, Hallstrand TS, Kaminsky DA, McCarthy K, McCormack MC, Oropez CE, Rosenfeld M, Stanojevic S, Swanney MP, Thompson BR. Standardization of Spirometry 2019 Update. An Official American Thoracic Society and European Respiratory Society Technical Statement. *Am J Respir Crit Care Med* 2019; 200: e70-e88.

13. Tan WC, Sin DD, Bourbeau J, Hernandez P, Chapman KR, Cowie R, FitzGerald JM, Marciniuk DD, Maltais F, Buist AS, Road J, Hogg JC, Kirby M, Coxson H, Hague C, Leipsic J, O'Donnell DE, Aaron SD, Can CCRG. Characteristics of COPD in never-smokers and ever-smokers in the general population: results from the CanCOLD study. *Thorax* 2015; 70: 822-829.

14. Han MK, Steenrod AW, Bacci ED, Leidy NK, Mannino DM, Thomashow BM, Barr RG, Make BJ, Bowler RP, Rennard SI, Houfek JF, Yawn BP, Meldrum CA, Walsh JW, Martinez FJ. Identifying Patients with Undiagnosed COPD in Primary Care Settings: Insight from Screening Tools and Epidemiologic Studies. *Chronic Obstr Pulm Dis* 2015; 2: 103-121.

15. Regan EA, Lynch DA, Curran-Everett D, Curtis JL, Austin JH, Grenier PA, Kauczor HU, Bailey WC, DeMeo DL, Casaburi RH, Friedman P, Van Beek EJ, Hokanson JE, Bowler RP, Beaty TH, Washko GR, Han MK, Kim V, Kim SS, Yagihashi K, Washington L, McEvoy CE, Tanner C, Mannino DM, Make BJ, Silverman EK, Crapo JD, Genetic Epidemiology of CI. Clinical and Radiologic Disease in Smokers With Normal Spirometry. *Jama Intern Med* 2015; 175: 1539-1549.

16. Chassagnon G, Vakalopolou M, Paragios N, Revel MP. Deep learning: definition and perspectives for thoracic imaging. *Eur Radiol* 2020; 30: 2021-2030.

17. Das N, Verstraete K, Stanojevic S, Topalovic M, Aerts JM, Janssens W. Deep-learning algorithm helps to standardise ATS/ERS spirometric acceptability and usability criteria. *Eur Respir J* 2020; 56.

18. Bodduluri S, Nakhmani A, Reinhardt JM, Wilson CG, McDonald ML, Rudraraju R, Jaeger BC, Bhakta NR, Castaldi PJ, Sciurba FC, Zhang C, Bangalore PV, Bhatt SP. Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease. *JCI Insight* 2020; 5.

19. Iwano S, Okada T, Satake H, Naganawa S. 3D-CT Volumetry of the Lung Using Multidetector Row CT: Comparison with Pulmonary Function Tests. *Academic Radiology* 2009; 16: 250-256.

20. Bhatt SP, Soler X, Wang X, Murray S, Anzueto AR, Beaty TH, Boriek AM, Casaburi R, Criner GJ, Diaz AA, Dransfield MT, Curran-Everett D, Galban CJ, Hoffman EA, Hogg JC, Kazerooni EA, Kim V, Kinney GL, Lagstein A, Lynch DA, Make BJ, Martinez FJ, Ramsdell JW, Reddy R, Ross BD, Rossiter HB, Steiner RM, Strand MJ, van Beek EJR, Wan ES, Washko GR, Wells JM, Wendt CH, Wise RA, Silverman EK, Crapo JD, Bowler RP, Han MLK, Investigators C. Association between Functional Small Airway Disease and FEV1 Decline in Chronic Obstructive Pulmonary Disease. *Am J Resp Crit Care* 2016; 194: 178-184.

21. Kwack WG, Kang YS, Jeong YJ, Oh JY, Cha YK, Kim JS, Yoon YS. Association between thoracic fat measured using computed tomography and lung function in a population without respiratory diseases. *Journal of Thoracic Disease* 2019; 11: 5300-5309.

