Radiology

Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks[1]

Paras Lakhani, MD
Baskaran Sundaram, MD

Purpose: To evaluate the efficacy of deep convolutional neural networks (DCNNs) for detecting tuberculosis (TB) on chest radiographs.

Materials and Methods: Four deidentified HIPAA-compliant datasets were used in this study that were exempted from review by the institutional review board, which consisted of 1007 posteroanterior chest radiographs. The datasets were split into training (68.0%), validation (17.1%), and test (14.9%). Two different DCNNs, AlexNet and GoogLeNet, were used to classify the images as having manifestations of pulmonary TB or as healthy. Both untrained and pretrained networks on ImageNet were used, and augmentation with multiple preprocessing techniques. Ensembles were performed on the best-performing algorithms. For cases where the classifiers were in disagreement, an independent board-certified cardiothoracic radiologist blindly interpreted the images to evaluate a potential radiologist-augmented workflow. Receiver operating characteristic curves and areas under the curve (AUCs) were used to assess model performance by using the DeLong method for statistical comparison of receiver operating characteristic curves.

Results: The best-performing classifier had an AUC of 0.99, which was an ensemble of the AlexNet and GoogLeNet DCNNs. The AUCs of the pretrained models were greater than that of the untrained models ($P < .001$). Augmenting the dataset further increased accuracy (P values for AlexNet and GoogLeNet were .03 and .02, respectively). The DCNNs had disagreement in 13 of the 150 test cases, which were blindly reviewed by a cardiothoracic radiologist, who correctly interpreted all 13 cases (100%). This radiologist-augmented approach resulted in a sensitivity of 97.3% and specificity 100%.

Conclusion: Deep learning with DCNNs can accurately classify TB at chest radiography with an AUC of 0.99. A radiologist-augmented approach for cases where there was disagreement among the classifiers further improved accuracy.

[1] From the Department of Radiology, Thomas Jefferson University Hospital, Sidney Kimmel Jefferson Medical College, 132 S 10th St, Room 1080A, Main Building, Philadelphia, PA 19107-5244. Received October 5, 2016; revision requested November 23; revision received December 12; accepted January 9, 2017; final version accepted January 19. **Address correspondence to** P.L. (e-mail: *paras.lakhani@jefferson.edu*).

Tuberculosis (TB) is an infectious disease caused by the bacillus *Mycobacterium tuberculosis*. TB is a leading cause of death by infectious disease worldwide, alongside human immunodeficiency virus–acquired immune deficiency syndrome (known as HIV-AIDS) (1). In 2014, approximately 9600000 people developed clinical TB, resulting in 1500000 deaths (1).

While indiscriminate mass screening for TB should be avoided, the World Health Organization recommends broader use of screening by chest radiography and rapid molecular diagnostics for selected high-risk groups (1). Posteroanterior chest radiography is an important part of many algorithms for worldwide screening of TB (1,2). In addition, imaging also plays the central role in the work-up of patients suspected of having pulmonary TB (2,3).

It has been reported (4,5) that there is a relative lack of radiology interpretation expertise in many TB-prevalent locations, which may impair screening efficacy and work-up efforts. An efficacious automated and cost-effective method could aid screening evaluation efforts in developing nations and facilitate earlier detection of disease. Therefore, there has been an interest in the use of computer-aided diagnosis for detection of pulmonary TB at chest radiography, with multiple approaches proposed (4,6,7).

Commercially available software (CAD4TB; Image Analysis Group, Nijmegen, the Netherlands) had an area under the curve (AUC) that ranged from 0.71 to 0.84 in five studies, according to one review (8). The software is based on machine-learning approaches and uses a combination of textural abnormality and shape detection (9). Another computer-aided diagnosis study for detection of pulmonary TB at chest radiography used lung segmentation, texture and shape feature extraction, and classification with support vector machines to achieve an AUC of 0.87–0.90 (10).

Currently, deep learning techniques are considered to be state of the art for classification of images, which arises from the recent success in the ImageNet Large Scale Visual Recognition Competition (11). Since 2012, all winning entries used deep convolutional neural networks (DCNN), a type of deep learning approach well suited for analyzing images. This resulted in a decrease in the classification error rate from approximately 25% in 2011 to 3.6% in 2015 (11,12). Convolutional neural networks have been around for some time; for example, in 1998 LeCun et al (13) used them to classify handwritten digits. However, it was only until relatively recently that such networks could be applied to everyday images because of the tremendous parallel processing power required, which became possible with modern graphics processing unit technology.

There is interest in applying deep learning in radiology because of the recent success, and with promising results. Some examples include detection of pleural effusion and cardiomegaly at chest radiography (14), mediastinal lymph nodes at computed tomography (CT) (15), lung nodules at CT (16), and pancreatic (17) and brain segmentation (18). Regarding classification of TB at chest radiography, one group recently used DCNN to achieve an AUC of 0.88–96 on three different datasets (19).

In this study, we evaluate the efficacy of DCNN for detection of TB on chest radiographs.

Advances in Knowledge

- Deep learning with convolutional neural networks can accurately classify tuberculosis (TB) at chest radiography with an area under the curve of 0.99.
- Pretrained neural networks ($P <$ .001) and augmented datasets (P = .02 and P = .03) resulted in greater accuracy.
- The most accurate model incorporated a radiologist overread when the machines were discrepant, which had a net sensitivity of 97.3% and a specificity of 100%.

