End-to-end domain knowledge-assisted automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using computed tomography (CT)

Wenxi Yu and Hua Zhou
Department of Biostatistics, University of California, Los Angeles, CA 90024, USA

Jonathan G. Goldin
Department of Radiology, University of California, Los Angeles, CA 90024, USA

Weng Kee Wong
Department of Biostatistics, University of California, Los Angeles, CA 90024, USA

Grace Hyun J. Kim[a)]
Department of Biostatistics, University of California, Los Angeles, CA 90024, USA
Department of Radiology, University of California, Los Angeles, CA 90024, USA

(Received 9 March 2020; revised 21 January 2021; accepted for publication 25 January 2021; published 19 March 2021)

Purpose: Domain knowledge (DK) acquired from prior studies is important for medical diagnosis. This paper leverages the population-level DK using an optimality design criterion to train a deep learning model in an end-to-end manner. In this study, the problem of interest is at the patient level to diagnose a subject with idiopathic pulmonary fibrosis (IPF) among subjects with interstitial lung disease (ILD) using a computed tomography (CT). IPF diagnosis is a complicated process with multidisciplinary discussion with experts and is subject to interobserver variability, even for experienced radiologists. To this end, we propose a new statistical method to construct a time/memory-efficient IPF diagnosis model using axial chest CT and DK, along with an optimality design criterion via a DK-enhanced loss function of deep learning.
Methods: Four state-of-the-art two-dimensional convolutional neural network (2D-CNN) architectures (MobileNet, VGG16, ResNet-50, and DenseNet-121) and one baseline 2D-CNN are implemented to automatically diagnose IPF among ILD patients. Axial lung CT images are retrospectively acquired from 389 IPF patients and 700 non-IPF ILD patients in five multicenter clinical trials. To enrich the sample size and boost model performance, we sample 20 three-slice samples (triplets) from each CT scan, where these three slices are randomly selected from the top, middle, and bottom of both lungs respectively. Model performance is evaluated using a fivefold cross-validation, where each fold was stratified using a fixed proportion of IPF vs non-IPF.
Results: Using DK-enhanced loss function increases the model performance of the baseline CNN model from 0.77 to 0.89 in terms of study-wise accuracy. Four other well-developed models reach satisfactory model performance with an overall accuracy >0.95 but the benefits brought on by the DK-enhanced loss function is not noticeable.
Conclusions: We believe this is the first attempt that (a) uses population-level DK with an optimal design criterion to train deep learning-based diagnostic models in an end-to-end manner and (b) focuses on patient-level IPF diagnosis. Further evaluation of using population-level DK on prospective studies is warranted and is underway. [https://doi.org/10.1002/mp.14754]

Key words: computed tomography, deep learning, idiopathic pulmonary fibrosis (IPF), optimal design

1. INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is defined as a specific form of chronic, progressive fibrosing interstitial pneumonia of unknown causes. IPF is limited to the lungs and usually occurs in older adults.[1] It is a rare disease with irreversible and unpredictable progression and survival.[1] The prevalence estimates of IPF in the USA varied between 14 and 27.9 cases per 100,000 in the population.[2] The median survival time ranges from 2 to 5 yr, but some patients live much longer.[1–3]

Idiopathic pulmonary fibrosis is associated with histopathologic and/or radiologic pattern of usual interstitial pneumonia (UIP).[1] Computed tomography (CT) chest images are used to determine the presence of the UIP pattern. UIP pattern is associated with some common CT representations, including honeycombing, ground glass opacity, reticular pattern with peripheral traction bronchiectasis or bronchiolectasis, etc.[3] Notably, these CT features usually occur in the subpleural and basal areas.

The diagnosis of IPF involves the collaboration of multidisciplinary discussion from specialists: clinicians,

radiologists, and pathologists. The up-to-date clinical practice guideline for IPF, published in 2018, provides a detailed explanation and flowchart regarding the overall diagnostic workflow.[3] According to the guideline, CT assessment has become a cornerstone in the diagnosis of IPF. However, using CT evaluation for IPF diagnosis is a difficult task and subject to interobserver variability, even for experienced radiologists.[4] Developing an automated diagnosis of IPF using CT can be helpful for a prototype of this task or a prescreening tool.

Additionally, in some cases where a definite diagnosis of IPF could not be made, surgical lung biopsy is suggested.[3] However, surgical lung biopsy is also known to be associated with an increasing risk of in-hospitalization or mortality.[5] In this context, investigating automated CT evaluation for IPF diagnosis may potentially reduce the need for lung biopsy in the long run.

Our aim is to develop an efficient and domain knowledge-assisted diagnosis model for IPF among ILD patients based on their axial lung CT scans. It is a time/memory efficient method and no lung segmentation is required. Domain knowledge (DK) based on previous studies and optimal design theory is incorporated in the training of diagnostic models in an end-to-end manner. An added advantage of our method is that it leverages the population level IPF prognostic trends (i.e., whether CT images indicate disease progression or not) across the lung positions, which is an important factor in the classification of IPF.

There are three potential clinical significance of this work: (a) it facilitates automatic diagnosis of IPF that saves time and reduces interobserver variability; (b) it enables early diagnosis and treatment, which may lead to early antifibrotic treatment and increase the likelihood of a slow disease progression; and (c) it potentially reduces the need for lung biopsy in the diagnosis process. The latter is an important consideration since biopsy is associated with increased in-hospital mortality.

There have been growing interests in providing IPF prognosis support after two proven effective therapeutic treatments.[6,7] Specifically, developing robust and sensitive biomarkers is meaningful for evaluating the efficacy of IPF clinical trials. Previous research used machine learning techniques (such as support vector machines) to construct quantitative CT scores from texture classification model and they have shown good clinical applicability.[8–10]

Our work, different from IPF prognosis, focuses on the diagnosis of IPF. Computer-aided diagnosis system has gained popularity over the past few years. Some attempts have been made to use deep learning methods for interstitial lung diseases classification problems on multiple input image scales. The input scales vary from image patches of size 32×32,[11] one axial slice,[12] and frontal-view chest CT image.[13]

Patient-level UIP diagnosis classifies patients into three categories: UIP, possible UIP, or inconsistent with UIP. Recent methods using deep learning tools have shown comparable performance when patients are diagnosed by radiologists.[14] Our work differs from this study conducted by Walsh et al.[14] in three ways: (a) our current work focuses on IPF rather than UIP diagnosis; (b) no lung segmentation is needed in our work; and (c) transfer learning and DK are incorporated in the present work, which also uses statistical optimization techniques. Using CT scans to automatically diagnose IPF is limited so far and we believe our proposed method can have a potential impact in patient-level classification of IPF with DK using volumetric CT scans.

