Journal of Cystic Fibrosis 19 (2020) 131–138

Contents lists available at ScienceDirect

Journal of Cystic Fibrosis

journal homepage: www.elsevier.com/locate/jcf

Original Article

# Deep learning to automate Brasfield chest radiographic scoring for cystic fibrosis

Evan J. Zucker[a,*], Zachary A. Barnes[b], Matthew P. Lungren[a], Yekaterina Shpanskaya[a], Jayne M. Seekins[a], Safwan S. Halabi[a], David B. Larson[a]

[a] Department of Radiology, Stanford University School of Medicine, 725 Welch Road, Stanford, CA 94305, USA
[b] Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA

ARTICLE INFO

*Article history:*
Received 25 January 2019
Revised 27 March 2019
Accepted 21 April 2019
Available online 2 May 2019

*Keywords:*
Deep learning
Deep convolutional neural network
Cystic fibrosis
Brasfield
Chest
Radiograph

ABSTRACT

*Background:* The aim of this study was to evaluate the hypothesis that a deep convolutional neural network (DCNN) model could facilitate automated Brasfield scoring of chest radiographs (CXRs) for patients with cystic fibrosis (CF), performing similarly to a pediatric radiologist.

*Methods:* All frontal/lateral chest radiographs (2058 exams) performed in CF patients at a single institution from January 2008–2018 were retrospectively identified, and ground-truth Brasfield scoring performed by a board-certified pediatric radiologist. 1858 exams (90.3%) were used to train and validate the DCNN model, while 200 exams (9.7%) were reserved for a test set. Five board-certified pediatric radiologists independently scored the test set according to the Brasfield method. DCNN model vs. radiologist performance was compared using Spearman correlation (ρ) as well as mean difference (MD), mean absolute difference (MAD), and root mean squared error (RMSE) estimation.

*Results:* For the total Brasfield score, ρ for the model-derived results computed pairwise with each radiologist's scores ranged from 0.79–0.83, compared to 0.85–0.90 for radiologist vs. radiologist scores. The MD between model estimates of the total Brasfield score and the average score of radiologists was −0.09. Based on MD, MAD, and RMSE, the model matched or exceeded radiologist performance for all subfeatures except air-trapping and large lesions.

*Conclusions:* A DCNN model is promising for predicting CF Brasfield scores with accuracy similar to that of a pediatric radiologist.



## 1. Introduction

Many imaging scoring systems have been validated to quantitate the severity of cystic fibrosis (CF) lung disease [1–10]. Among them, the Brasfield chest radiograph method may be preferred in practice because it performs similarly to other methods and yet is comparatively simpler and faster to use and interpret. In fact, Brasfield scores are routinely requested by pediatric pulmonologists and other providers treating CF patients at our institution. The Brasfield system is reported to have high rates of intra- and interrater reliability based on prior studies of pediatric radiologists and pediatricians at CF centers [4–7]. However, among generalists with only limited exposure to sporadic CF cases, accurate and consistent scoring may be difficult to achieve. Moreover, even when specialists are available, scoring is a time-consuming task that may be infeasible during a busy workday.

In recent years, deep learning utilizing convolutional neural networks, a form of machine learning, has been successfully employed to accurately perform a variety of image recognition and classification tasks by encoding hierarchies of spatial features through adaptive mathematical models [11–14]. Emerging medical applications range from automated bone age assessment to automated CXR diagnosis [11–21]. Deep learning thus also has potential to automate Brasfield scoring, thereby reducing the need for subspecialized readers to perform tedious processes while maintaining reliable quantitative metrics. The purpose of this study was to develop a deep convolutional neural network (DCNN) model for automated Brasfield scoring and to compare its performance to that of pediatric radiologists.

## 2. Methods

The Institutional Review Board (IRB) approved the review of radiologic and clinical data for this retrospective study. Informed

* Corresponding author at: Department of Radiology, Stanford University School of Medicine, 725 Welch Road, Stanford, CA 94305, USA.
*E-mail address:* zucker@post.harvard.edu (E.J. Zucker).

https://doi.org/10.1016/j.jcf.2019.04.016
1569-1993/

consent was waived, but patient confidentiality was protected in accordance with Health Insurance Portability and Accountability Act (HIPAA) guidelines.

### 2.1. Data acquisition

Using the authors' picture archiving and communication system (PACS) based at a single tertiary care children's hospital, all consecutive two-view (frontal/lateral) chest radiographic examinations performed in patients with CF for any indication (either routine, e.g., as part of the annual check-up, or non-routine, e.g., before exacerbation treatment) from January 1, 2008 through January 1, 2018 were retrospectively extracted. CXRs were performed on a GE Discovery XR650 digital radiography system (GE Healthcare, Chicago, IL), using exposure charts and automatic exposure control (AEC) to account for variations in required exposure with patient size. Exams were performed in inspiration for patients who could understand and execute radiographer instructions and otherwise during quiet breathing. At the authors' institution, Brasfield scores are frequently requested by providers ordering two-view CF CXRs and subsequently reported by the interpreting pediatric radiologists on request. However, not every two-view CF CXR report contains a Brasfield score or all elements of the Brasfield score. Thus, the decision was made to extract all two-view CF CXRs regardless of the presence of Brasfield scores in the report to increase the number of available exams. Among all available two-view CXRs performed during the 10-year time period, those performed in patients with CF were identified by reviewing all corresponding radiology reports and selecting those with mention of "cystic fibrosis," "CF," or "Brasfield" in the exam history/indication section. No demographic or other exclusion criteria applied. However, single-view (e.g., frontal-only) CXRs in CF patients were not extracted, as the Brasfield system was designed only for frontal/lateral CXR exams [4]. The final study dataset consisted of 2058 exams obtained in 451 unique patients.