22. Chen F, Kubo T, Shoji T, Fujinaga T, Bando T, Date H. Comparison of pulmonary function test and computed tomography volumetry in living lung donors. *J Heart Lung Transplant* 2011; 30: 572-575.

23. Lafata KJ, Zhou Z, Liu JG, Hong J, Kelsey CR, Yin FF. An Exploratory Radiomics Approach to Quantifying Pulmonary Function in CT Images. *Sci Rep* 2019; 9: 11509.

24. Chen K, Hoffman EA, Seetharaman I, Jiao F, Lin CL, Chan KS. Linking Lung Airway Structure to Pulmonary Function Via Composite Bridge Regression. *Ann Appl Stat* 2016; 10: 1880-1906.

25. Koo HJ, Lee SM, Seo JB, Lee SM, Kim N, Oh SY, Lee JS, Oh YM. Prediction of Pulmonary Function in Patients with Chronic Obstructive Pulmonary Disease: Correlation with Quantitative CT Parameters. *Korean J Radiol* 2019; 20: 683-692.

26. McDonald ML, Diaz AA, Ross JC, San Jose Estepar R, Zhou L, Regan EA, Eckbo E, Muralidhar N, Come CE, Cho MH, Hersh CP, Lange C, Wouters E, Casaburi RH, Coxson HO, Macnee W, Rennard SI, Lomas DA, Agusti A, Celli BR, Black-Shinn JL, Kinney GL, Lutz SM, Hokanson JE, Silverman EK, Washko GR. Quantitative computed tomography measures of pectoralis muscle area and disease severity in chronic obstructive pulmonary disease. A cross-sectional study. *Ann Am Thorac Soc* 2014; 11: 326-334.

27. Pu J, Leader JK, Meng X, Whiting B, Wilson D, Sciurba FC, Reilly JJ, Bigbee WL, Siegfried J, Gur D. Three-dimensional airway tree architecture and pulmonary function. *Acad Radiol* 2012; 19: 1395-1401.

28. Szegedy C, Liu W, Jia YQ, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going Deeper with Convolutions. *2015 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)* 2015: 1-9.

29. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. *2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)* 2016: 2818-2826.

30. Choi JK PD, Lee JO, . Normal predictive values of spirometry in Korean population. Tuberculosis and Respiratory Diseases. *Tuber Resp Dis* 2005; 58: 230-242.

31. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 6299-6308.

32. Quanjer PH, Hall GL, Stanojevic S, Cole TJ, Stocks J. Age-and height-based prediction bias in spirometry reference equations. *European Respiratory Journal* 2012; 40: 190-197.

33. Haynes JM, Kaminsky DA, Stanojevic S, Ruppel GL. Pulmonary function reference equations: a brief history to explain all the confusion. *Respiratory care* 2020; 65: 1030-1038.

34. Miller MR, Quanjer PH, Swanney MP, Ruppel G, Enright PL. Interpreting lung function data using 80% predicted and fixed thresholds misclassifies more than 20% of patients. *Chest* 2011; 139: 52-59.

35. Ostridge K, Williams NP, Kim V, Harden S, Bourne S, Clarke SC, Aris E, Mesia-Vela S, Devaster JM, Tuck A, Williams A, Wootton S, Staples KJ, Wilkinson TMA, Group AS. Relationship of CT-quantified emphysema, small airways disease and bronchial wall dimensions with physiological, inflammatory and infective measures in COPD. *Respir Res* 2018; 19: 31.

# Abstract (In Korean)

폐암 선별 진단에 있어서 기존의 검진 방법인 흉부 X 선에 비해 저선량 방사선 CT 가 그 선별능력이 우월함을 입증하다는 내용을 담은 두 번의 대규모 무작위 대조 실험(RCT) 결과가 보고된 이후, 여러 폐암 관련 단체들은 고위험 군에 저선량 CT 를 이용한 폐암검진 권고안을 마련하였다. 국내에서도 2 년 간의 시범사업을 거쳐 2019 년부터 국가 암 건진 사업의 대상에 폐암을 포함하였다. 이로써 54 세-74 세 연령 군 중, 30 갑년 이상의 흡연력을 가진 이는 국가의 지원을 받아 저선량 CT 를 이용한 폐암 검진을 받을 수 있게 되었다. 많은 나라들이 이에 대응하는 정책적 노력을 기울이고 있어 향후 저선량 CT 의 시행 횟수가 증가할 것으로 예상되며 이러한 영상 정보를 이용하여 자동화된 방식으로 진단에 도움을 줄 수 있는 알고리즘의 필요성이 증대될 것으로 생각된다.