Implication for Patient Care

- Automated detection of pulmonary TB at chest radiography may facilitate screening and evaluation efforts in TB-prevalent areas with limited access to radiologists.

Materials and Methods

Datasets

All datasets were deidentified and compliant with the Health Insurance Portability and Accountability Act. The Belarus and Thomas Jefferson University datasets were exempted from institutional review board review at Thomas Jefferson University Hospital. The National Institutes of Health datasets were exempted from review by the institutional review board (No. 5357) by the National Institutes of Health Office of Human Research Protection Programs. This was a retrospective study that involved four datasets (Table 1). This includes two publicly available datasets maintained by the National Institutes of Health, which are from Montgomery County, Maryland, and Shenzhen, China (20). The other two datasets are from Thomas Jefferson University Hospital, Philadelphia, and the Belarus Tuberculosis Portal maintained by the Belarus TB public health program (21). For the Thomas Jefferson University and Belarus datasets, the positive

Published online before print
10.1148/radiol.2017162326 **Content code:** CH

Radiology 2017; 000:1–9

Abbreviations:
AUC = area under the curve
DCNN = deep convolutional neural network
TB = tuberculosis

Author contributions:
Guarantor of integrity of entire study, P.L.; study concepts/study design or data acquisition or data analysis/interpretation, P.L., B.S.; manuscript drafting or manuscript revision for important intellectual content, P.L., B.S.; approval of final version of submitted manuscript, P.L., B.S.; agrees to ensure any questions related to the work are appropriately resolved, P.L., B.S.; literature research, P.L., B.S.; clinical studies, P.L., B.S.; experimental studies, P.L.; statistical analysis, P.L.; and manuscript editing, P.L., B.S.

Conflicts of interest are listed at the end of this article.

Table 1

TB Chest X-Ray Datasets

Dataset Origin	No. of Cases Positive for TB	No. of Healthy Control Patients	File Type	Bit Depth	CR/DR	Resolution	Average Age (y)	Men (%)	Positive Cases with Pleural Effusion (%)	Positive Cases with Miliary TB (%)	Positive Cases with Cavitation (%)
Montgomery County, Md	58	80	PNG	8 bit	CR	4020 × 4892	33.1 ± 18.1	44.2	20.7 (12/58)	3.5 (2/58)	19.0 (11/58)
Shenzhen, China	336	326	PNG	8 bit	DR	948–3001 × 1130–3001	33.4 ± 14.0	66.4	6.5 (22/336)	0.9 (3/336)	10.1 (34/336)
Belarus TB Public Health Program	88	0	DICOM	16 bit	DR	2248 × 2248–2724	52.7 ± 9.5	76.1	11.4 (10/88)	4.6 (4/88)	18.2 (16/88)
Thomas Jefferson University Hospital	10	109	DICOM	16 bit	CR + DR	1994–3280 × 2428–4248	53.1 ± 16.9	47.9	10 (1/10)	0 (0/10)	40 (4/10)

Note.—Data in parentheses are numerator and denominator. There were a total of 1007 cases (492 cases positive for TB and 515 healthy control patients). "Positive cases" refer to cases that were positive for TB. CR = computed radiography, DICOM = Digital Communications in Medicine, DR = digital radiography, PNG = Portable Network Graphics.

cases with radiologic manifestations of pulmonary TB were confirmed with pathologic findings of sputum, original authors of the radiology reports, and an independent radiologist (P.L., with 10 years of experience). For the Thomas Jefferson University Dataset, the healthy control patients were established from the original authors of the radiology reports and an independent radiologist (P.L.). For the National Institutes of Health datasets, patients who were positive for TB and healthy control patients were established from clinical records and expert readers. For the Belarus dataset, the first 88 consecutive cases (of 420 in the portal) were downloaded for patients who underwent posteroanterior chest radiography at the time of initial diagnosis and pathologic analysis. Because the Belarus dataset consisted of patients who were positive for TB, a similar number of healthy control patients were obtained from Thomas Jefferson University Hospital so that the cumulative total of all datasets would have a similar number of patients who were positive for TB and healthy patients (Table 1). The dataset from China included a minority of pediatric images (21 pediatric, 641 adults) so the image sizes had a larger range (Table 1). The patient demographics for the datasets and additional pertinent findings such as pleural effusion, military pattern of disease, and presence of cavitation for positive cases are also provided in Table 1.

Methods

The chest radiographic images were resized to a 256 × 256 matrix and converted into Portable Network Graphics format. The images were loaded onto a computer with a Linux operating system (Ubuntu 14.04; Canonical, London, England) and with the Caffe deep learning framework (*http://caffe.berkeleyvision.org;* BVLC, Berkeley, Calif) (22), with CUDA 7.5/cuDNN 5.0 (Nvidia Corporation, Santa Clara, Calif) dependencies for graphics processing unit acceleration. The computer contained an Intel i5 3570k 3.4-gHz processor (Intel, Santa Clara, Calif), 4 TB of hard disk space, 32 GB of RAM, and a CUDA-enabled Nvidia Titan ×12 GB graphics processing unit (Nvidia).