### 2.2. Image labeling and data partitioning

A single board-certified pediatric radiologist (E.J.Z.) with additional cardiothoracic subspecialty expertise, 5 years post-fellowship experience, and prior research training in quantitative CF CXR evaluation, herein denoted the universal reader, retrospectively scored all CXRs according to the Brasfield system in random order over a 1 month period. The Brasfield method consists of assigning 5 subscores to each paired frontal/lateral CXR according to 5 features, with higher scores indicating more severe radiographic disease. The features scored are: air trapping (0–4), linear markings (0–4), nodular cystic lesions/bronchiectasis (0–4), large lesions (0 = absent, 3 = segmental/lobar atelectasis or pneumonia, 5 = multiple atelectasis or pneumonia), and general severity (0–5, with scores of 4–5 reserved for cases with complications such as cardiac enlargement or pneumothorax). The subscores are subtracted from 25 to generate the total score, with lower scores indicating more severe disease [4–6]. The decision was made to rescore all available exams due to absent or inconsistent application of the Brasfield system in the original radiology reports.

From the 2058 available exams, a test set consisting of 200 CXRs (9.7%s) performed in 130 unique patients was assembled using stratified random sampling, facilitating approximately equal distributions of each Brasfield subscore based on the universal reader's scoring evaluations. Due to a preponderance of mild radiographic disease among the available CXRs, more severe cases (higher Brasfield subscores) were oversampled. The remaining 1858 exams (90.3%) performed in 451 unique patients comprised the training/validation set used to construct deep learning model, as detailed in the next section. The approximately 90%/10% distribution of training to test set data was selected to preserve as much data as possible for effective model learning while conserving an evaluation set reflective of the range radiographic severity for each Brasfield feature; a similar distribution has been used in prior studies of deep learning for medical imaging [19,22].

### 2.3. Deep learning model

The problem of generating Brasfield total and subscores for a CF patient was formulated as an image classification machine learning problem that could be approached using a DCNN model. The goal was to develop a model that could predict a set of labels, that is, the Brasfield total and subscores, for a newly encountered CF CXR exam, based on example CXRs that had already been labeled with the correct Brasfield scores as inputs. This process is akin to a radiologist prospectively scoring a CXR according to the Brasfield system based on previous experience scoring different CXRs. In this case, the 1858 CXRs reserved for model training/validation, which had been "labeled" by the universal reader with their corresponding Brasfield total and subscores, served as the examples from which the model could "learn."

The DCNN model for predicting the Brasfield total and subscores was constructed using the open source TensorFlow framework (API r1.12, Google LLC, Mountain View, CA) on a single NVIDIA GeForce GTX 1080 GPU (Santa Clara, CA). All frontal/lateral CXR images were collected from the 2058 available exams. These images were first converted from DICOM to portable network graphics (PNG) format and rescaled to $256 \times 256$ pixels. A Histogram-based intensity standardization algorithm (CLAHE) was applied to each image, improving contrast via setting the range of potential pixel values to 0–255 [11,23].

The model was based on the ResNet-18 architecture, an 18-layer-deep CNN [24,25]. This architecture is pre-trained on 1.2 million images across 1000 classes from the ImageNet database, a publicly available online repository containing labeled images of everyday day objects [26]. This pre-training or transfer learning process permits optimization of the early DCNN layers responsible for recognition of generic image features, such as edges, which are shared between even seemingly disparate objects, and has been utilized in prior studies of medical image classification [11,19,20,27]. To refine the model, the pre-trained weights on the first 15 layers of the model were maintained. Then, the model weights for the outer layers were iteratively adjusted through training process, which involved passing only the labeled CXR images reserved for training (i.e., 1858 exams in total), after further rescaling to $224 \times 224$ pixels, through the model a total of 100 times (epochs) in batches of 32 CXRs; the goal in each iteration was to further converge on the model parameters that could best predict the Brasfield subscores, using an Adam stochastic optimization algorithm to minimize an ordinal regression loss function (degree of error from desired output) [11,28]. A random data augmentation input scheme for the training set images was performed using a combination of rotation, shear, cropping, brightening transformations, effectively increasing the training set size, while preserving image labels [11,29]. Tuning of hyperparameters (fixed model weights not "learned" through the training process) was performed using a random 10% selection of the training data, which served to validate the performance of the model. The final model outputs consisted of the probability of each potential subscore for each Brasfield feature, from which the total Brasfield score could be derived.

To provide visual validation of the quantitative model results, class activation mappings (CAMs) were also implemented. In these saliency or "heat" maps, the absolute magnitude of the partial derivative of the loss function associated with a particular ground-truth label (Brasfield subscore) is displayed with respect to each input pixel. As a result, pixels with greater magnitude in the map