2. MATERIALS AND METHODS

2.A. Datasets

Axial lung CT scans are retrospectively acquired from five multicenter studies, including two IPF studies and three non-IPF studies. The inclusion criterion is that each patient has been clinically diagnosed as interstitial lung diseases. CT scans with IPF diagnosis were confirmed by multidisciplinary clinical teams.[1,3] CT images of IPF patients were collected from December 2004 to July 2016; CT images of non-IPF patients were collected from May 1997 to May 2018. For each patient, only the first available total lung capacity (TLC) scans are used for the algorithm development and testing. In total, there are 1089 patients, including 389 IPF and 700 non-IPF patients, collectively obtained from the five multicenter studies. CT images were acquired under different CT scanners and protocols, which are summarized in the Supporting Information A. Figure 1 shows the data flow of image preprocessing and model construction and Table I summarizes the disease diagnosis, the number of patient visits, and the number of CT slices per visit for the five studies with study 1 and 2 involving IPF patients, and study 3, 4, and 5 involving non-IPF ILD patients. CT scans from study 1 and 2 were confirmed as IPF with the IPF diagnostic criteria.[1,3] CT scans from study 3, 4, and 5 were clinically confirmed as other ILD diseases. CT scans from study 1, 4 and 5 were anonymized images from multicenter studies, whereas CT scans from study 2 and 3 were each collected from a single center. We note that some scans (13.3%, N = 60) from study 3 are non-volumetric scans, where the spacing between each adjacent CT slice along the z-dimension is not consistent. Therefore, the average number of CT slices in study 3 is fewer than that of other studies.

2.B. Problem statement

Our main research problem is a binary classification task to determine whether a CT scan is from a subject with IPF vs non-IPF. The model input is the axial lung CT images of one patient visit, which are usually of dimension $512 \times 512 \times N_s$. Here 512 is the image resolution and N_s is the number of slices, which varies from different CT scans. The output is a binary label $y \in \{0, 1\}$ indicating whether the CT scan is from a subject with IPF or not. Further clinical information, such as gender and age, cannot be retrieved due to the

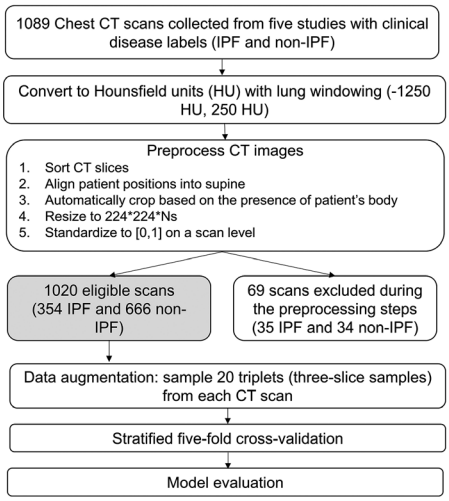

FIG. 1. Data flow of image preparation and model construction. Ns: the number of computed tomography slices for each scan, which varies for each scan.

TABLE I. Basic information of the five studies.

Study	Type	Disease diagnosis	Number of subjects	Number of CT slices per visit (mean ± standard deviation)
1	IPF	IPF	245	359 ± 106
2	IPF	IPF	144	280 ± 46
3	Non-IPF	Other ILDs	449	53 ± 25
4	Non-IPF	Myositis	81	253 ± 75
5	Non-IPF	Systemic sclerosis	170	106 ± 83

anonymization process, and thus is not provided for the automatic diagnosis system.

In clinical settings, the classification task needs to be carried out in a timely manner with limited training samples and computational storage. Due to the weak supervision nature of this task (i.e., one ground truth label per CT scan) and the relatively limited number of images available, we propose to use two-dimensional convolutional neural network (2D-CNN) models, rather than 3D-CNN, for this work. 2D-CNN models are commonly used for other medical-related tasks.[15–17]

Dimensionality reduction is necessary before implementing the 2D-CNN models. These models constrain the third dimension of the input to be three, corresponding to the RGB channels. We propose to reduce the input dimension to $224 \times 224 \times 3$ by the incorporation of DK and optimal design theory. Thus, for each training and testing sample, only three lung CT slices are used as model inputs. We refer the three CT slices as a *triplet* throughout the rest of the manuscript.

For illustration, Fig. 2 shows four representative triplets in terms of their original and rescaled images, with different clinical diagnoses. After preprocessing, we automatically remove the information that is outside of the body. Each CT slice is rescaled to a uniform dimension of 224×224, which is the commonly used as the default size of CNN architectures, to normalize patients with different sizes along the anteroposterior and lateral dimensions. Additionally, for prone CT scans, we rotate the scans 180 degrees to align scans with different patient positions. More details of the preprocessing steps are described in Section 2.E.

It is well-known that deep learning models usually require a large amount of training data; accordingly, for each scan, we randomly sample a user-selected number M of triplets to enrich the number of training and testing samples. In our study, we select $M = 20$. At the same time, we include some additional experiments by setting an adaptive number for M based on the number of CT slices for each scan. More details are provided in the section 2.G.

2.C. Domain knowledge (DK)

We leverage DK in the selection of triplet locations using a statistical optimality design criterion and the training of the classification model in an end-to-end manner.

Specifically, we utilize the population-level disease trends of IPF in our classification task. Previous studies used quantum particle swarm optimization incorporated with a resampling technique and a random forest method to predict the pixel-level IPF progression status (i.e., whether the pixel of the segmented CT lung image suggests progressive or not progressive).[18] Intuitively, CT slices that contain more progressive pixels have more disease patterns of IPF and thus could be useful information in the classification task. Therefore, we assign higher weights for triplets which have well-represented IPF progressive trends, and vice versa. The weights for each triplet are then evaluated using an optimal design criterion.

Before discussing technical details, we first define standardized slice position (SSP) to align patient visits with a varying number of CT slices. We define $SSP = \frac{n^{th}CT\,slice\,number}{N_s - 1}$, where N_s is the number of slices for that patient visit. SSP ranges from 0 to 1, where 0 is the first CT slice at the very top of the lung and 1 is the last CT slice at the very bottom of the lung.