Two different deep convolutional neural network architectures were evaluated in this study, AlexNet (23) and GoogLeNet (24), including pretrained and untrained models. Pretrained networks were already trained on 1.2 million everyday color images from ImageNet (*http://www.image-net.org/*) that consisted of 1000 categories before learning from the chest radiographs in this study (referred to as pretrained). Untrained networks were not trained before they were used (referred to as untrained). This included AlexNet untrained (Alex-Net-U), AlexNet pretrained (AlexNet-T), GoogLeNet untrained (GoogLeNet-U), and GoogLeNet pretrained (GoogLeNet-T). Pretrained networks were obtained from the Caffe Model Zoo, an open-access repository of pretrained models for use with Caffe. The following solver parameters were used for training: 120 epochs; base learning rate for untrained models and for pretrained models, 0.01 and 0.001, respectively; stochastic gradient descent; step-down, 33%; and γ, 0.1.

All images were augmented by using random cropping of 227 × 227 pixels, mean subtraction, and mirror images, which were prebuilt options within the Caffe framework. Further augmentation was performed in training some of the DCNNs, including rotations of 90°, 180°, and 270°, and Contrast Limited Adaptive Histogram Equalization processing by using ImageJ v. 1.50i (NIH, Bethesda, Md) (25). The DCNNs that used this additional augmentation are labeled AlexNet-TA and GoogLeNet-TA when pretrained on ImageNet, and AlexNet-UA and GoogLeNet-UA when untrained.

Of the 1007 patients in the total dataset (Table 1), 150 random patients (14.9%) were selected for testing. Randomization was performed by using pseudorandom numbers generated from the random function in the Python Standard Library (Python 2.7.13, Python Software Foundation, Wilmington, Del). Of these 150 test patients, 75 were positive for TB and 75 were healthy. Among the remaining 857 patients, they were

randomly split into an 80%:20% ratio into training (685 patients) and validation (172 patients). The training set was used to train the algorithm, the validation set was for model selection, and the test set was for assessment of the final chosen model. In deciding the percent split, the goal is to keep enough data for the algorithms to train from but have enough validation and test cases to maintain a reasonable confidence interval of the accuracy of the model (26).

The 75 test patients positive for TB were analyzed by a cardiothoracic radiologist (P.L.) for degree of pulmonary parenchymal involvement by TB and placed into one of the following three categories: subtle (pulmonary parenchymal involvement, <4%), intermediate (pulmonary parenchymal involvement, 4%–8%), and readily apparent (pulmonary parenchymal involvement, >8%) (Table 2). To determine this, the right and left lungs were divided into three zones (upper, middle, and lower). Opacities that occupied half or more of one zone were considered readily apparent. Opacities occupying a fourth to half of a zone were considered intermediate. Opacities occupying less than a fourth of a zone were considered subtle.

Statistical and Data Analysis

All statistical analyses were performed by using software (MedCalc v. 16.8; MedCalc Software, Ostend, Belgium). On the test datasets, receiver operating characteristic curves and AUCs were determined (27). Contingency tables, accuracy, sensitivity, and specificity were determined from the optimal threshold by the Youden index, which is the following equation: [1 − (false-positive rate + false-negative rate)]. For the receiver operating characteristic curves, standard error, 95% confidence intervals, and comparisons between AUCs were made by using a nonparametric approach (28–31). The adjusted Wald method was used to determine 95% confidence intervals on the accuracy, sensitivity, and specificity from the contingency tables (32). *P* values less than .05 were considered to indicate statistical significance.

Ensembles were performed by taking different weighted averages of the probability scores generated by the classifiers (AlexNet and GoogLeNet). This ranged from using equal weighting (50% AlexNet and 50% GoogLeNet) to up to 10-fold weighting biased toward either classifier. Receiver operating characteristic curves, AUC, and optimal sensitivity and specificity values were then determined for various ensemble approaches.

For cases where the AlexNet and GoogLeNet classifiers had disagreement, an independent board-certified cardiothoracic radiologist (B.S., with 18 years of experience) blindly interpreted the images as either having manifestations of TB or as normal. Contingency tables and sensitivity and specificity values were then created from these results (Fig 1).

Results

A summary of the results is provided in Table 3. For both deep neural networks, the AUCs of the pretrained models (AlexNet-T, GoogLeNet-T) were greater than that of the untrained models (AlexNet-U, GoogLeNet-U) ($P < .001$). In addition, augmentation of the dataset with additional transformations,

Table 2

Distribution of Test Cases Positive for TB

Degree of Conspicuity	Cases (%)
Subtle (<4% pulmonary parenchymal involvement)	33.3 (25/75)
Intermediate (4%–8% pulmonary parenchymal involvement)	37.3 (28/75)
Readily apparent (>8% pulmonary parenchymal involvement)	29.3 (22/75)

Note.—Data in parentheses are numerator and denominator.

