Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs

Mark Cicero, MD, BESc, Alexander Bilbily, MD, BHSc,* Errol Colak, MD, FRCPC, HBSc,* Tim Dowdell, MD, CCFP, FRCPC,* Bruce Gray, MD, FRCPC, BSc,* Kuhan Perampaladas, MSc, BSc,† and Joseph Barfett, MD, FRCPC, MSc, BESc**

Objectives: Convolutional neural networks (CNNs) are a subtype of artificial neural network that have shown strong performance in computer vision tasks including image classification. To date, there has been limited application of CNNs to chest radiographs, the most frequently performed medical imaging study. We hypothesize CNNs can learn to classify frontal chest radiographs according to common findings from a sufficiently large data set.

Materials and Methods: Our institution's research ethics board approved a single-center retrospective review of 35,038 adult posterior-anterior chest radiographs and final reports performed between 2005 and 2015 (56% men, average age of 56, patient type: 24% inpatient, 39% outpatient, 37% emergency department) with a waiver for informed consent. The GoogLeNet CNN was trained using 3 graphics processing units to automatically classify radiographs as normal (n = 11,702) or into 1 or more of cardiomegaly (n = 9240), consolidation (n = 6788), pleural effusion (n = 7786), pulmonary edema (n = 1286), or pneumothorax (n = 1299). The network's performance was evaluated using receiver operating curve analysis on a test set of 2443 radiographs with the criterion standard being board-certified radiologist interpretation.

Results: Using 256 × 256-pixel images as input, the network achieved an overall sensitivity and specificity of 91% with an area under the curve of 0.964 for classifying a study as normal (n = 1203). For the abnormal categories, the sensitivity, specificity, and area under the curve, respectively, were 91%, 91%, and 0.962 for pleural effusion (n = 782), 82%, 82%, and 0.868 for pulmonary edema (n = 356), 74%, 75%, and 0.850 for consolidation (n = 214), 81%, 80%, and 0.875 for cardiomegaly (n = 482), and 78%, 78%, and 0.861 for pneumothorax (n = 167).

Conclusions: Current deep CNN architectures can be trained with modest-sized medical data sets to achieve clinically useful performance at detecting and excluding common pathology on chest radiographs.

Key Words: machine learning, neural network, x-ray, cardiomegaly, pleural effusion, pulmonary edema, pneumothorax, pneumonia

(*Invest Radiol* 2017;52: 281–287)

In recent years, there has been tremendous advancement in the field of machine learning and computer vision through the use of artificial neural networks. In 2012, dramatic improvements in image recognition during the ImageNet Large-Scale Visual Recognition Competition were achieved using convolutional neural network (CNN) architectures.[1] These convolutional architectures, loosely based on our understanding of the human visual cortex, have been adopted by other researchers and have achieved further improvements, surpassing human-level performance on certain image recognition tasks.[2] Central to these advancements has been the use of graphics processing units (GPUs) and large training data sets.

Two popular CNN architectures that have shown excellent results on the ImageNet data set are AlexNet published by Krishevsky et al[1] in 2012 and GoogLeNet published by Szegedy et al[3] in 2014. These architectures have been designed to process 256 × 256-pixel RGB color images through a series of convolutional and pooling layers using rectified linear units as their neuron activation functions. The output of the network is a probability distribution for each of the classes defined during training. In brief, the training process involves showing the network training data many times in small batches. With each batch, the network makes guesses at the correct category and compares it to the specified labels. The difference between the network prediction and the correct label is represented by a cost function. Small adjustments to the weights connecting each neuron of the network are made iteratively to minimize the cost function.[4] Using this approach, imaging features are learned by the network rather than being explicitly defined by the programmer.