Based on the predictive results,[18] we plot the percentage of progressive lung area vs SSP based on the population level, see Fig. 3(a). The blue line represents the median curve on a population level and the gray area represents the 95% quantiles.

We observe that except for the boundaries (i.e., the apex and base of the lungs), which are defined by the top and bottom 10%, the percentage of progressive lung areas gradually increases as the slice moves toward the base of the lungs. This is consistent with previous findings for UIP patterns, which are indicative of IPF and usually reside in the base of lung parenchyma. We note that, at the boundaries (the first and last few CT slices), the number of segmented lung area voxels are much smaller than that of other areas. Also, there is a high level of noise effect due to the proton refection near scapula. Based on these two reasons, the prediction results at the boundaries are unstable with a wide quantile for the

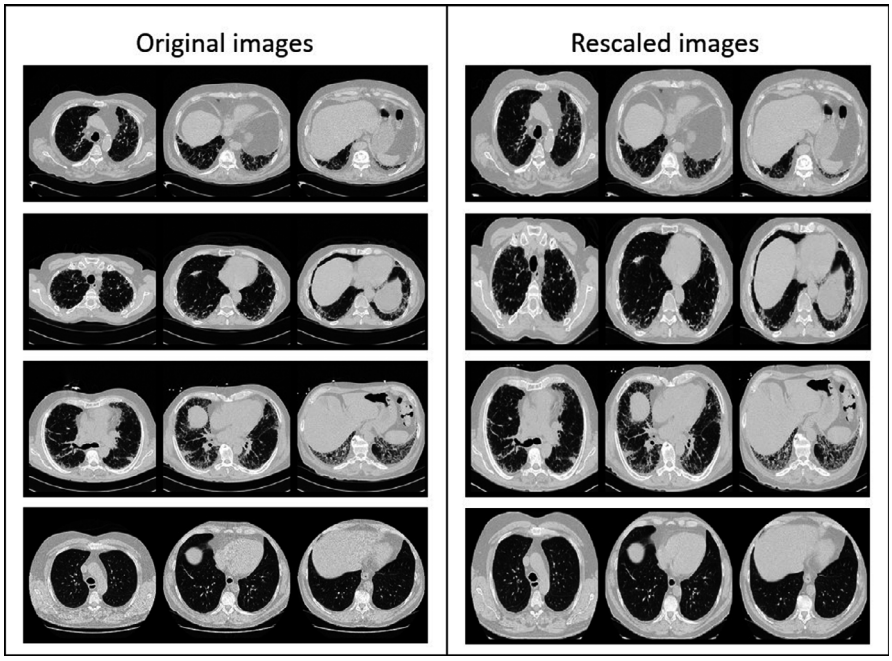

FIG. 2. Four representative triplets of original images and rescaled computed tomography images. The top row is one idiopathic pulmonary fibrosis (IPF) patient with radiological diagnosis of usual interstitial pneumonia (UIP) pattern; the second row is one IPF patient with possible UIP diagnosis; the third row is a non-IPF patient with possible UIP pattern; and the bottom row is one non-IPF patient.

percentage of progressive lung areas. We therefore remove the boundaries for future analysis.

Figure 3(a) shows four vertical orange dotted lines, which are the SSP locations at 0.1, 0.37, 0.64, and 0.9. They are obtained by removing the top and bottom 10% to avoid the boundary effects, and then evenly dividing the rest of the lung positions into three zones, indicated as zone 1, 2, and 3 in the figure. Specifically, zone 1, 2, and 3 represent SSP locations from 0.1 to 0.37, from 0.37 to 0.64, and from 0.64 to 0.9, respectively, and they capture the upper, middle, and lower of the lungs respectively.

For each triplet, we sample one slice from each zone. We test the model performance with and without DK-enhanced loss function in Fig. 3. Without DK, we treat each triplet identically and assign the same weights for all triplets. With DK, we assign greater weights to triplets that are more representative of the population level IPF progressive trends; see for example, triplet 2 shown in Fig. 3(c) for calculating the loss function. Thus, these triplets play an important role in estimating parameters in the IPF diagnostic model when the entire process is conducted in an end-to-end manner. We provide the detailed steps on how to calculate the D-criterion value of triplet 1, shown in the Fig. 3(b), in the Supporting Information C.

2.D. D-optimal design

Model-based optimal design theory has numerous and useful applications in medical research, engineering, and many other disciplines.[19,20] When we have a statistical model to describe the relationship between the mean response variable and covariates, optimal design theory provides guidance on how to judiciously design an experiment to optimize the criterion. One common criterion is that model parameters be estimated as accurately as possible with minimal cost. Such an objective is attained by a D-optimal and described in more details below. For our project, a D-optimal design helps us determine the weights to be used in each triplet to assess the overall trends of the population-level IPF progressive curve using information from prior studies [see Fig. 3(a)] via a DK-enhanced loss function shown as $D(Z_i)$ in the formula (d) in Fig. 3. Additional background information on optimal designs can be found in Berger and Wong.[20]

We now provide some fundamentals on constructing D-optimal designs. Suppose we have N independent responses from an assumed statistical model given by $y_i = f(x_i)^T\beta + \varepsilon_i, i = 1,\ldots,N$. Here y_i is the univariate response variable from subject i, $f(x_i)$ is a design vector of dimension $p \times 1$, β is the unknown parameter of dimension $p \times 1$, and the error term ε_i is normally distributed with mean 0 and constant variance. For example, we may have two covariates age and gender in our study and the regression function $f(x_i) = (1, \text{age}_i, \text{gender}_i)^T$ has $p = 3$ parameters. If the interest is to estimate the three parameters in the model, two common design criteria are D-optimality and A-optimality, and if interest is to estimate the entire response surface, G-optimality is frequently used.[19] Here D, A, and G stand for the determinant (Det), average variance, and global criterion, respectively and the resulting optimal designs have different properties. The D-optimality criterion is the most popular for estimating model parameters and mathematically, it is