Figure 1

A

GoogLeNet-TA

	Diagnosis +	Diagnosis –	Total
Test +	69	1	70
Test –	6	74	80
Total	75	75	150

B

AlexNet-TA

	Diagnosis +	Diagnosis –	Total
Test +	69	4	73
Test –	6	71	77
Total	75	75	150

C

Ensemble

	Diagnosis +	Diagnosis –	Total
Test +	73	4	77
Test –	2	71	73
Total	75	75	150

D

Radiologist Augmented

	Diagnosis +	Diagnosis –	Total
Test +	73	0	73
Test –	2	75	77
Total	75	75	150

Figure 1: Contingency tables. *A,* Sensitivity, 92.0% (95% confidence interval: 83.3%, 96.6%); specificity, 98.7% (95% confidence interval: 92.1%, 100%); accuracy, 95.3% (95% confidence interval: 90.5%, 97.9%). *B,* Sensitivity, 92.0% (95% confidence interval: 83.3%, 96.6%); specificity, 94.7% (95% confidence interval: 86.7%, 98.3%); accuracy, 93.3% (95% confidence interval: 88.0%, 96.5%). *C,* Sensitivity, 97.3% (95% confidence interval: 90.2%, 99.8%); specificity, 94.7% (95% confidence interval: 86.7%, 98.3%); accuracy, 96.0% (95% confidence interval: 91.4%, 98.3%). *D,* Sensitivity, 97.3% (95% confidence interval: 90.2%, 99.8%); specificity, 100% (95% confidence interval: 95.8%, 100%); accuracy, 98.7% (95% confidence interval: 95.0%, 99.9%).

Table 3

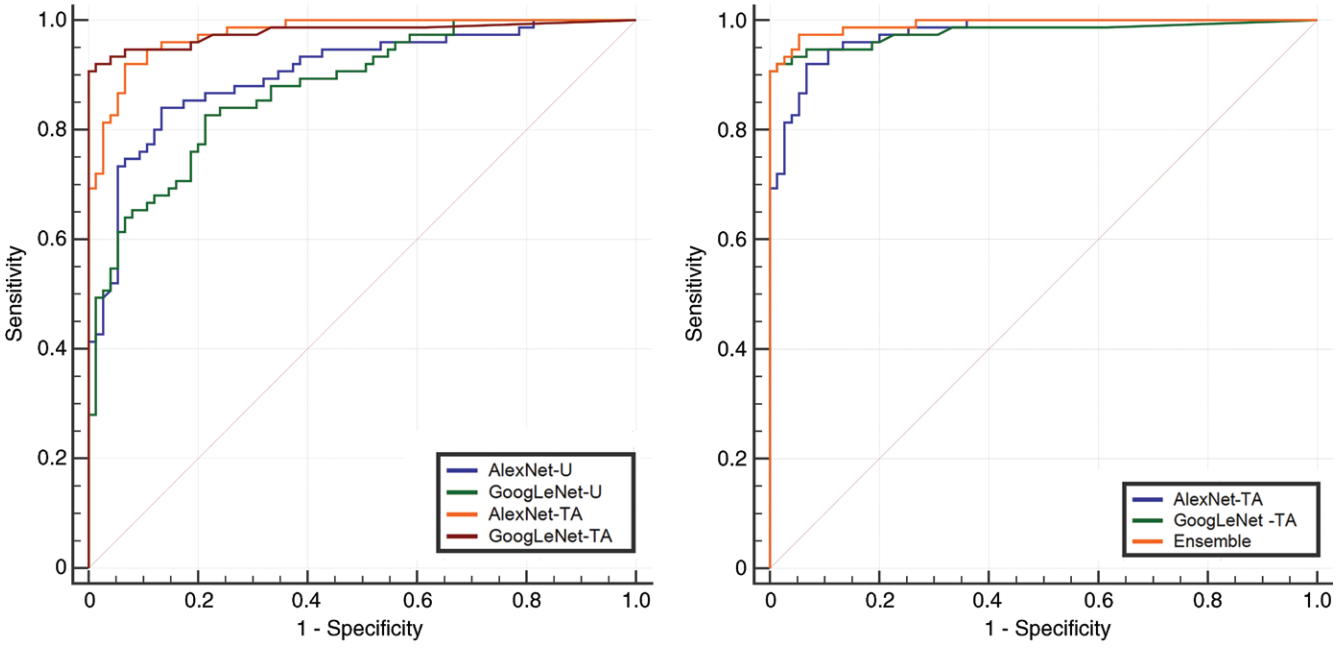
AUC Test Dataset

Parameter	Untrained	Pretrained	Untrained with Augmentation*	Pretrained with Augmentation*
AlexNet	0.90 (0.84, 0.95)	0.98 (0.95, 1.00)	0.95 (0.90, 0.98)	0.98 (0.94, 0.99)
GoogLeNet	0.88 (0.81, 0.92)	0.97 (0.93, 0.99)	0.94 (0.89, 0.97)	0.98 (0.94, 1.00)
Ensemble				0.99 (0.96, 1.00)

Note.—Data in parentheses are 95% confidence interval.

* Additional augmentation of 90, 180, 270 rotations, and Contrast Limited Adaptive Histogram Equalization processing.

Figure 2



a. b.