To date, there has been limited application of this technology using large data sets for diagnosis in medical imaging and chest radiographs in particular. This is in part due to the challenge of accessing high fidelity labeled training data within siloed hospital information technology systems.[5] Bar et al[6] presented a method of recognizing pathology on chest radiographs using a nonmedical training data set in 2015. Depeursinge et al[7] applied texture analysis to automate classification of usual interstitial pneumonia on high-resolution computer tomography (CT) in 33 patients and achieved an area under curve (AUC) of 0.81. Another challenge specific to medical imaging is dealing with the high resolution with which images are acquired at for diagnostic purposes. These images, usually greater than 2000 pixels in each dimension, are much larger than what is acceptable in other domains CNNs have had success in and this leads to increased computational requirements. Chest radiographs are an appropriate initial application for deep learning algorithms because of their frequency in practice, the availability of large data sets, and standardized acquisition technique. We propose that CNNs can be trained to detect abnormalities on chest radiographs and hence can assume a crucial role in computer-aided detection and diagnosis. Applications include using CNNs to alert clinicians of potential findings immediately after image acquisition before a final radiologist report. For radiologists, CNNs have many applications including a means for work list triaging to reduce the interval from image acquisition to reporting of abnormal studies.

This article discusses the training and validation of a CNN to detect cardiomegaly, pulmonary edema, consolidation, pleural effusion, and pneumothorax on single-view posterior-anterior (PA) frontal chest radiographs.

Received for publication September 17, 2016; and accepted for publication, after revision, October 24, 2016.

From the *Department of Medical Imaging, St Michael's Hospital, and †Department of Pharmaceutical Sciences, University of Toronto, Toronto, Ontario, Canada.

Correspondence to: Mark Cicero, MD, BESc, Department of Medical Imaging, St Michael's Hospital, 30 Bond St, Toronto, Ontario, Canada M5B 1W8. E-mail: mark.cicero@mail.utoronto.ca.

The authors have no conflicts of interest.

ISSN: 0020-9996/17/5205–0281
DOI: 10.1097/RLI.0000000000000341

MATERIALS AND METHODS

Study Population

Our institutional review board approved this single-center retrospective study with a waiver for informed consent. A search of our tertiary care center's Radiology Information System (Syngo; Siemens Medical Solutions USA Inc, Malvern, PA) for 2-view chest radiograph examinations was performed using Montage Search and Analytics (Montage Healthcare Solutions, Philadelphia, PA). This search identified 269,799 two-view chest radiographic examinations performed between January 1, 2005, and December 31, 2015, on patients at least 18 years of age.

A custom Python (Python Programming Foundation) computer script was used to classify these studies according to the radiology report. Studies were classified as normal or into 5 nonexclusive categories including cardiomegaly, pleural effusion, pulmonary edema, consolidation, and pneumothorax if the finding was described as being present with certainty. This was achieved by parsing each sentence of the report and searching it for inclusion key words. If the same sentence also included an exclusion key word suggesting negative sentiment, it was excluded from the original inclusion word category, but was still considered for other categories. The exclusion key words were determined by iteratively reviewing all the sentences that met the inclusion criteria and adding exclusion key words such that all the qualifying sentences described a positive finding with certainty. Modifiers like "small," "mild," and "trace" were added to exclusion key words to enhance feature learning (see Discussion). Table 1 shows the inclusion and exclusion key words used within the computer script with the corresponding number of PA studies that met the inclusion criteria listed in the rightmost column. The authors performed quality assurance of this method by manually reviewing 1000 randomly selected sentences meeting the criteria for each category until all sentences described the presence of the finding in question.

A total of 61,192 studies were identified by the above classification script and were exported from the PACS repository. Another Python computer script was written to check the DICOM metadata for the acquisition technique (PA, anterior-posterior [AP], or lateral), convert the exported PA studies from 16-bit grayscale DICOM images to 8-bit grayscale portable network graphics (PNGs) images, and name the converted image according to the study identification number for deidentification purposes. The images maintained their original resolution and default window and level settings stored in the DICOM metadata. The converted images were sorted into separate folders for each of the 6 classification categories. The AP and lateral projections were excluded. This resulted in a total of 35,038 unique PA radiographs.