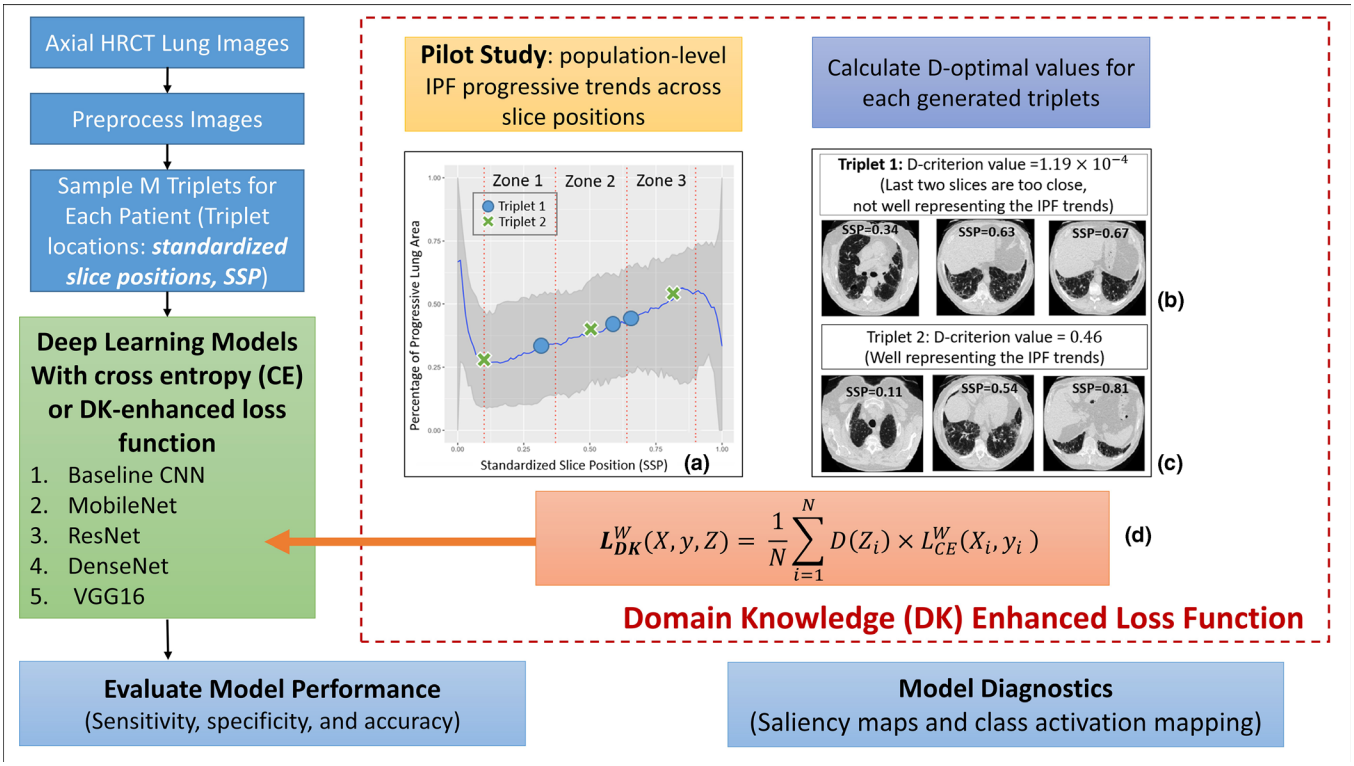

FIG. 3. Flowchart of the study design. SSP: standardized slice position, DK: domain knowledge with optimization, CE: cross entropy without optimization in selecting slices. [Color figure can be viewed at wileyonlinelibrary.com]

defined by $\mathrm{Det}[\mathrm{Cov}(\hat{\beta})]$. A design that achieves the smallest D-criterion value among all designs is *D-optimal* and such a design estimates the model parameters with the smallest volume of the confidence ellipsoid for β.

For nonlinear models, the criterion depends on the unknown parameters that we want to estimate and they have to be replaced by an initial set of estimates for the model parameters before the D-optimality criterion can be optimized. The resulting designs are, strictly speaking, locally D-optimal designs because they depend on the initial set of model parameters estimates.

Our response variable is the population trends of the percentage of progressive lung area over SSP and we estimated it using data acquired from the pilot study.[21] We used the generalized linear model (GLM) with a logit link function since the response variable, the percentage of progressive pixels, is not normally distributed.

We used data and fitted several what we thought are plausible models: they include polynomial models of degrees 3 and 4 and more flexible models like fractional polynomials. The latter class models the mean response as a polynomial but additionally allows for fractional powers in each nominal. Fractional polynomials were proposed by Royston et al.[22] where they showed via many examples that fractional polynomials can fit univariate response variables in the biomedical sciences much better than polynomials. They further recommended that for practical applications, it suffices to consider a set consisting of positive and non-negative powers only. For this reason, we also used fractional polynomials to estimate the median population level disease progression. Akaike information criterion (AIC) and visual examination were used as criteria for model selection.[23] Both criteria suggest that FP is the best model that describes the median population trends of IPF progression among all the models we have considered. Details on the model comparisons and estimated parameters are in the Supporting Information B.

In a nutshell, for each randomly sampled triplet, we evaluate its D-criterion value based on the determinant of the information matrix. Triplets with a larger D-criterion value better represent the overall population level IPF progressive trends. The Supporting Information C and D contain further discussion on the D-optimal design under a GLM setting and the distribution of D-criterion values.

2.E. Two-dimensional convolutional neural network (2D-CNN)

Before implementing 2D-CNN models, we normalized each CT scan if the scan did not meet the study-level criteria. Four main study-level criteria are: (a) align patient's position into supine, (b) center a patient position, (c) automatically remove the location of table information, and (d) rescale to a uniform image size. If a CT scan was deviated from the general platform, we normalized the images prior to the algorithm development. As a result, the processed image has the

uniform property of creating a consistent lung windowing based on Hounsfield units, aligning patients' positions, automatically cropping the scans based on the presence of the body by canny edge detector using Python library scikit-image,[24] resizing to a uniform scale of 224 × 224 by cubic spline interpolation, and standardizing to a scale of zero to one.

Traditional 2D-CNNs are designed for processing RGB images (three channels), which are usually of size 224 × 224 × 3. We use each triplet as one training or testing sample, where three CT slices correspond to three RGB channels.

Four state-of-the-art 2D-CNN structures are implemented for this disease classification task, which are MobileNet,[25] VGG16,[26] ResNet-50,[27] and DenseNet-121.[28]

To compare, a baseline CNN model is also designed with two convolutional modules and one decision module. The architecture of the baseline CNN model is provided in Fig. 4.