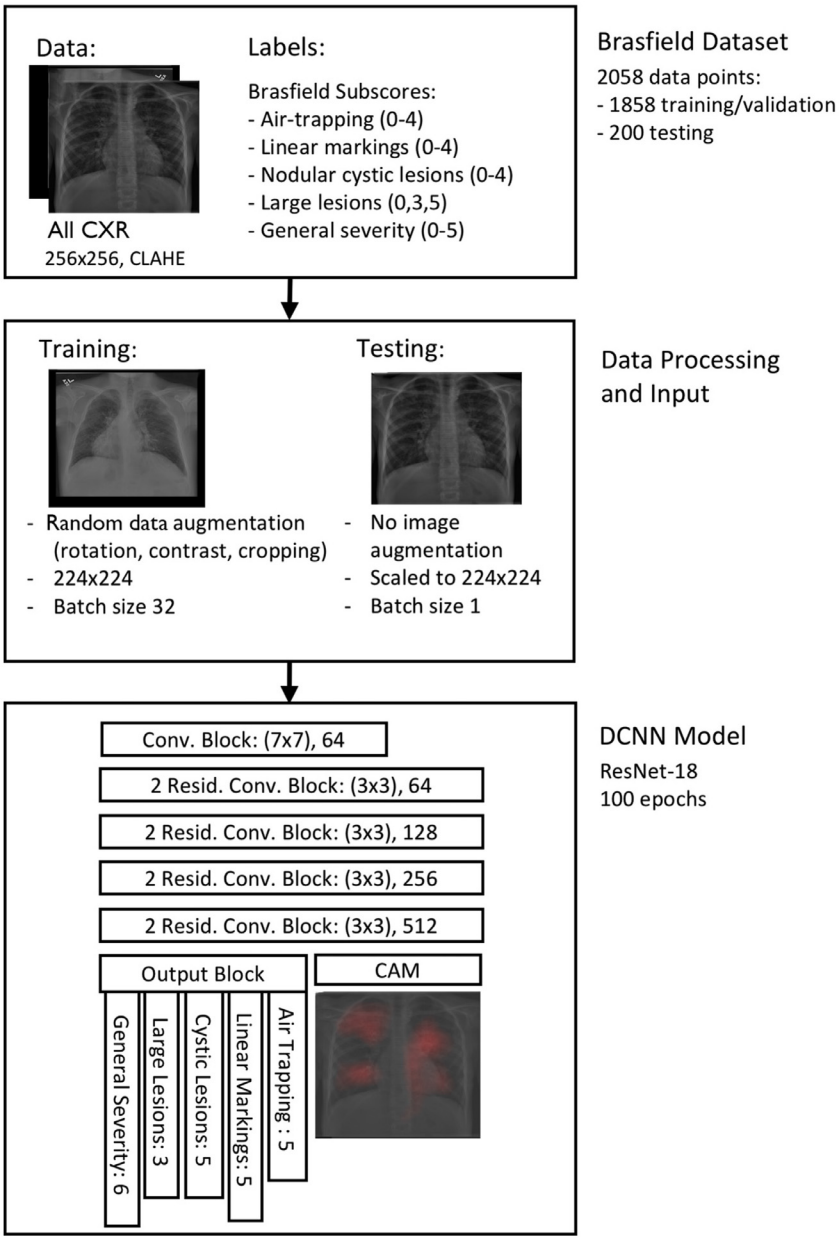


**Fig. 1.** Schematic representation of the DCNN construction process for predicting Brasfield scores. The available input data consists of 2058 total CXRs. Each CXR has been labeled by the universal reader with a subscore for each Brasfield feature (air-trapping, linear markings, nodular cystic lesions, large lesions, and general severity). During data preprocessing, the CXRs are rescaled to $256 \times 256$ pixels. A histogram-based intensity standardization algorithm (CLAHE) is applied to normalize the range of potential pixel values to 0–255. Then, the CXRs are divided into a set of 1858 exams used to train and validate the DCNN model and a set of 200 exams used to test the performance of the model after it is trained. To improve the efficiency of the model, the images are further scaled down to $224 \times 224$ pixels. The DCNN is based on the ResNet-18 architecture, an 18-layer-deep model, which has been pretrained on 1.2 million images of everyday objects available from the ImageNet database. The inner layers of the model, which are attuned to detecting low-level features such as edges, are left in place. During the training process, random operations are applied to the CXRs not reserved for testing such as rotation and contrast adjustment, which effectively augment the dataset while preserving the assigned labels (i.e., Brasfield subscores). Then, through an iterative process, random batches of 32 CXRs are continually passed to the model. During each iteration, the weighting factors of the model's outer layers are adjusted, improving the ability of the model to predict the Brasfield subscores for a given CXR. This is achieved through a mathematical optimization process; the goal is to minimize a loss function that computes how discrepant the model output (predicted Brasfield subscores) and the expected output (labeled Brasfield scores) are for a given set of inputs (CXRs). This process is repeated for 100 complete passes through the training data, or epochs, after which the model converges to an empirically optimal solution. Once the model has been fully trained, it is run on the test set. For each input to the model (a CXR it has not previously encountered), the model provides, for each Brasfield feature, the probability of each Brasfield subscore. The subscore with highest probability is taken as the predicted score. In addition, class activation maps (CAMs) are generated for each Brasfield feature; these show the portions of the image that are most salient in discerning the model's predicted subscores.

correspond to those of most importance in determining the model output for that subtype [11,19,20,30].

A schematic representation of the DCNN construction process for Brasfield scoring is shown in Fig. 1. Further details on machine learning principles and their application to automating Brasfield scoring can be found in the online supplement (Appendix A).

### 2.4. Model evaluation

A reference standard document was first created in consensus between the universal reader and 4 additional board-certified pediatric radiologists (M.P.L., D.B.L., J.M.S., and S.S.H., with 6, 12, 12, and 11 years post-fellowship experience, respectively) to

improve consistency in application of the Brasfield system. This document consisted of 3 anonymized and randomly selected prototype image examples of each Brasfield subscore, drawn from the CXRs reserved for DCNN model training and based on the universal reader's scores, accompanied by text denoting the specific features justifying a particular score along the severity scale. Then, the additional radiologists independently scored each exam in the test set cohort in random order according to the Brasfield system, using the reference standard as a guide, and blinded to patient demographics, specific exam history (other than CF), and other radiologists' scores.

The model was then also applied to the test set cohort. For each case, the DCNN-predicted subscore for each Brasfield feature was automatically derived by summing the weighted probabilities of each potential score, then rounding to the nearest whole number. The total Brasfield score for each case was derived by subtracting the sum of the (rounded) predicted subscores from 25. Heat maps were also created for each test set case and visually reviewed to establish concordance between the model's predicted Brasfield subscores and the image features contributing to those predictions.

### 2.5. Statistical analysis

To assess the performance of the DCNN model relative to that of the pediatric radiologists, Spearman correlation analysis was utilized. Pairwise model-radiologist correlation coefficients were calculated for each subscore as well as the total Brasfield score. In addition, pairwise radiologist-radiologist correlation coefficients were calculated to discern radiologist interobserver differences in scoring the test set.

The performance of the model was also assessed by calculating the mean difference (MD), mean absolute difference (MAD) and root mean squared error (RMSE) between the model estimates for each Brasfield subscore and total score and the mean sub- and total scores of the 5 pediatric radiologists, used as the reference standard. Moreover, in pairwise fashion, the MD, MAD, and RMSE of each radiologist's scores was calculated relative to the mean of the other human reviewers' scores. At the same time, the MD, MAD, and RMSE of the model's scores relative to the mean of the 4 radiologists' scores under consideration were also calculated, thus allowing simultaneous comparison of each radiologist's and the model's deviation in scoring relative to that of the remaining radiologists.