폐활량계를 이용한 폐기능검사(spirometry)는 COPD 및 여러 호흡기 질환을 진단하고, 중증도를 판정하는 도구로서 활용되고 있다. CT 에서 여러 정량적 지표를 추출하여 폐활량 지표와의 상관관계를 분석하고, 해당 지표를 예측하는 연구 결과가 많이 제시되었다. 그러나 기존 연구는 정량적 지표를 추출하거나 추출된 지표를 이용하여 모델을 수립하는 과정에 인간의 노력과 시간이 많이 드는 한계가 존재하였다. 합성곱 기반 신경망(Convolutional neural network)을 이용한 심층 학습 알고리즘(deep learning; 이하 딥러닝)은 학습과정에서 역전파 알고리즘을 이용하여, 영상의 특성 지표를 추출하고 모델을 구성하는 과정을 자동화하며 기존 알고리즘에 비해 월등한 성능을 보여 주고 있다. 본 학위논문에서는 이러한 CNN 기반의 딥러닝 알고리즘을 이용하여 저선량 CT 에서 폐활량 지표 예측 가능성을 탐색하였다.

서울아산병원에서 건강검진을 받은 피검사자를 대상으로 한 본 연구에서는 폐활량계로 측정한 노력성 폐활량(forced vital capacity; 이하, FVC)와 1 초간 노력성 폐활량(forced expiratory volume in one second; 이하, $FEV_1$)의 측정값을 예측하는 모델을 각각 구성하였다. 이렇게 예측한 FVC 와 $FEV_1$ 은 해당 대상자와 같은 성별, 연령, 키, 몸무게를 지닌 정상군에 대한 예측값에 대한 비율, FVC 에 대한 $FEV_1$ 의 비율로 변환되어 임상에서 위험군을 진단할 때 쓰이는 기준을 적용한 분류 모델을 구축하는 데에 이용되었다. 위험군을 선별하는 기준은 FVC%, $FEV_1$% <80 그리고 $FEV_1$/FVC <70%이다. 또한, 측정값과 딥러닝 모델 예측값 사이의 일치도 평가지표도 계산하여 비교하였다. 학습된 딥러닝 모델은 폐활량 측정값 모두에 좋은 일치도 성능을 보여 주었으며, FVC 가 $FEV_1$ 에 비해 좋은 성능을 보였으나, 기준값에 대해 보정된 지표들(FVC%, $FEV_1$%)에 대해서는 측정치보다 낮은 성능을 보였다. 또한, 분류 성능에 대해서는 임상에 쓰일 만한 좋은 성능을 보여 주지는 못했다. FVC, FEV1 각 예측 모델의 결과에 영향을

주는 영역을 GradCAM 을 이용해 살펴본 바, FVC 예측 모델은 우측 폐의 앞부분과, 좌측 폐의 중간 부분을 강조하였으며, $FEV_1$ 예측 모델에서는 양측 모델의 중간 부분과, 앞/뒤 영역이 강조된 결과가 도출되었다.

결론적으로, 본 학위 논문의 연구과정에서는 CNN 을 기반으로 한 딥러닝 모델을 이용하여 저선량 CT 영상에서 폐 기능, 그 중 폐활량 지표를 예측하는 모델을 개발하였다. 모델을 개발하는 과정에서, 연구 결과에 영향을 주는 입력 변수들의 조합에 대해서도 실험하였다. 본 연구의 결과가 CT 를 이용한 폐 기능 예측 연구를 진행함에 있어 시작점이 될 수 있을 것으로 기대된다.