It is important to recognize that if a report met inclusion criteria for 2 or more categories the corresponding frontal radiograph was included in each of these categories. This occurred 3063 times, resulting in 38,101 total training radiographs. It should also be noted that a large portion of studies (approximately 208,607) were excluded because they were reported in a way that did not meet the inclusion criteria and therefore the findings could not be verified without manual review. Figure 1 summarizes the data acquisition and categorization process. The training set consisted of 56% men and 44% women with an average patient age of 56 years. Inpatients accounted for 24% of studies, outpatients accounted for 39%, and patients from the emergency department accounted for 37% of studies.

Neural Network Architecture

We used the winning CNN from the 2014 ImageNet Competition. This network, submitted by a team from Google Inc (Mountain View, CA), named GoogLeNet, is 22-parameter layers deep and uses the inception architecture described in detail by Szegedy et al.[3] The advantage of this architecture is its computational efficiency, enabling a large neural network capable of learning complex features while being less prone to overfitting than other similarly sized networks. The first layer of the network involves a 7×7 convolutional layer with a stride of 2. Convolutional layers consist of a receptive field (ie, a patch), in this case 7×7 pixels, which moves through the image at a specified interval, in this case 2 pixels. The pixels in the patch get multiplied by weight values in a convolutional filter, the number of which is arbitrarily defined, and a bias term is added. The output is a value that forms a single value in the matrix of the next layer. As the patch moves through the image, each of the values in the matrix is computed. This process is repeated at each layer. At the core of this architecture is an "inception" module, which takes the output of the previous layer and performs 1×1 convolution, 1×1 followed by 3×3 convolution, and 1×1 followed by 5×5 convolution. The previous layer also feeds into a 3×3 max pooling followed by a 1×1 convolution. Filter concatenation combines each of these results, which are then used as inputs in the next module. Full details on the inception module can be found in the original article published by Szegedy et al.

Training Details

This architecture was designed for 256×256 pixel images; therefore, our training set was downsampled to this resolution. For

TABLE 1. Classification Criteria and No. Satisfying PA Chest Radiographs

Inclusion Key Words	Exclusion Key Words	Total No. PA Studies
Normal examination, normal study, unremarkable examination, normal chest x-ray		11,702
Cardiomegaly, cardiac silhouette	No, not enlarged, is normal, unremarkable, small, mild, slight, borderline	9240
Effusion	No, without, resolved, resolution, pericardial, indication:, small, likely, tiny, trace, follow, assess, rule	7786
Consolidation, pneumonia	No, free of, without, clear, rule out, query, assess, indication:, history:, clinical:, resolved, resolution, improv-, early, mild, possible, developing, definitely, exclude, residual, suspicion, suspicious, difficult, subtle	6788
Edema	No, R/O, rule, without, assess, query, indication:, history:, clinical:, resolved, resolving, resolution, improv-, mild, may, likely, exclude	1286
Pneumothorax	No, without, exclude, negative, definite, mimic, suspect, clinical, rule out, r/o, query, assess, indication:, history:, clinical:, resolved, resolution, small, trace, tiny, decrease, improv-, reduction	1299

PA indicates posterior-anterior.