Categorical data were compared using the Fisher exact test. Means were compared using the Mann-Whitney *U* test. Variances (i.e., RMSE) were compared with the Levene test. Statistical significance was set at the $p \leq 0.05$ level. All statistical analysis was carried out using Microsoft Excel (version 2016, Microsoft Corporation, Redmond, WA) and Stata software (version 14, StataCorp LP, College Station, TX). Percentages were rounded to the nearest 10th of a percent.

## 3. Results

Summary characteristics of the training/validation and test sets are presented in Table 1. The mean ± SD age of test patients

**Table 1**
Summary characteristics of the training/validation and test image data sets.

| Variable | Training/Validation set | Test set | *P*-Value for difference |
|---|---|---|---|
| No. of exams (%total exams) | 1858 (90.3%) | 200 (9.7%) | N/A |
| No. of unique patients | 451 | 130 | N/A |
| Mean age (SD), years | 10.4 (7.7) | 12.0 (7.9) | $p = 0.001^{\dagger}$ |
| Gender | | | |
| Male | 826 (44.5%) | 87 (43.5%) | $p = 0.82$ |
| Female | 1032 (55.5%) | 113 (56.5%) | |
| Brasfield air-trapping score | | | |
| No. with score of 0 (%) | 322 (17.3%) | 34 (17.0%) | $p < 0.001^{\dagger}$ |
| No. with score of 1 (%) | 834 (44.9%) | 54 (27.0%) | |
| No. with score of 2 (%) | 334 (18.0%) | 43 (21.5%) | |
| No. with score of 3 (%) | 303 (16.3%) | 43 (21.5%) | |
| No. with score of 4 (%) | 65 (3.5%) | 26 (13.0%) | |
| Brasfield linear markings score | | | |
| No. with score of 0 (%) | 196 (10.6%) | 32 (16.0%) | $p < 0.001^{\dagger}$ |
| No. with score of 1 (%) | 961 (51.7%) | 54 (27.0%) | |
| No. with score of 2 (%) | 312 (16.8%) | 52 (26.0%) | |
| No. with score of 3 (%) | 333 (17.9%) | 42 (21.0%) | |
| No. with score of 4 (%) | 56 (3.0%) | 20 (10.0%) | |
| Brasfield nodular cystic lesions score | | | |
| No. with score of 0 (%) | 1093 (78.6%) | 81 (40.5%) | $p < 0.001^{\dagger}$ |
| No. with score of 1 (%) | 163 (8.8%) | 13 (6.5%) | |
| No. with score of 2 (%) | 200 (10.8%) | 36 (18.0%) | |
| No. with score of 3 (%) | 316 (17.0%) | 39 (19.5%) | |
| No. with score of 4 (%) | 86 (4.6%) | 31 (15.5%) | |
| Brasfield large lesions score | | | |
| No. with score of 0 (%) | 1610 (86.7%) | 147 (73.5%) | $p < 0.001^{\dagger}$ |
| No. with score of 3 (%) | 202 (10.9%) | 34 (17.0%) | |
| No. with score of 5 (%) | 46 (2.5%) | 19 (9.5%) | |
| Brasfield general severity score | | | |
| No. with score of 0 (%) | 124 (6.7%) | 23 (11.5%) | $p < 0.001^{\dagger}$ |
| No. with score of 1 (%) | 1067 (57.4%) | 69 (34.5%) | |
| No. with score of 2 (%) | 276 (14.9%) | 41 (20.5%) | |
| No. with score of 3 (%) | 390 (21.0%) | 62 (31.0%) | |
| No. with score of 4 (%) | 1 (0.1%) | 5 (2.5%) | |
| No. with score of 5 (%) | 0 (0.0%) | 0 (0.0%) | |
| Brasfield total score (SD) | 19.1 (4.7) | 16.9 (6.0) | $p < 0.001^{\dagger}$ |

Abbreviations: N/A = not applicable; SD = standard deviation.
Note: Mean (SD) age and gender percentages are calculated based on an exam-level (not unique patient-level) basis. Age is defined as the time in years elapsed from date of birth to date of exam. Brasfield scores shown correspond to the universal reader's evaluation.
$^{\dagger}p \leq 0.05$.

was 12.0 ± 7.9 years, in comparison to 10.4 ± 7.7 years for the training/validation set ($p < 0.001$). Gender distributions were not statistically different (44.5% male for the training/validation set vs. 43.5% for the test set, $p = 0.822$). The mean ± SD total Brasfield score, as labeled by the universal reader, was 16.9 ± 6.0 for the test set, compared to 19.1 ± 4.7 for the training/validation set ($p < 0.001$). All subscore groupings (air-trapping, linear markings, nodular cystic lesions, large lesions, and general severity) were also statistically different in the test set group ($p < 0.001$), which was prospectively chosen to incorporate more abnormal cases, compared to the training/validation set, as detailed in Table 1. The DCNN model completed training in approximately 14.5 h and the test set in approximately 5 s.

Results of the correlation analysis are presented in Table 2. For the total Brasfield score, Spearman correlation coefficients of the model-derived results compared pairwise with each radiologist's scores ranged from 0.79–0.83. In comparison, when each radiologist was compared pairwise with every other radiologist, correlation coefficients for the total Brasfield score ranged from 0.85–0.90. For air-trapping, linear markings, nodular cystic lesions, large lesions, and general severity, respectively, model vs. radiologist score correlations ranged from 0.51–0.78, 0.70–0.78, 0.75–0.83, 0.29–0.38, and 0.76–0.82, respectively. In comparison, radiologist vs. radiologist score correlations ranges for these subscores in the same order were 0.48–0.78, 0.71–0.82, 0.71–0.87, 0.74–0.84, and 0.82–0.89, respectively.

Results of MD, MAD, and RMSE analyses are presented in Table 3, Supplementary Table S1, and Supplementary Table S2, respectively. For the total Brasfield score, the MD, MAD, and RMSE of the model compared to the average of all 5 radiologists were − 0.09, 2.03, and 2.71, respectively. There was no systematic trend between overall radiographic disease severity (based on total Brasfield score) and model performance; based on MD, the model performed best for intermediate disease and worst for the most severe disease (Supplementary Table S3). For subscores, the MD, MAD, and RMSE of the model's estimates compared to the average subscores of all 5 radiologists were in the range of −0.19-0.24, 0.40–0.86, and 0.56–1.56, respectively.