For all of the aforementioned models (baseline CNN, MobileNet, VGG16, ResNet-50, and DenseNet-121), we run 40 epochs using batch size of 10. We use Adam optimizer with learning rate 0.0001 for all scenarios. These hyper-parameters are selected based on exploratory attempts. Model parameters are pretrained by ImageNet[29] and updated using medical images for this task. All models are implemented using Keras.[30]

2.F. DK-enhanced training of 2D-CNN

We add a dense layer at the last layer of the CNN for all models, producing two CNN scores (IPF and non-IPF) for each input triplet. The softmax function is applied afterwards to normalize the CNN scores from two real numbers into two probabilities that sum up to 1. The two probabilities are the probabilities of the patient being classified into one of two classes: IPF ($l = 1$) or non-IPF ($l = 0$) based on their specific input triplet. Let s_{i0} and s_{i1} be the CNN scores after the last dense layers for triplet i being classified as non-IPF or IPF, respectively. Softmax function is used to calculate the normalized CNN score:

$$f(s_{il}) = \frac{\exp(s_{il})}{\exp(s_{i0}) + exp(s_{i1})}, l = 0, 1.$$

Without leveraging DK, categorical cross entropy is used as the loss function. The categorical cross entropy evaluated with deep learning model weights W at triplet i and is presented below:

$$L_{CE}^{W}(X_i, y_i) = -y_i \log(f(s_{i1})) - (1 - y_i)\log(f(s_{i0})).$$

Let X_i be the CT input triplet i, let $X = (X_1, \ldots, X_N)$ be the set of all triplets and let $y = (y_1, \ldots, y_N)$, where y_i is the label of ground truth for triplet i with $y_i = 1$ if the triplet i is sampled from an IPF patient and $y_i = 0$ if the triplet i is sampled from a non-IPF patient. The overall categorical cross entropy is calculated by averaging the categorical cross entropy across all N triplets:

$$L_{CE}^{W}(X, y) = \frac{1}{N}\sum_{i=1}^{N} L_{CE}^{W}(X_i, y_i),$$

where $N = n \times M$, n and M are the total number of patients and the number of sampled triplets from each patient, respectively (n = 1089 and M = 20 in our research).

With DK, we designed a DK-enhanced loss function, where we weigh each triplet by its D-criterion value $D(Z_i)$ and $Z_i = (z_{i1}, z_{i2}, z_{i3})$ is a 3 × 1 vector representing the SSP for triplet i, and $Z = (Z_1, \ldots, Z_N)$ is the set of SSPs for all N triplets. The DK-enhanced loss function is

$$L_{DK}^{W}(X, y, Z) = \frac{1}{N}\sum_{i=1}^{N} D(Z_i) L_{CE}^{W}(X_i, y_i).$$

Two sample proportion tests between DK and CE were conducted for the overall sensitivity, specificity, and accuracy on all five models (baseline CNN, MobileNet, VGG16, ResNet-50, and DenseNet-121), respectively. We set the significant level to be 0.05. To account for multiple hypothesis testing, we used the Bonferroni correction to set the significance cutoff for each statistical test at $\frac{0.05}{3} = 0.017$, where 3 is the number of tests for each model, that is, the overall sensitivity, specificity, and accuracy.[31]

2.G. Sensitivity analysis

Sensitivity analysis is defined as a method to determine the quality of a model by evaluating the extent to which results are impacted by changing model assumptions, methods, or certain model inputs. We design three scenarios to assess whether altering one of the preprocessing steps may lead to a different model performance, including sampling different number of triplets for each scan (scenario 1), adding an isotropic resampling step (scenario 2), and sampling triplets only from lower zones (scenario 3).

Under scenario 1, we sample a varying number of triplets (i.e., using an adaptive selection of M) for each scan based on the number of CT slices. This tests if the number of triplets should vary in scans which contain different numbers of CT slices. We empirically set M_k=0.1∗Ns_k, where M_k is the number of sampled triplets and Ns_k is the number of CT slices for patient k. For example, if one CT scan contains 250 CT slices (Ns_k= 250), we set M_k=25 for this patient k, that is, sample 25 triplets from this scan. The DK-enhanced loss function under scenario 1 is.

$$L_{DK,S1}^{W}(X, y, Z) = \frac{1}{N'}\sum_{i=1}^{N'} D(Z_i) L_{CE}^{W}(X_i, y_i),$$

where $N' = \sum_{k=1}^{n} M_k = \sum_{k=1}^{n} 0.1 \times Ns_k$, N' is the total number of triplets under scenario 1, M_k is the number of triplets for patient k,n is the total number of patients, and Ns_k is the number of CT slices for patient k.

Under scenario 2, in order to mitigate the possible confounding effects caused by varying slice thicknesses and pixel spacing, we resample all CT scans to a uniform isotropic cube of volume 1 × 1 × 1 mm^3 by cubic spline interpolation. In this step, we exclude scans which have inconsistent spacing

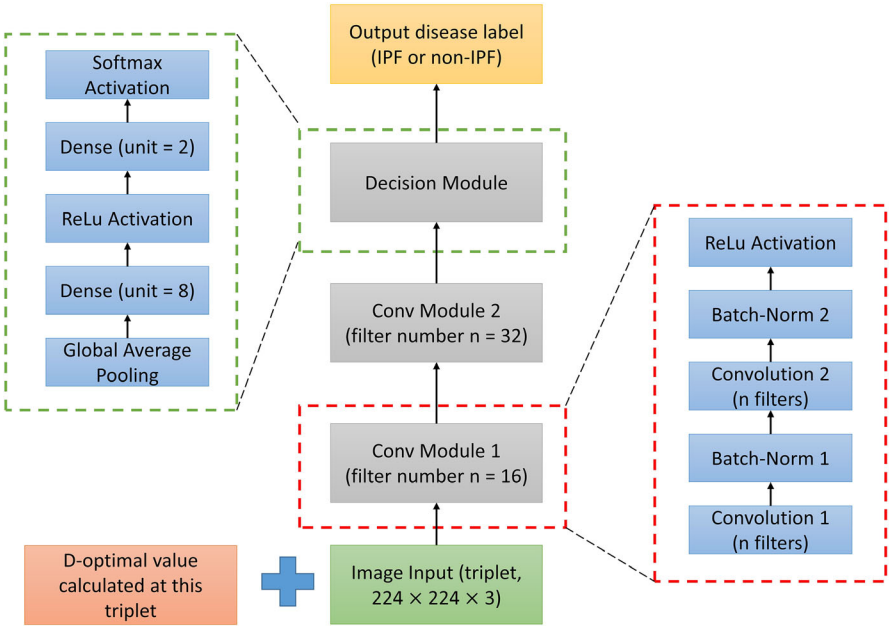

FIG. 4. Baseline CNN architecture. [Color figure can be viewed at wileyonlinelibrary.com]

along the z-dimension across all CT slices (nonvolumetric scans, N = 68, 6.2%). This step aims to align scans with different pixel spacing and slice thicknesses. The DK-enhanced loss function under scenario 2 is.

$$L^{W}_{DK,S2}(X',y,Z) = \frac{1}{N}\sum_{i=1}^{N} D(Z_i) L^{W}_{CE}(X'_i, y_i),$$

where X'_i is the CT input triplet i using X_i after isotropic resampling.