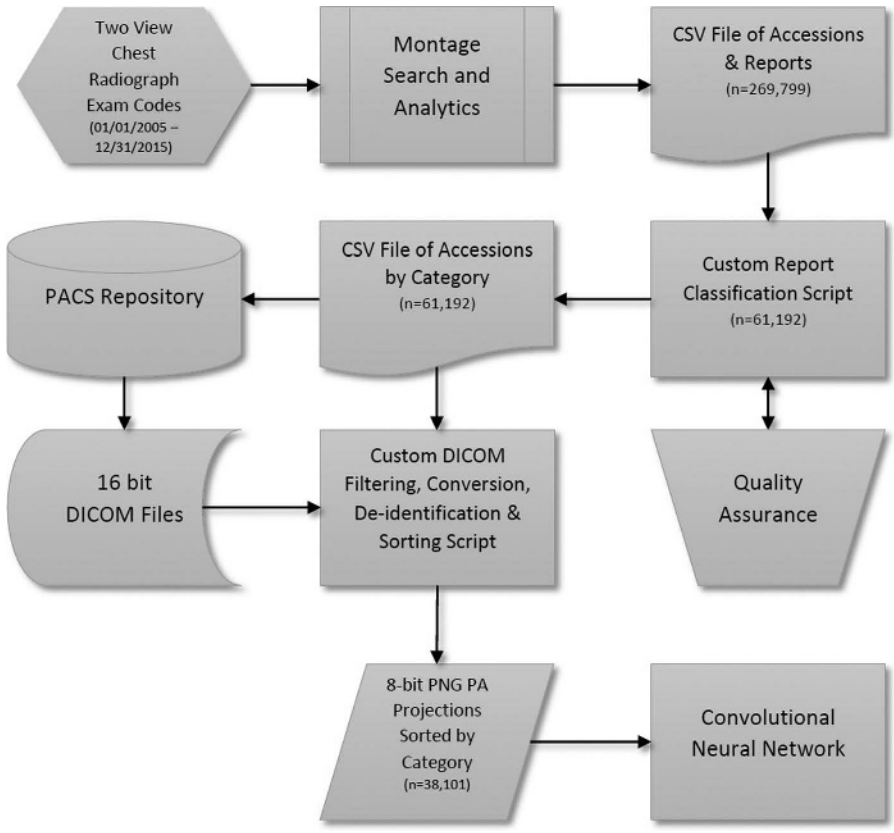

FIGURE 1. Summary of the data acquisition and categorization process.

aspect ratio imbalance, a half-crop half-fill method was used, which resizes the largest dimension to the target and then pads the shorter dimension with noise to achieve a square aspect ratio. During training, 224 × 224 pixel random crops of the image were taken along with their mirrored version. This technique is done to amplify the training data set, a sampling technique used previously.[1] The input layer is therefore 224 × 224 × 1. Rectified linear units are used as the neuron activation function.

Training took just over 1 hour using 3 NVIDIA Titan X 12GB GPUs (NVIDIA Corporation, Santa Clara, CA) and the NVIDIA Deep Learning GPU Training System (DIGITS 4.0) using the Caffe deep learning framework by the Berkeley Learning and Vision Center (BLVC). Stochastic gradient descent optimization was used with a batch size of 72 and an initial learning rate of 0.01, which was decreased according to a sigmoid decay function over the 60 iterations (epochs) through the training set. Twenty percent of the images in the training set were used for validation during training. The model reached a maximum top-1 accuracy of 67% on the validation set.

Statistical Analysis

To test the trained neural network, we estimated a sample size calculation for both sensitivity and specificity for each classification category using the equation (1) from Naing et al[8] with an expected sensitivity and specificity of 88% and a precision of 0.05 with a confidence interval of 95%. We approximated disease prevalence by performing searches in Montage over a 10-year period utilizing the "optimize for positive findings" filter for each of the 5 disease categories. The prevalence estimates were 1% for pneumothorax, 3% for edema, 4% for consolidation, 7% for pleural effusion, and 24% for cardiomegaly. Using the prevalence of pneumothorax, edema or consolidation resulted in a prohibitory number of test cases. Therefore, a prevalence of 7% for pleural effusion was used and resulted in a required sample size of 2329.

$$n = \frac{z^2 P(1-P)}{d^2} \quad (1)$$

Randomly allocating approximately 7% of the total classified data set to the test set resulted in an adequate sample size for testing. The remainder of images (32,586) were allocated to the training set. Images in the test set were not used for training purposes.

The findings in the test set were validated by 2 board-certified radiologists: one with 3 and one with 8 years of experience. Reviewers did not have access to the final report, the clinical history, or prior studies. Reviewers looked at the full-resolution PA views, usually greater than 2000 pixels in each dimension. There were 2 test cases with indeterminate findings. These cases were discussed by the 2 reviewers and a joint conclusion was made. Images that excluded structures of a full frontal chest radiograph (ie, cutoff costophrenic angles or lung apices) were removed from the test set. The final total number of unique radiographs in the test set was 2443. Table 2 shows the distribution of findings in the test set.