**Table 2**
Radiologist v. model correlation in brasfield score estimates.

| | Rad 1 | Rad 2 | Rad 3 | Rad 4 | Rad 5 |
|---|---|---|---|---|---|
| Model | | | | | |
| Air-trapping | 0.78 | 0.51 | 0.70 | 0.68 | 0.62 |
| Linear markings | 0.74 | 0.78 | 0.76 | 0.70 | 0.78 |
| Nodular cystic lesions | 0.83 | 0.80 | 0.82 | 0.75 | 0.81 |
| Large lesions | 0.29 | 0.35 | 0.38 | 0.28 | 0.30 |
| General severity | 0.82 | 0.76 | 0.79 | 0.81 | 0.78 |
| Total score | 0.83 | 0.79 | 0.83 | 0.83 | 0.83 |
| Rad 1 | | | | | |
| Air-trapping | | 0.58 | 0.78 | 0.69 | 0.69 |
| Linear markings | | 0.79 | 0.79 | 0.76 | 0.83 |
| Nodular cystic lesions | | 0.87 | 0.87 | 0.77 | 0.87 |
| Large lesions | | 0.77 | 0.76 | 0.78 | 0.74 |
| General severity | | 0.85 | 0.84 | 0.82 | 0.84 |
| Total score | | 0.87 | 0.89 | 0.88 | 0.88 |
| Rad 2 | | | | | |
| Air-trapping | | | 0.59 | 0.48 | 0.58 |
| Linear markings | | | 0.79 | 0.75 | 0.82 |
| Nodular cystic lesions | | | 0.85 | 0.77 | 0.87 |
| Large lesions | | | 0.83 | 0.76 | 0.84 |
| General severity | | | 0.85 | 0.85 | 0.89 |
| Total score | | | 0.88 | 0.85 | 0.89 |
| Rad 3 | | | | | |
| Air-trapping | | | | 0.75 | 0.71 |
| Linear markings | | | | 0.71 | 0.78 |
| Nodular cystic lesions | | | | 0.71 | 0.86 |
| Large lesions | | | | 0.75 | 0.80 |
| General severity | | | | 0.84 | 0.84 |
| Total score | | | | 0.90 | 0.90 |
| Rad 4 | | | | | |
| Air-trapping | | | | | 0.63 |
| Linear markings | | | | | 0.74 |
| Nodular cystic lesions | | | | | 0.75 |
| Large lesions | | | | | 0.74 |
| General severity | | | | | 0.88 |
| Total score | | | | | 0.89 |

Abbreviations: Rad = radiologist.
Note: Values correspond to Spearman correlation coefficients.

Results of MD, MAD, and RMSE analyses are presented in Table 3, Supplementary Table S1, and Supplementary Table S2, respectively. For the total Brasfield score, the MD, MAD, and RMSE of the model compared to the average of all 5 radiologists were − 0.09, 2.03, and 2.71, respectively. There was no systematic trend between overall radiographic disease severity (based on total Brasfield score) and model performance; based on MD, the model performed best for intermediate disease and worst for the most severe disease (Supplementary Table S3). For subscores, the MD, MAD, and RMSE of the model's estimates compared to the average subscores of all 5 radiologists were in the range of −0.19-0.24, 0.40–0.86, and 0.56–1.56, respectively.

Based on MD (closest to 0), the model when compared pairwise to each radiologist vs. the average of the remaining radiologists performed statistically at least as well as 2/5 readers for air-trapping, 3/5 readers for linear markings, 5/5 for nodular cystic lesions, 3/5 readers for large lesions, 5/5 readers for general severity, and 5/5 readers for total score. Based on MAD (closest to 0), the model when compared pairwise to each radiologist performed statistically at least as well as 4/5 radiologists for air-trapping, 5/5 radiologists for linear markings, 5/5 radiologists for nodular cystic lesions, 0/5 radiologists for large lesions, 5/5 radiologists for general severity, and 3/5 radiologists for total score. Based on RMSE (closest to 0), the model when compared pairwise to each radiologist performed statistically at least as well as 5/5 radiologists for air-trapping, 5/5 radiologists for linear markings, 5/5 radiologists for nodular cystic lesions, 0/5 radiologists for large lesions, 5/5 radiologists for general severity, and 2/5 radiologists for total score.

Heat maps generated for the test set cases demonstrated visual correlation between salient features identified by the model and the severity of CF lung disease according to radiographic subscore. Examples are shown in Fig. 2.

## 4. Discussion

In this study, we have found that our DCNN model for estimating Brasfield scores in patients with cystic fibrosis achieved correlations of 0.79–0.83 for the total Brasfield score when compared to 5 pediatric radiologists, who achieved correlations of 0.85–0.90. The mean difference in total score between the model and the average of the 5 radiologists was −0.09. In addition, based on correlation analysis as well as MD, MAD, and RMSE estimates, the model at least matched radiologist performance for a majority of subfeatures (linear markings, nodular cystic lesions, general severity). Thus, we have demonstrated the feasibility and potential of a DCNN model for predicting Brasfield total and subscores at a level approaching that of pediatric radiologists.