Figure 2: **(a)** Comparison of receiver operating characteristic curves for the untrained AlexNet-U and GoogLeNet-U models and pretrained with augmentation AlexNet-TA and GoogLeNet-TA models. The receiver operating characteristic curves for the AlexNet-TA and GoogLeNet-TA models had an AUC that was significantly greater than that for the untrained AlexNet-U and GoogLeNet-U models ($P < .001$) (Table 3). **(b)** Comparison of receiver operating characteristic curves for the AlexNet-TA, GoogLeNet-TA, and ensemble of the two models. The ensemble provided the best AUC (Table 3).

such as rotations and Contrast Limited Adaptive Histogram Equalization, further increased accuracy for both neural networks (AlexNet-UA, GoogLeNet-UA) over untrained models (AlexNet-U, GoogLeNet-U) ($P = .03$ for AlexNet and $P = .02$ for GoogLeNet). The best-performing ensemble model had an AUC of 0.99, which was significantly greater than that of the untrained AlexNet-U and GoogLeNet-U models, which had AUCs of 0.90 and 0.88, respectively ($P < .001$).

A comparison of receiver operating characteristic curves for the untrained and pretrained augmented models for AlexNet and GoogLeNet, as well as ensemble approaches, are provided in Figure 2.

The contingency tables for the best-performing models, including GoogLeNet-TA, AlexNet-TA, ensemble of AlexNet-TA, and GoogLeNet-TA are provided in Figure 1. The sensitivity of AlexNet-TA was 92.0% and the specificity was 94.7%. The sensitivity of GoogLeNet-TA was 92.0% and the specificity was 98.7%. The sensitivity of the ensemble was 97.3% and the specificity was 94.7%.

The distribution of the conspicuity of the 75 test patients who were positive for TB is provided in Table 2.

Radiologist-augmented Approach

The classifiers (AlexNet-TA and GoogLeNet-TA) had disagreement in 13 of the 150 test cases. The 13 discordant cases were then blindly reviewed by a

Figure 3

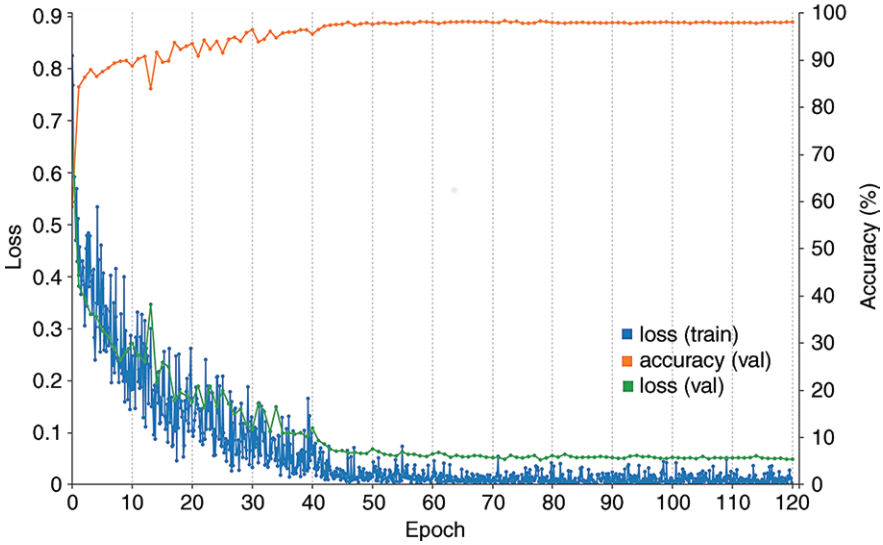


Figure 3: Training curve of AlexNet-TA classifier. The orange line represents the accuracy over the course of training, which increases over time, with a final accuracy of 98.2% at the final epoch. Training was performed for 120 epochs, and each epoch represents one pass through the entire training dataset. The blue and green curves represent the loss on the training and validation datasets, which decreases over time. The loss represents the fit between a prediction and the ground truth label. As expected, there is a reduction of loss over the course of training as accuracy improves. The loss on the validation is similar to the training, which indicates that there is no appreciable overfitting. These training curves are used for model selection. In this case, the best performing model at epoch 120 was used on the test data for final assessment. *Val* = validation.

cardiothoracic radiologist, who correctly interpreted all 13 cases (100%). The contingency table of this radiologist-augmented approach is provided in Figure 1, with a sensitivity of 97.3% and a specificity of 100%. This approach uses the classifier's answers for the 137 cases where there is agreement and the radiologist's answers for the 13 cases with disagreement.

Discussion

Machine learning is a branch of artificial intelligence in which computers are not explicitly programmed but can perform tasks by analyzing relationships of existing data (33). In this study, we use supervised DCNNs, a type of deep learning that employs multiple hidden layers and has been remarkably successful for image classification (34). It is referred to as supervised because the machine was trained on many prelabeled examples.

One of the advantages of deep learning is its ability to excel with high-dimensional datasets, such as images, which can be represented at multiple levels. For example, regarding images, DCNNs can be represented at lower levels with pixel intensity values, edges, and blobs; at intermediate levels, with parts of objects; and at higher levels, the object as a whole.