Regarding scenario 3, since IPF-related radiological features usually occur in the lower lungs, it is instructive to add one experiment to use triplets only collected from lower lungs (i.e., zone 3 in Fig. 3 (a)). Under this circumstance, the DK-enhanced loss function is.

$$L^{W}_{DK,S3}\left(\tilde{X},y,Z\right) = \frac{1}{N}\sum_{i=1}^{N} D(\tilde{Z}_i) L^{W}_{CE}(\tilde{X}_i, y_i),$$

where $\tilde{Z}_i = (z_{i\tilde{1}}, z_{i\tilde{2}}, z_{i\tilde{3}})^T$ is the 3∗1 standardized slice position for triplet i which are sampled from zone 3 only, that is, $z_{i\tilde{j}} \in (0.64, 0.9], j = 1,2,3$ for all triplet i.$\tilde{X}_i$ is the CT input triplet i collected based on the standardized slice position $\tilde{Z}_i$.

3. RESULTS

In this section, we summarize the main results and the sensitivity analysis results in 3.A and 3.B, respectively.

3.A. Main results

We pooled CT images from all five studies (two IPF studies and three non-IPF studies) together for the training and testing of the model. We performed a stratified fivefold cross-validation, a commonly used technique to separate training and testing sets, where the proportion of IPF vs non-IPF is fixed across all folds. During cross-validation, these five folds were separated at the patient level, therefore, no triplets from the same patient are evaluated in both training and testing samples. During the testing phase, M triplets were sampled from each scan following the manner as discussed, producing M predictive results (IPF vs non-IPF) for each scan. The final predictive result for each scan was decided based on majority vote of all M triplets. We set M = 20 for our task. We use sensitivity, specificity, and accuracy as statistical measures. Sensitivity is defined as the number of scans which are correctly classified as IPF divided by the total number of IPF scans. Specificity is defined as the number of scans which are correctly classified as non-IPF divided by the total number of non-IPF ILD scans. Accuracy measures the proportion of CT scans that are correctly classified.

Table II summarizes the study-wise and overall model performance using five models (Baseline CNN, MobileNet, VGG16, ResNet-50, and DenseNet-121) under two loss functions, that is, cross-entropy loss (CE) and DK-enhanced loss function (DK). Note that study 1 and study 2 include IPF patients, which is referred to as positives in this research, with sensitivity information only. Similarly, study 3, 4, and 5 contain non-IPF ILD patients, which is defined as negatives, with specificity information only. For baseline CNN model, using DK significantly increases the overall sensitivity ($P < 0.001$), but decreases the overall specificity ($P < 0.01$). There is no significant difference between DK and CE for other methods under this scenario.

3.B. Sensitivity analysis results

The complete results for scenario 1 (selecting a varying number of triplets per scan), 2 (adding isotropic resampling), and 3 (sampling from lower zones only) are provided in the Supporting Information Tables S4–S6, respectively. For each of the scenario, we calculate the absolute difference in terms of the overall model sensitivity, specificity, and accuracy between the main results (Table II) and that of each scenario. We calculate the median and interquartile range (IQR) across all ten models for each metric, under each scenario.

Under scenario 1, the median (±IQR) for the overall model sensitivity, specificity, and accuracy between the main results and that of scenario 1 across all ten model architectures is 0.04 (±0.04), 0.01 (± 0.03), and 0.02 (±0.03), respectively.

Under scenario 2, the median (±IQR) for the overall model sensitivity, specificity, and accuracy between the main results and that of scenario 2 across all models is 0.01 (±0.03), 0.01 (±0.01), and 0.01 (±0.02), respectively.

Under scenario 3, the median (±IQR) for the overall model sensitivity, specificity, and accuracy between the main results and that of scenario 3 across ten models is 0.03 (± 0.03), 0.01 (±0.01), and 0.02 (±0.01), respectively.

4. DISCUSSION

We developed a deep learning-based model for IPF diagnosis: (1) from a clinical perspective, by incorporating DK regarding the disease pattern distribution of IPF; (2) from a methodological perspective, by including optimal design methods in building a loss function. Methodologically, to the best of our knowledge, this is the first work that leverages the merits of optimal design in the training of deep learning methods in an end-to-end manner. Clinically, providing automatic IPF diagnosis support is timely and meaningful because the proposed method (1) facilitates automated IPF diagnosis and reduces inter- and intrareader disagreement; (2) enables early antifibrotic treatment and so may prolong patient's survival time; (3) decreases the likelihood of requiring of lung biopsy in the long run and its attendant's risks.

In medical imaging domain, as contrary to natural imaging, well-labeled and high-quality images are time-consuming and expensive to acquire. Therefore, many researches aim to tackle the limited sample size problem in medical imaging by utilizing DK.[32,33] Unlike previous work, we now focus on the population-level information acquired from the previous studies and utilize both DK and optimal design guidelines in the training process of the deep learning models.

Each of the earlier studies used in this research contains either IPF patients in study 1 and study 2 or non-IPF patients in study 3, study 4, and study 5, and one may argue that the diagnosis model captures confounding effects (or batch effects) rather than IPF-related CT features. Admittedly, this is one limitation of this work due to the availability of imaging data and the nature of retrospective data collection. However, we note that each study is conducted at multiple sites with different protocols and a variety of experimental conditions that likely involve CT scanners, slice thickness, reconstruction kernel, and patient positions, see the Supporting Information for an expanded list of potential confounders. This heterogeneous experimental setup contributes to a fair model that concentrates on the underlying CT features of IPF rather than picking up other confounding factors.