We then classified the test set using the trained neural network and performed receiver operating curve (ROC) analysis for each category; hence, a total of 6 ROC analyses were performed. In each case, the cutoff for a positive result was increased by increments of 10 and the sensitivity and false-positive rates of the neural network were plotted. Fine-tuning of the cutoffs was done in smaller increments to

TABLE 2. ROC Analysis Results at Threshold to Maximize Sensitivity and Specificity With AUC Measurements

Category	Number	Prevalence, %	Sensitivity, %	Specificity, %	AUC	Threshold, %	Accuracy, %	PPV, %	NPV, %
Normal	1203	49	91	91	0.964	20.0	91	90	91
Cardiomegaly	482	20	81	80	0.875	23.0	80	49	95
Effusion	782	32	91	91	0.962	8.0	91	82	95
Consolidation	214	9	74	75	0.850	15.5	75	23	97
Edema	356	15	82	82	0.868	1.1	82	43	96
Pneumothorax	167	7	78	78	0.861	3.8	78	49	95

ROC indicates receiver operating curve; AUC, area under curve.

maximize the sum of the sensitivity and specificity for each category independently.

RESULTS

The network achieved the highest AUC in the normal category (AUC = 0.964) with sensitivity and specificity of 91% when greater than or equal to 20% output probability was used as the threshold for a positive result. Figure 2 shows the ROC curve for the normal category. Results of other categories are summarized in Table 2.

Figure 3 shows six cases where the neural network classifications were correct. That is, the probabilities of each finding were above the thresholds obtained from the ROC analysis. In all cases, the top-1 prediction is concordant with a true finding. In Figure 3, D and E, effusion and cardiomegaly are the second predictions which are also respectively present. In Figure 3F, a left apical pneumothorax is present, although difficult to see at low resolution, the network is able to detect it. It also activates the effusion and consolidation categories above their respective thresholds.

In Figure 4, we show cases where the network made incorrect predictions. In Figure 4A, the dominant feature is a pleural effusion with cardiomegaly. While the network readily identifies the left pleural effusion, it is unable to confidently predict cardiomegaly. In Figure 4B, the network correctly detects cardiomegaly and suspects an effusion; however, it does not activate high enough for pulmonary edema that is present. The network classifies Figure 4C as highly normal and misses a subtle streaky opacity in the right lower lung, suggesting pneumonia. Figure 4D highlights the network's challenge with pneumothorax. Figure 4E magnifies the pneumothorax present in Figure 4D, which is arguably not humanly perceptible at 256 × 256 pixels. The network is unable to detect this and instead classifies the image as normal. Figure 4F depicts a pitfall in patients with breast shadows where a normal x-ray is classified as possibly being normal or possibly having a pneumothorax, consolidation, or an effusion.

DISCUSSION

The results of this study indicate the considerable potential of CNNs to provide an initial interpretation of medical images such as chest radiographs. One challenge, as with other machine learning endeavors, is obtaining a large quality labeled data set for training. In this study, searching reports by their mention of pertinent findings proved to be an effective initial means of obtaining a training data set; however, this form of labeling images is still considered to be a source of error as the report may be incorrect. Adoption of structured reporting in daily practice would assist in future machine learning endeavors as it would simplify text parsing and facilitate the labeling process. Correlation with patients who also underwent CT would have been ideal to use as a criterion standard; however, this would result in too few training examples to learn meaningful representations. In addition, we elected to exclude reports describing equivocal findings to enhance feature learning, which may decrease the network's performance on cases where