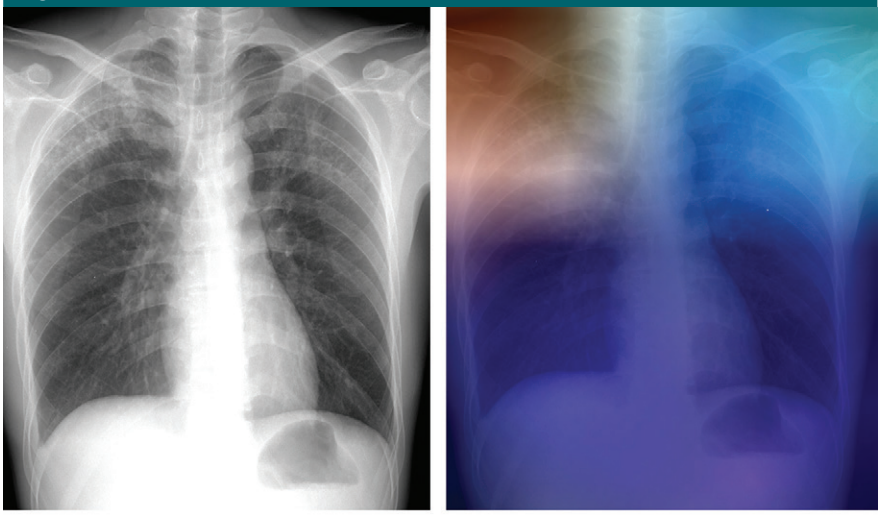
In this study, the DCNNs pretrained with everyday images on ImageNet performed better than the untrained networks, concordant with previously published works (Table 3, Fig 2) (14,15,19). This concept is called transfer learning. Although how the use of pretrained networks with nonmedical images would aid in a classification task of medical images at first may not seem intuitive, there are elements to all images that are similar, including edges and blobs that compose the initial layers of the neural network. By applying the pretrained networks to medical images, the fully connected layers were set to random initialization of weights so that they could relearn from the medical images provided.

Augmentation of the dataset with rotated images and image contrast enhancement with Contrast Limited Adaptive Histogram Equalization further improved performance (Table 3, Fig 2). It was shown (35) that more variations supplied to the neural network can improve generalization and performance of the DCNN.

One of the problems with machine learning, including deep learning, is overfitting (36). Overfitting occurs when the trained model does not generalize well to unseen cases, but fits the training data well. This becomes more apparent when the training sample size is small. Both of the DCNNs in this studied used dropout or model regularization strategies to help overcome this issue (23,24,36). Assessment of the training curve (Fig 3) can be used to assess the possibility of overfitting. From the curve, it is apparent that the data loss is similar for both validation and training datasets, which indicates well-fit curves. If there were overfitting, the loss on the training data would be much greater than that of the validation data. In addition, for this reason, the cases were split three ways (training, validation, and test). The AUCs for the receiver operating characteristic curves of the classifiers were based on the test dataset, which had not been seen by the trained networks (Table 3, Fig 2). This shows that the algorithm is generalizable and could provide accurate results with cases not previously seen.

The use of ensembles is another method to improve performance. This involves blending multiple algorithms to improve the predictive performance compared with any one algorithm alone (37). Ensembles are more effective when individual classifiers are not as correlative, and they work by removing uncorrelated errors of individual classifiers by using averaging (37). For ensemble methods in this study, we used weighted averages of the probability scores for both the AlexNet and GoogLeNet algorithms, with up to 10-fold weighting in each direction. A 10-fold weighted

Figure 4



a. b.

Figure 4: **(a)** Posteroanterior chest radiograph shows upper lobe opacities with pathologic analysis–proven active TB. **(b)** Same posteroanterior chest radiograph, with a heat map overlay of one of the strongest activations obtained from the fifth convolutional layer after it was passed through the GoogLeNet-TA classifier. The red and light blue regions in the upper lobes represent areas activated by the deep neural network. The dark purple background represents areas that are not activated. This shows that the network is focusing on parts of the image where the disease is present (both upper lobes).

average toward GoogLeNet (stronger weighting of GoogLeNet) provided better accuracy than the other tested choices, with an AUC of 0.992 (standard error, 0.0046; 95% confidence interval: 0.961, 1.000).

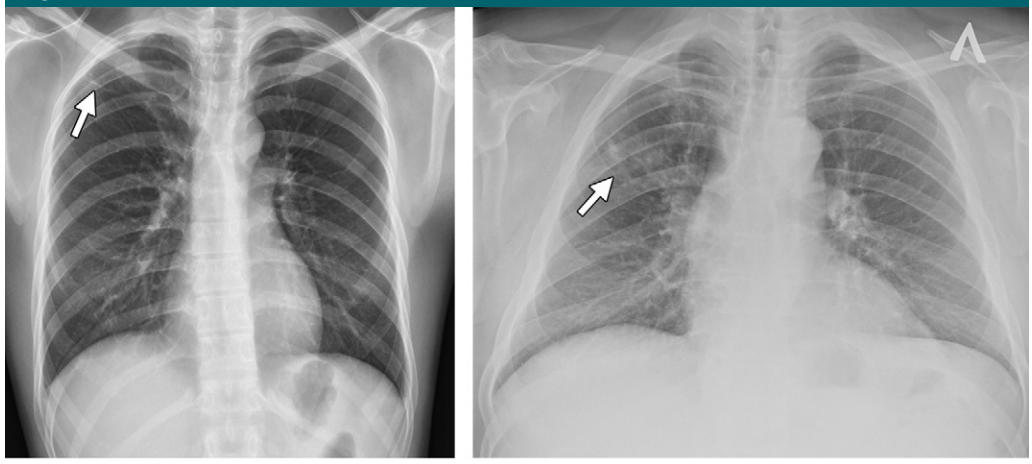
It was previously described (38) that neural networks, particularly deep neural networks, are functionally so-called black boxes, meaning it is difficult to determine how the network arrived at its conclusion. This is an important consideration because one would want to know that DCNN was looking at a parenchymal abnormality in the lung apices consistent with TB, as in the case of this study, rather than a nonrelevant part of the image per se. The functional black-box effect is complicated by the sheer size of DCNNs; for example, AlexNet has 60 000 000 trained parameters, and it would be practically infeasible to analyze them individually (20). However, there are tools that can help aid visualization of a neural network (38), which can give more confidence that a DCNN is activated by the appropriate part of the image. For example, Figure 4 shows one of the strongest activations within the GoogLeNet classifier, after being given a chest radiograph that is positive for TB. In this example, the lung apices have some of the strongest activations, which correspond to areas of parenchymal disease. One can feel more confident in the ability of the classifier through such visualization methods.