In addition, to address this concern of confounding effects, we have added multiple model generalizability experiments (see Supporting Information E for more details). By setting aside one study as the holdout test set at one time, we evaluate the generalizability of the constructed model to unseen domains (i.e., institutions and clinical diagnoses) using MobileNet. The results suggest that, most experiments can

TABLE II. Study-wise model performance and overall model performance.

	Sensitivity (IPF patients)		Specificity (Non-IPF ILD patients)			Overall model performance		
Model (Loss function)	Study 1	Study 2	Study 3	Study 4	Study 5	Sensitivity	Specificity	Accuracy
Baseline CNN (CE)	0.77 (0.38)	0.68 (0.39)	0.96 (0.04)	0.94 (0.09)	0.98 (0.02)	0.74 (0.38)	**0.97 (0.03)**	0.89 (0.12)
Baseline CNN (DK)	0.89 (0.13)	0.81 (0.20)	0.91 (0.07)	0.88 (0.19)	0.96 (0.03)	**0.86 (0.15)**	0.94 (0.05)	0.91 (0.04)
MobileNet (CE)	0.97 (0.01)	0.96 (0.07)	1 (0)	0.96 (0.05)	0.99 (0.02)	0.97 (0.02)	0.98 (0)	0.98 (0.01)
MobileNet (DK)	0.98 (0.02)	0.94 (0.06)	1 (0)	0.96 (0.04)	0.98 (0.01)	0.96 (0.02)	0.98 (0.01)	0.97 (0.01)
VGG16 (CE)	0.96 (0.03)	0.87 (0.07)	0.99 (0.02)	0.95 (0.06)	0.99 (0.01)	0.93 (0.04)	0.98 (0.01)	0.96 (0.01)
VGG16 (DK)	0.95 (0.04)	0.86 (0.09)	0.99 (0.02)	0.95 (0.06)	0.99 (0.01)	0.92 (0.05)	0.98 (0.01)	0.96 (0.01)
ResNet-50 (CE)	0.96 (0.02)	0.92 (0.05)	0.98 (0.05)	0.97 (0.03)	0.99 (0.01)	0.95 (0.02)	0.98 (0.01)	0.97 (0.01)
ResNet-50 (DK)	0.96 (0.02)	0.90 (0.09)	1 (0)	0.96 (0.05)	0.99 (0.01)	0.94 (0.03)	0.98 (0.01)	0.97 (0.01)
DenseNet-121 (CE)	0.97 (0.02)	0.98 (0.02)	1 (0)	0.97 (0.04)	0.98 (0)	0.97 (0.01)	0.98 (0.01)	0.98 (0)
DenseNet-121 (DK)	0.96 (0.04)	0.94 (0.06)	1 (0)	0.97 (0.04)	0.99 (0)	0.95 (0.02)	0.99 (0.01)	0.97 (0)

Note: Mean and standard deviations shown in brackets are calculated across the results from each testing fold. CE: cross entropy loss without domain knowledge-enhanced loss function; DK: domain knowledge-enhanced loss function. Statistically significant results ($P < 0.017$) are highlighted in bold font. The significance cutoff 0.017 is decided by Bonferroni correction for multiple testing, which is dividing the prespecified significance level 0.05 by the number of tests (3, including the overall sensitivity, specificity, and accuracy) for each model.

successfully classify more than 90% of patients in the holdout study (accuracies greater than 90%). This suggests that most experiments are able to generalize well to unseen domains. Notably, there is a certain level of decrease in overall model accuracy compared to results provided in the Table II, when using one study as the holdout study at a time. For example, for six of eight generalizability experiments, we observe a 1%–4% degradation in model accuracy; for two of eight experiments, we observe a 25%-26% decrease in model accuracy, which we provide some explanations in the Supporting Information E. This degradation in performance may due to the fact that the number of training and testing samples are fewer since we set one study aside as the holdout set. At the same time, this lack of generalizability is not surprising as such findings are frequently reported in many areas of research when deep learning models are applied to unseen domains.[34] This provides a warning that when deploying the developed model to scans collected from other institutions or ILD patients with different clinical diagnoses, some decrease in model performance is to be expected. Many domain adaptation and domain generalization techniques have been developed to tackle this problem, but they are out of the scope for this paper.[35]

In summary, we have, for the first time, incorporated the population-level DK (i.e., IPF progression trends across the lung position acquired from pilot studies) with ideas of optimal design methodology into the training of deep learning models. Specifically, we sample 20 triplets from each patient visit to augment the number of training data and boost model performance. These triplets were randomly sampled with one from each zone (the top, middle, and bottom of the lungs). Intuitively, these 20 triplets should not be treated identically, as these randomly sampled CT slices might not be fully representative and reflect the disease characteristics fairly. Some triplets might contain three slices which are adjacent to each other, and thus contain less disease information. To this end, we estimated the population-level disease trends across lung positions from previous studies and evaluated the importance of each triplet by its D-optimality value. The triplet with a larger value is "a better design" for estimating the parameters of the population-level trends, and consequently, it is believed to be more representative of the overall disease trends. We then design the DK-enhanced loss function, where the D-criterion value of each triplet is used as a weight to evaluate the importance of each triplet. This process is incorporated into the training of the deep learning models in an end-to-end manner.

Current experiments show that incorporating DK in the training of deep learning models increases the overall accuracy from 0.89 to 0.91 for the baseline CNN model. However, this increase in the overall accuracy using DK is not observed for other well-known model architectures, including MobileNet, VGG16, ResNet50, and DenseNet-121. This may occur due to the existence of ceiling effect, since other well-developed deep learning architectures have already achieved a satisfactory model performance with overall accuracy greater than 0.95. We also expect the proposed methodology is generally applicable to tackle other similar problems in the medical arena as well, even though our work here only concerns IPF diagnosis.

Sensitivity analysis experiments suggest that (1) selecting a flexible number of triplets per scan, (2) isotropic resampling each scan to a constant size of 1 mm^3 cube, and (3) sampling triplets only from lower zones may change the overall model sensitivity, specificity, accuracy in a reasonable range.

Our future work includes exploring the constructed model on prospective studies, where IPF and non-IPF ILD patients are collected under the same imaging protocols. This is a more accurate reflection of the clinical applicability of the developed model, as contrary to using fivefold cross-validation without independent studies.