The model's deficits in predicting the total Brasfield score were likely primarily driven by its weak performance for scoring large lesions. Model correlation for large lesions subscores were 0.29–0.38 vs. 0.74–0.84 for radiologists, and the majority of radiologists outperformed the model based on MD, MAD, and RMSE. Unlike other Brasfield features, "large lesions" comprise a relatively heterogeneous group of potential abnormalities. Furthermore, while other Brasfield features (e.g., linear markings and nodular cystic lesions) tend to increase in severity concurrently with progressive CF lung disease, "large lesions" are more sporadic and asyn-

**Table 3**
Radiologist v. model mean difference in brasfield score estimates.

| Group comparison | Air-trapping | Linear markings | Nodular cystic lesions | Large lesions | General severity | Total score |
|---|---|---|---|---|---|---|
| Model v. All Rad Avg | 0.24 | 0.15 | −0.10 | −0.19 | 0.00 | −0.09 |
| Model v. Rad 2–5 Avg | 0.33 | 0.19 | **−0.04** | **−0.16** | 0.02 | **−0.33** |
| Rad 1 v. Rad 2–5 Avg | 0.43 | 0.21 | 0.33 | 0.17 | 0.08 | −1.21 |
| | p = 0.16 | p = 0.83 | p < 0.001† | p = 0.006† | p = 0.89 | p = 0.01† |
| Model v. Rad 1,3–5 Avg | 0.26 | 0.16 | **0.00** | −0.20 | 0.02 | −0.24 |
| Rad 2 v. Rad 1,3–5 Avg | **0.11** | **0.06** | 0.52 | −0.03 | 0.10 | −0.74 |
| | p = 0.006† | p = 0.039† | p < 0.001† | p = 0.18 | p = 0.84 | p = 0.28 |
| Model v. Rad 1–2,4–5 Avg | 0.26 | 0.14 | −0.12 | −0.20 | −0.01 | **−0.07** |
| Rad 3 v. Rad 1–2,4–5 Avg | **0.11** | **−0.06** | −0.07 | **−0.04** | −0.05 | 0.12 |
| | p = 0.003† | p = 0.002† | p = 0.27 | p < 0.001† | p = 0.11 | p = 0.03† |
| Model v. Rad 1–3,5 Avg | 0.24 | 0.16 | **−0.22** | −0.16 | **0.06** | −0.08 |
| Rad 4 v. Rad 1–3,5 Avg | **−0.02** | 0.04 | −0.57 | **0.15** | 0.30 | 0.06 |
| | p < 0.001† | p = 0.40 | p < 0.001† | p = 0.02† | p < 0.001† | p = 0.09 |
| Model v. Rad 1–4 Avg | **0.12** | **0.10** | −0.15 | −0.24 | **−0.09** | **0.27** |
| Rad 5 v. Rad 1–4 Avg | −0.63 | −0.26 | −0.20 | −0.25 | −0.43 | 1.78 |
| | p < 0.001† | p < 0.001† | p = 0.38 | p = 0.97 | p < 0.001† | p < 0.001† |

Abbreviations: Avg = average (mean); Rad = radiologist.
Note: Values with least dispersion from the mean (closest to 0) achieving a statistically significant difference from the comparator in the same table row and column using the Mann- Whitney *U* test are bolded.
† $p \leq 0.05$.

chronous, potentially reflect more acute disease that could improve or even disappear with treatment [31–33]. Moreover, as opposed to other Brasfield categories in which severity increases are gradual and separated by only 1 point, large lesions can only be scored 0, 3, or 5; thus, any "errors" in large lesion classification appear more pronounced in assessing the model's overall performance. Finally, the interrater reproducibility of large lesions scores reported in the literature is substantially lower than that achieved by the pediatric radiologists in our study, with correlations in the range of 0.49–0.52 [4]. The better-than-average reproducibility of large

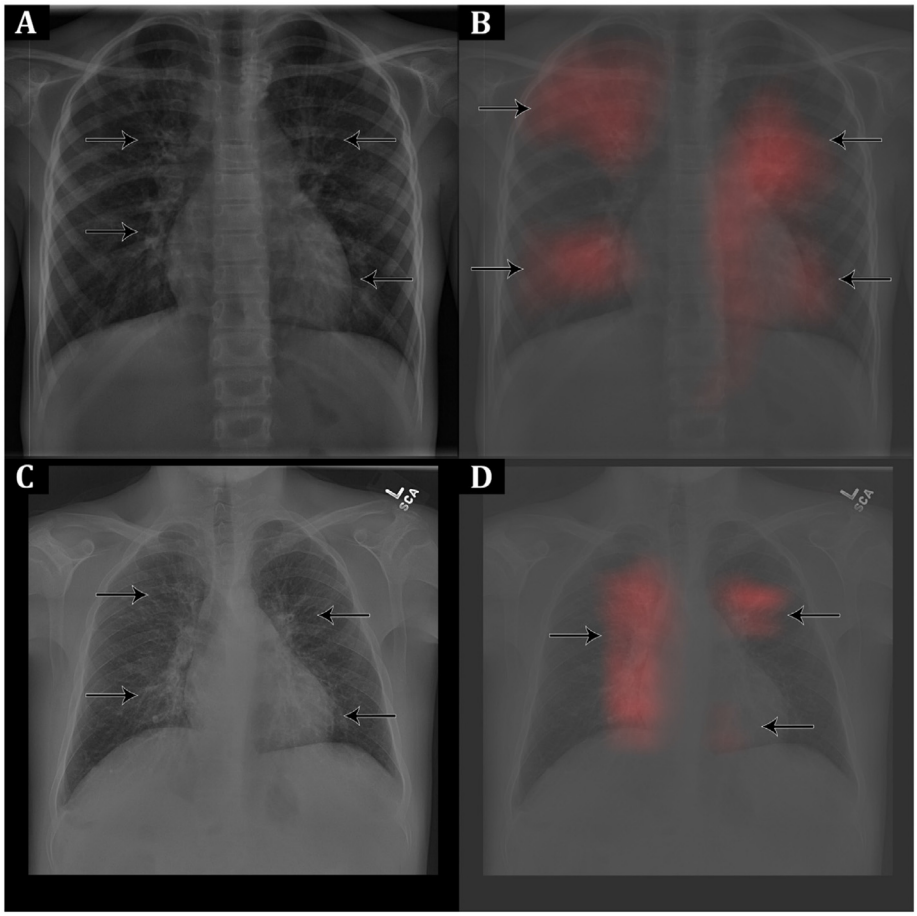


**Fig. 2.** Example class activation mappings (CAMs) generated by the DCNN model using chest radiographs obtained in cystic fibrosis (CF) patients from the test set cohort. Pixels of higher intensity indicate features of greater salience to the model in determining its output. (A) Frontal chest radiograph in a 6-year-old female with CF shows bilateral peribronchial airway thickening (arrows). (B) Corresponding CAM generated by the model for discerning Brasfield "linear markings" shows high pixel intensity in most of the same areas of abnormality (arrows). (C) Frontal chest radiograph in a 14-year-old male with CF shows bilateral bronchiectasis (arrows). (D) Corresponding CAM generated by the model for discerning Brasfield "nodular cystic lesions" shows high pixel intensity in most of the same areas of abnormality (arrows).

lesions scores achieved by the radiologists in our study may have been facilitated by adherence to an image-rich reference standard document, which is not generally available to Brasfield scorers.