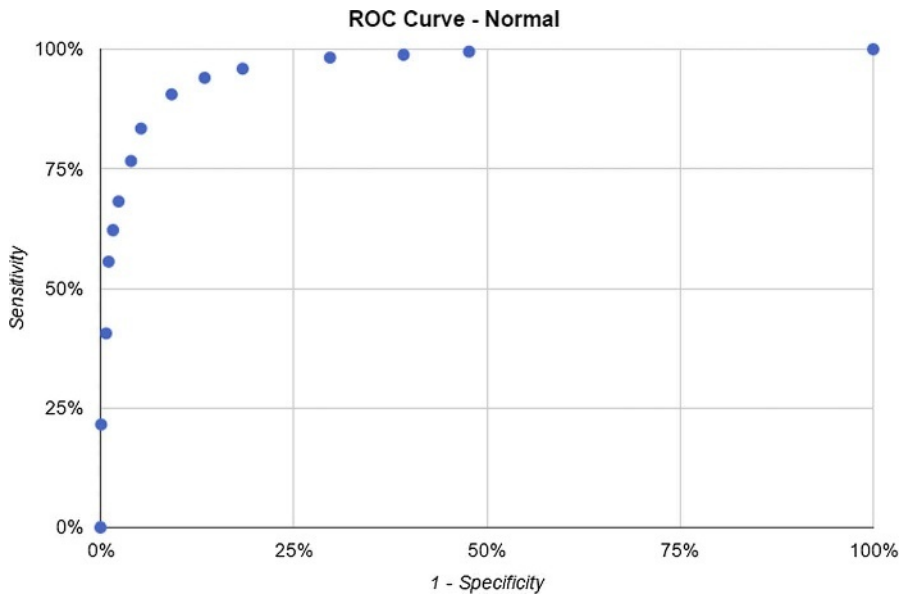

FIGURE 2. ROC curve for the "normal" category. Figure 2 can be viewed online in color at www.investigativeradiology.com.