5. CONCLUSIONS

We develop an efficient IPF diagnosis model using DK (i.e., population-level disease information) and optimal design theory. This study shows satisfactory performance using various well-known deep learning models in the task of IPF diagnosis using CT images. To the best of our knowledge, this is the first work that (1) leverages population DK with optimal design criterion to train deep learning models in an end-to-end fashion; (2) focuses on patient-level IPF diagnosis solely based on CT images.

ACKNOWLEDGMENTS

This research is supported by NIH, NHLBI-R21-HL140465. Hua Zhou was supported by grants from the National Human Genome Research Institute (HG006139) and the National Institute of General Medical Sciences (GM053275).

CONFLICT OF INTEREST

The authors have no conflict to disclose.

DATA AVAILABILITY STATEMENT

Raw data were generated from UCLA. Derived data from HRCT images supporting the findings of this study are available from the corresponding author GK on reasonable request.

[a)]Author to whom correspondence should be addressed. Electronic email: GraceKim@mednet.ucla.edu.

REFERENCES

1. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med*. 2011;183:788–824.
2. Nalysnyk L, Cid-Ruzafa J, Rotella P, Esser D. Incidence and prevalence of idiopathic pulmonary fibrosis: review of the literature. *Eur Respir Rev*. 2012;21:355–361.

3. Raghu G, Remy-Jardin M, Myers JL, et al. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med*. 2018;198:e44–e68.
4. Walsh SL, Calandriello L, Sverzellati N, Wells AU, Hansell DM. Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT. *Thorax*. 2016;71:45–51.
5. Hutchinson JP, Fogarty AW, McKeever TM, Hubbard RB. In-hospital mortality after surgical lung biopsy for interstitial lung disease in the United States. 2000 to 2011. *Am J Respir Crit Care Med*. 2000; 2016:1161–1167.
6. King TE, Bradford WZ, Castro-Bernardini S, et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med*. 2014;370:2083–2092.
7. Richeldi L, du Bois RM, Raghu G, et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med*. 2014;370: 2071–2082.
8. Kim HJ, Brown MS, Chong D, et al. Comparison of the quantitative CT imaging biomarkers of idiopathic pulmonary fibrosis at baseline and early change with an interval of 7 months. *Acad Radiol*. 2015;22:70–80.
9. Wu X, Kim GH, Salisbury ML, et al. Computed tomographic biomarkers in idiopathic pulmonary fibrosis. The future of quantitative analysis. *Am J Respir Crit Care Med*. 2019;199:12–21.
10. Chong DY, Kim HJ, Lo P, et al. Robustness-driven feature selection in classification of fibrotic interstitial lung disease patterns in computed tomography using 3D texture features. *IEEE Trans Med Imaging*. 2015;35:144–157.
11. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans Med Imaging*. 2016;35:1207–1216.
12. Shin H-C, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35:1285–1298.
13. Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. ArXiv Prepr ArXiv171105225. Published online. 2017.
14. Walsh SL, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *Lancet Respir Med*. 2018;6:837–845.
15. Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng*. 2019;3:173.
16. Pereira M, Fantini I, Lotufo R, Rittner L. An extended-2D CNN for multiclass Alzheimer's Disease diagnosis through structural MRI. In: Medical Imaging 2020: Computer-Aided Diagnosis. Vol 11314. International Society for Optics and Photonics; 2020:113141V.
17. Zhang S, Han F, Liang Z, et al. An investigation of CNN models for differentiating malignant from benign lesions using small pathologically proven datasets. *Comput Med Imaging Graph*. 2019;77:101645.
18. Shi Y, Wong WK, Goldin JG, Brown MS, Kim GHJ. Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization-Random forest approach. *Artif Intell Med*. 2019;100:101709.
19. Berger MP, Wong W-K. *An Introduction to Optimal Designs for Social and Biomedical Research*, vol. 83. John Wiley & Sons; 2009.
20. Berger MP, Wong W-K. *Applied Optimal Designs*. Hoboken: John Wiley & Sons; 2005.
21. Shi Y, Kee WW, Goldin JG, Brown MS, Kim HJ. Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: a quantum particle swarm optimization - random forest approach. *Artif Intell Med*. 2019;100:101709.
22. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *J R Stat Soc Ser C Appl Stat*. 1994;43:429–453.
23. Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statistics. *Dordr Neth Reidel*. 1986;81.
24. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ*. 2014;2:e453.
25. Howard AG, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv Prepr ArXiv170404861. Published online. 2017.
26. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv Prepr ArXiv14091556. Published online. 2014.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016:770–778.
28. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017:4700–4708.
29. Deng J, Dong W, Socher R, Li L-J, Li K, Imagenet F-F. A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee; 2009:248–255.
30. Keras CF. GitHub; 2015. https://github.com/fchollet/keras.
31. Shaffer JP. Multiple hypothesis testing. *Annu Rev Psychol*. 1995;46: 561–584.
32. Chai Y, Liu H, Xu J. Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowl-Based Syst*. 2018;161:147–156.
33. Pape C, Matskevych A, Wolny A, et al. Leveraging domain knowledge to improve microscopy image segmentation with lifted multicuts. *Front Comput Sci*. 2019;1:6.
34. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15:e1002683.
35. Dou Q, de Castro DC, Kamnitsas K, Glocker B. Domain generalization via model-agnostic learning of semantic features. *Adv Neural Inf Process Syst*. 2019:6450–6461.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig S1. The true median curve in blue shows the percentage of progressive pixels vs standardized slice position (SSP). The other colored curves are the best fits to the overall population trends from the other three models. The gray area represents the 95% quantile of the true curve and the two dotted vertical lines at SSP = 0.10 and 0.90 represent the noticeable boundary effects.

Fig S2. Distribution of criterion values while fixing z_1, z_2, z_3 respectively.

Table S1. CT acquisition and image reconstruction conditions of the five studies.

Table S2. Model fitting performance: three preselected models and their corresponding estimated parameters and AIC values.

Table S3. Experimental setup and results for model generalizability testing by *using one study at a time as the holdout test study.*

Table S4. Study-wise model performance and overall model performance with *an adaptive selection of triplets per scan*.

Table S5. Study-wise model performance and overall model performance by *adding a resampling step* during the preprocessing procedure.

Table S6. Study-wise model performance and overall model performance using triplets collected from *lower zones only.*