In the end, the model output depends on the variety, quantity, and quality of the data on which it was trained. Thus, we might similarly expect it to perform less well when faced with other uncommon image patterns, such as severe nodular cystic lesions in the absence of linear markings. Indeed, the model overall performed worst relative to the mean radiologist for the most severe radiographic disease (Supplementary Table S3), for which training examples were most scarce. Moreover, such severe disease (total score $< 11$) would typically require the presence of large lesions, for which the model underperformed, as noted above.

In the end, the model output depends on the variety, quantity, and quality of the data on which it was trained. Thus, we might similarly expect it to perform less well when faced with other uncommon image patterns, such as severe nodular cystic lesions in the absence of linear markings. Indeed, the model overall performed worst relative to the mean radiologist for the most severe radiographic disease (Supplementary Table S3), for which training examples were most scarce. Moreover, such severe disease (total score $< 11$) would typically require the presence of large lesions, for which the model underperformed, as noted above.

As such, the DCNN model could likely be further enhanced with access to a larger training dataset. It is well understood that DCNN model accuracy generally increases with more training data, although the marginal performance returns decrease as the quantity of training data grows larger [11,21]. However, acquiring training data is challenging in this context given the relative rarity of CF with only a finite number of CXRs and laborious nature of image evaluation on a large scale. The use of a carefully devised reference standard to improve the consistency of scoring likely assisted in maximizing the DCNN model performance despite a smaller training set. Ultimately, the availability of multicenter data, encompassing a broad range of patient disease states and demographics as well as radiographic techniques, might best ensure both the accuracy and generalizability of the model. For example, recent studies have shown that DCNNs developed for classifying CXRs and trained on local data perform less well when tested on external rather than internal data sources [34,35].

Our study has several limitations. First, as alluded to above, we recognize that the performance of our DCNN model, based on retrospective evaluation of a single institution's limited number of CF CXRs, including some in non-unique patients, may not necessarily be generalizable to other CF cohorts. Although our study included a range of Brasfield subscores, larger scale, multi-institutional data including a broad spectrum of CF patients as well as imaging equipment with variable radiographic parameters, would likely assist in improving the performance and more widespread applicability of the model, which could then be validated in prospective fashion. However, such coordinated efforts pose multiple challenges ranging from privacy concerns in image sharing to workload considerations involved in image scoring and are beyond the scope of this initial work.

Second, the model was trained based on a single pediatric radiologist's scores and then assessed based on a group of 5 pediatric radiologists. As there are no absolute "ground truth" Brasfield scores, we instead assessed the model against human readers with specialty expertise (in this case, pediatric radiologists), an approach consistent with prior studies on DCNNs for medical imaging [11,19,20]. Due to the arduous nature of scoring training data, only a single radiologist was assigned to this task. Nonetheless, high rates of radiologist interrater reliability were maintained, likely through the rigorous creation and application of the Brasfield reference standard document. Third, neither the DCNN model nor the radiologists scored images with reference to comparison exams or reported prior Brasfield scores, if available, in accordance with the design of the Brasfield system based only on image features [4]. However, we recognize that in clinical practice, current scores may be influenced by prior images or scores in order to convey changes in radiographic disease; the ability of the model to predict change is an area for potential future investigation. Finally, we stress that the DCNN model is only applicable to the generation of Brasfield scores. While other quantitative CF scoring systems exist, DCNNs models would have to be specifically constructed to produce the appropriate score metrics, although overlap between systems might facilitate some synergies using existing training data.

Many further steps are prudent prior to the implementation of the Brasfield DCNN in clinical practice. First, as alluded to above, retraining the model using more data supplied by multiple CF centers with patients of varying disease severity and age and CXRs performed with variable radiographic equipment and techniques might improve the generalizability of the model's performance. If multiple readers were involved in scoring training data, adherence to a reference standard would likely prove beneficial in ensuring scoring consistency. Once retrained, the model could similarly be retested on a diverse, multicenter CF CXR cohort. After additional refinement, the model could be tested in a clinical setting by routing (either manually or automatically) newly obtained digital CXR images in CF patients to a server hosting the pretrained DCNN. The new CXR images would be evaluated in real-time, and the Brasfield scores could be prepopulated into the radiology report, for review by the radiologist either before or after the scores had been manually derived. Associated CAMs could also be made available. The model output would be compared with the radiologist's manually derived scores, and the concordance rates determined. Additionally, the potential clinical benefits of auto-generating Brasfield scores could be measured, such as faster exam turnaround time. After validation of the refined DCNN's performance and benefits in a clinical setting, the model could be distributed via locally installed software or an external website/server. In addition, regulatory approval as a new technological device could be pursued. While these future project phases would likely be costly, once implemented, the incremental cost of evaluating a new CXR (e.g., via uploading images to a secure website) would likely be much less, although continued system maintenance would be necessary.

In conclusion, we have successfully developed and demonstrated the potential of a DCNN model for automating Brasfield scoring. Within the limitations of this study, the model's performance approaches that of pediatric radiologists. Further refinement and multicenter validation followed by distribution of such software could enable more accessible, rapid, accurate, and consistent radiographic quantification of CF lung involvement, in turn enhancing the precision of disease monitoring.

The following are the supplementary data related to this article.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcf.2019.04.016.

## Contributors

E.J.Z., Z.A.B., and D.B.L. contributed to the design of the study and literature research. All authors were involved in data collection. E.J.Z., Z.A.B., and D.B.L. were involved in data analysis and interpretation and manuscript drafting. All authors critically revised the article for important intellectual content and approved the final version for submission. E.J.Z. is the guarantor of integrity of the entire study.