One potential method to improve accuracy is a radiologist-augmented system, in which some of the images are sent to a radiologist for a so-called overread. In this system, images that were discordant (classified by one DCNN as positive for TB and the other as negative for TB) were sent to the radiologist for the final interpretation. Of the 150 test images, the best AlexNet and GoogLeNet classifiers agreed 137 times (91.3%) and disagreed 13 times (8.7%). A blinded board-certified radiologist then reviewed the 13 discordant images and correctly classified all 13 images (100%). This radiologist-augmented approach increased the sensitivity to 97.3% and specificity to 100% (Table 2). There were two false-negative findings, which were findings missed by both the DCNNs, and therefore never made it to the radiologist for review. The false-negative findings are shown in Figure 5. In the first case, the opacity is subtle in the right upper lobe. The other case shows a more apparent opacity in the right suprahilar region. It is conceivable that the use of larger training datasets, additional image augmentation methods, and additional machine learning approaches with more ensembles could improve this result. This workflow may be helpful in certain TB-prevalent regions where access to radiologists is lacking or cost prohibitive (4,5,20), in which an automated method could solely interpret a large portion of the cases, and only the equivocal cases are sent to a radiologist. One could also imagine a system with multiple (ie, more than two) classifiers, which may further improve accuracy, because a highly accurate automated system would be more desirable in this regard.

It is interesting to note that the DCNNs in this study outperformed that described by Hwang et al (19), which showed AUCs that ranged from 0.88 to 0.96, despite many more training cases. It is unclear if this is related to the different DCNN architectures or augmentation strategies used in this study. We also used two distinct DCNNs and provided an ensemble to further improve performance. Finally, we randomized the images from all of our datasets and split them into training, validation, and tests. This included images from many sites, with chest radiographs generated by both digital radiographic and computed radiographic technologies, which may have aided generalization and performance of the DCNNs. On the other hand, Hwang et al used only one large dataset for training (from Korea), consisting of only digital radiography images, and tested cross-over performance on the additional National Institutes of Health datasets from Shenzhen, China, and Montgomery County, Maryland, which had digital radiography and computed radiography images, respectively.

There are limitations to this work. The DCNNs do not replace human

Figure 5



a. b.

Figure 5: Two images with false-negative findings missed by both classifiers. **(a)** An opacity in the right upper lobe (arrow) on a posteroanterior radiograph. **(b)** A more apparent right suprahilar opacity (arrow) on a posteroanterior radiograph.

radiologic interpretation beyond that of TB because they are not tailored to evaluate other pathologic findings. In this study, the 75 images with tests positive for TB had a relatively even distribution of subtle, intermediate, and readily apparent opacities (Table 2). However, more research is needed to determine the performance of classifiers on only subtle opacities because obvious changes should be easier to detect than subtle ones. Another important factor to consider is that system performance will be affected by the selection of cases (percent of normal vs abnormal). Also, the system is designed for use in TB-prevalent regions with the goal of differentiating normal versus abnormal regarding TB evaluation, potentially part of a chest radiography screening program. If the algorithms were used in non-TB-prevalent locations and not solely for the purpose of TB evaluation, other pathologic findings that had a similar radiographic appearance, such as lung cancers and bacterial pneumonia, may be flagged as positive. As with multiple other studies that use deep learning (11,12,14–18), the images were down-sampled to 256 × 256 pixels before they were fed into the network. This is because of the sheer number of parameters that are inherent to these networks. Larger matrix sizes will increase the training time and will require more robust systems and graphics processing unit memory. However, accuracy may improve by using higher resolution images, particularly for subtle findings, and more research is needed in this regard. Finally, this was a retrospective study on datasets that were available at the time of the study. Further prospective investigation on the use of DCNNs in a clinical practice for pulmonary TB evaluation would be valuable, particularly in TB-prevalent regions. In addition, it would be interesting to evaluate the effect of additional training cases and other augmentation strategies on accuracy. Further work also needs to be performed to evaluate the effect with Digital Imaging and Communications in Medicine files directly and other deep neural network architectures. In conclusion, deep learning with DCNNs can accurately classify TB at chest radiography with an AUC of 0.99. A radiologist-augmented approach for cases where there was disagreement among the DCNNs further improved accuracy, with a sensitivity of 97% and a specificity of 100%.