## Declaration of interests

E.J.Z., Z.A.B., Y.S., J.M.S., and S.S.H have no relevant relationships. M.P.L. has received grants from Philips and the Stanford Child Health Research Institute, not related to the present article.

D.B.L. has grants pending or received to his institution by Philips and Siemens, not related to the present article.

## Funding

This work was supported by an internal grant from Stanford University, Department of Radiology, Division of Pediatric Radiology (grant 1204154–100-JHAIJ).

## Conflict of interest statement

The authors have no conflict of interest to declare relevant to this study. M.P.L. reports grants from Philips and the Stanford Child Health Research Institute, outside the submitted work. D.B.L. reports grants pending or received from Philips and Siemens, outside the submitted work. The remaining authors have nothing to disclose.

## Acknowledgements

None.

## References

[1] Kerem E, Conway S, Elborn S, Heijerman H, Consensus Committee. Standards of care for patients with cystic fibrosis: a European consensus. J Cyst Fibros 2005;4(1):7–26.
[2] Pittman JE, Ferkol TW. The evolution of cystic fibrosis care. Chest 2015;148(2):533–42.
[3] Loeve M, Krestin GP, Rosenfeld M, de Bruijne M, Stick SM, Tiddens HA. Chest computed tomography: a validated surrogate endpoint of cystic fibrosis lung disease? Eur Respir J 2013;42(3):844–57.
[4] Brasfield D, Hicks G, Soong S, Peters J, Tiller R. Evaluation of scoring system of the chest radiograph in cystic fibrosis: a collaborative study. AJR Am J Roentgenol 1980;134(6):1195–8.
[5] Cleveland RH, Neish AS, Zurakowski D, Nichols DP, Wohl ME, Colin AA. Cystic fibrosis: predictors of accelerated decline and distribution of disease in 230 patients. AJR Am J Roentgenol 1998;171(5):1311–15.
[6] Cleveland RH, Stamoulis C, Sawicki G, et al. Brasfield and Wisconsin scoring systems have equal value as outcome assessment tools of cystic fibrosis lung disease. Pediatr Radiol 2014;44(5):529–34.
[7] Cleveland RH, Sawicki GS, Stamoulis C. Similar performance of Brasfield and Wisconsin scoring systems in young children with cystic fibrosis. Pediatr Radiol 2015;45(11):16241628.
[8] Sanders DB, Li Z, Brody AS, Farrell PM. Chest computed tomography scores of severity are associated with future lung disease progression in children with cystic fibrosis. Am J Respir Crit Care Med 2011;184(7):816–21.
[9] Farrell PM, Li Z, Kosorok MR, et al. Longitudinal evaluation of bronchopulmonary disease in children with cystic fibrosis. Pediatr Pulmonol 2003;36(3):230–40.
[10] Sanders DB, Li Z, Rock MJ, Brody AS, Farrell PM. The sensitivity of lung disease surrogates in detecting chest CT abnormalities in children with cystic fibrosis. Pediatr Pulmonol 2012;47(6):567–73.
[11] Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287(1):313–22.
[12] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018. doi:10.1007/s13244-018-0639-9. Epub ahead of print.
[13] Rezaeilouyeh H, Mollahosseini A, Mahoor MH. Microscopic medical image classification framework via deep learning and shearlet transform. J Med Imaging (Bellingham) 2016;3(4):044501.
[14] Cheng JZ, Ni D, Chou YH, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2016;6:24454.
[15] Kunapuli G, Varghese BA, Ganapathy P, et al. A decision-support tool for renal mass classification. J Digit Imaging 2018;31(6):929–39.
[16] Wang C, Elazab A, Jia F, Wu J, Hu Q. Automated chest screening based on a hybrid model of transfer learning and convolutional sparse denoising autoencoder. Biomed Eng Online 2018;17(1):63.
[17] Choy G, Khalilzadeh O, Michalski M, et al. Current applications and future impact of machine learning in radiology. Radiology 2018;288(2):318–28.
[18] Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology 2018. doi:10.1148/radiol.2018180887. Epub ahead of print.
[19] Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018;15(11):e1002699.
[20] Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15(11):e1002686.
[21] Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 2018. doi:10.1148/radiol.2018181422. Epub ahead of print.
[22] Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A. Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiol Artif Intell 2019;1(1):e180015.
[23] Pisano ED, Zong S, Hemminger BM, et al. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. J Digit Imaging 1998;11(4):193–200.
[24] Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505–15.
[25] Yamashita R, Nishio M, Do RK, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9(4):611–29.
[26] Deng J, Dong W, Socher R, Li L, Li K, Imagenet Fei-Fei L. A large-scale hierarchical image database. In: Proceedings of the 12th IEEE conference on computer vision and pattern recognition; 2009. p. 248–55. Miami, Florida, June 20–25.
[27] Shin HC, 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(5):1285–98.
[28] Kingma D, Ba J. Adam: a method for stochastic optimization. Presented at the International Conference on Representational Learning, Banff, Canada, April 14–16, 2014. arXiv preprint July 23, 2015: arXiv:1412.6980v8. https://arxiv.org/abs/1412.6980. Published December 22, 2014. Accessed January 6, 2019.
[29] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates; 2012. p. 1097–105. Lake Tahoe, Nevada, December 3–6, 2012.
[30] Philbrick KA, Yoshida K, Inoue D, et al. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. AJR Am J Roentgenol 2018;211(6):1184–93.
[31] Wielpütz MO, Eichinger M, Biederer J, et al. Imaging of cystic fibrosis lung disease and clinical interpretation. Rofo 2016;188(9):834–45.
[32] Peroni DG, Boner AL. Atelectasis: mechanisms, diagnosis and management. Paediatr Respir Rev 2000;1(3):274–8.
[33] Schäfer J, Griese M, Chandrasekaran R, Chotirmall SH, Hartl D. Pathogenesis, imaging and clinical characteristics of CF and non-CF bronchiectasis. BMC Pulm Med 2018;18(1):79.
[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(11):e1002683.
[35] Pan I, Agarwal S, Merck D. Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks. J Digit Imaging 2019. doi:10.1007/s10278-019-00180-9. Epub ahead of print.