Disclosures of Conflicts of Interest: P.L. disclosed no relevant relationships. **B.S.** disclosed no relevant relationships.

References

1. World Health Organization. Global tuberculosis report 2015. http://apps.who.int/iris/bitstream/10665/191102/1/9789241565059_eng.pdf. Published October 28, 2015. Accessed September 20, 2016.
2. World Health Organization. Systematic screening for active tuberculosis: Principles and recommendations. http://www.who.int/tb/publications/Final_TB_Screening_guidelines.pdf. Published April 2013. Accessed September 20, 2016.
3. Bhalla AS, Goyal A, Guleria R, Gupta AK. Chest tuberculosis: Radiological review and imaging recommendations. Indian J Radiol Imaging 2015;25(3):213–225.
4. Melendez J, Sánchez CI, Philipsen RH, et al. An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. Sci Rep 2016;6:25265.
5. Hoog AH, Meme HK, van Deutekom H, et al. High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey. Int J Tuberc Lung Dis 2011;15(10):1308–1314.
6. Antani S. Automated Detection of Lung Diseases in Chest X-Rays. A Report to the Board of Scientific Counselors. US National Library of Medicine. https://lhncbc.nlm.nih.gov/system/files/pub9126.pdf. Published April 2015. Accessed September 20, 2016.
7. Jaeger S, Karargyris A, Candemir S, et al. Automatic screening for tuberculosis in chest radiographs: a survey. Quant Imaging Med Surg 2013;3(2):89–99.

8. Pande T, Cohen C, Pai M, Ahmad Khan F. Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: a systematic review. Int J Tuberc Lung Dis 2016;20(9):1226–1230.
9. Maduskar P, Muyoyeta M, Ayles H, Hogeweg L, Peters-Bax L, van Ginneken B. Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. Int J Tuberc Lung Dis 2013;17(12):1613–1620.
10. Jaeger S, Karargyris A, Candemir S, et al. Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging 2014;33(2):233–245.
11. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–252.
12. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv preprint. https://arxiv.org/abs/1512.03385. Published December 10, 2015. Accessed September 20, 2016.
13. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–2324.
14. Bar Y, Diamant I, Wolf L, Greenspan H. Deep learning with non-medical training used for chest pathology identification. In: Hadjiiski LM, Tourassi GD, eds. Proceedings of SPIE: medical imaging 2015—computer-aided diagnosis. Vol 9414. Bellingham, Wash: International Society for Optics and Photonics, 2015; 94140V.
15. 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):1285–1298.
16. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther 2015;8:2015–2022.
17. Roth HR, Farag A, Lu L, Turkbey EB, Summers RM. Deep convolutional networks for pancreas segmentation in CT imaging. In: Ourselin S, Styner MA, eds. Proceedings of SPIE: medical imaging 2015—image processing. Vol 9413. Bellingham, Wash: International Society for Optics and Photonics, 2015; 94131G.
18. Zhang W, Li R, Deng H, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. Neuroimage 2015;108:214–224.
19. Hwang S, Kim HE, Jeong J, Kim HJ. A novel approach for tuberculosis screening based on deep convolutional neural networks. In: Tourassi GD, Armato SG, eds. Proceedings of SPIE: medical imaging 2016—title. Vol 9785. Bellingham, Wash: International Society for Optics and Photonics, 2016; 97852W.
20. Jaeger S, Candemir S, Antani S, Wáng YX, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 2014;4(6):475–477.
21. Belarus Tuberculosis Portal. Belarus Public Health Web site. http://obsolete.tuberculosis.by/. Published September 1, 2011. Updated July 17, 2015. Accessed August 20, 2016.
22. Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia 2014. New York, NY: ACM, 2014.
23. 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; 1097–1105.
24. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015; 1–9.
25. Rasband WS. Image J. U.S. National Institutes of Health, Bethesda, Maryland, USA. http://imagej.nih.gov/ij/. 1997-2016.
26. Hastie T, Tibshirani R, Friedman J. Model assessment and selection. In: The elements of statistical learning. 2nd ed. New York, NY: Springer, 2009; 219–259.
27. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229(1):3–8.
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
29. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128–138.
30. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30 (7):1145–1159.
31. Fawcett T. ROC graphs: Notes and practical considerations for researchers. Mach Learn 2004;31(1):1–38.
32. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998;52(2):119–126.
33. Wang S, Summers RM. Machine learning and radiology. Med Image Anal 2012;16(5):933–951.
34. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
35. Wu R, Yan S, Shan Y, Dang Q, Sun G. Deep image: Scaling up image recognition. arXiv preprint. https://arxiv.org/abs/1501.02876. Published January 13, 2015. Updated July 6, 2015. Accessed September 21, 2016.
36. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–1958.
37. Dietterich TG. Ensemble methods in machine learning. Lect Notes Comput Sci 2000;1857:1–15.
38. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. arXiv preprint. https://arxiv.org/abs/1506.06579. Published June 22, 2015. Accessed September 21, 2016.