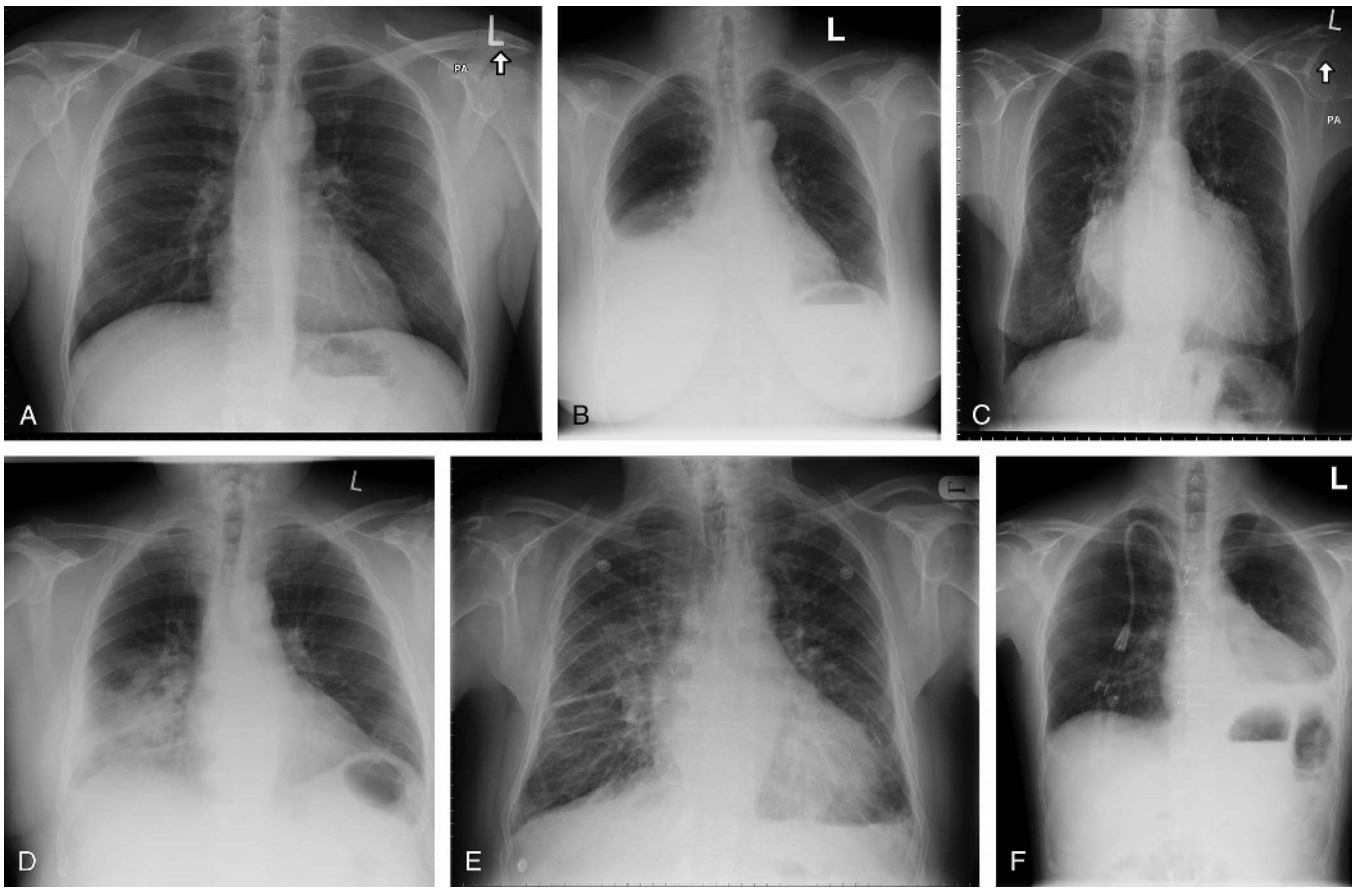

FIGURE 3. Selected true-positive examples. Positive results according to ROC analysis are provided in parentheses. A, Normal, 99.14% (>20.0%); consolidation, 0.46% (>15.5%); cardiomegaly, 0.13% (>23.0%); effusion, 0.13% (>8.0%); pneumothorax, 0.12% (>3.8%). B, Effusion, 77.91% (>8.0%); consolidation, 11.75% (>15.5%); cardiomegaly, 9.1% (>23.0%); pneumothorax, 0.96% (>3.8%); edema, 0.2% (>1.1%). C, Cardiomegaly, 90.4% (>23.0%); normal, 7.12% (>20.0%); edema, 1.0% (>1.1%); effusion, 0.69% (>8.0%); consolidation, 0.74% (>15.5%). D, Consolidation, 47.2% (>15.5%); effusion, 25.43% (>8.0%); cardiomegaly, 19.84% (>23.0%); pneumothorax, 4.8% (>3.8%); normal, 1.4% (>20.0%). E, Edema, 63.63% (>1.1%); cardiomegaly, 26.84% (>23.0%); effusion, 5.29% (>8.0%); consolidation, 4.13% (>15.5%); normal, 0.07% (>20.0%). F, Pneumothorax, 57.27% (>3.8%); effusion, 21.35% (>8.0%); consolidation, 18.34% (>15.5%); cardiomegaly, 1.30% (>23.0%); normal, 1.28% (>20.0%).

findings are subtle. For this particular application, there was correlation between many of the findings. We feel that this decreased the network's accuracy in predicting the correct category as the first prediction (top-1 accuracy) on the validation set during training because the same image was presented to the network with more than 1 correct label. We suspect this would be less of a problem as the size of the data set and number of possible categories increase.

Differences in image quality and acquisition technique also play a role in the design of a robust model. Radiologists must make diagnoses despite suboptimal image quality, and CNNs have gained a reputation at performing well even when variability is present. For example, they can correctly classify objects in images despite differences in object position, orientation, and magnification even if the entire object is not present in the image. However, they do exhibit potentially undesirable properties when it comes to small nonrandom perturbations, even if imperceptible.[9] The likelihood that the kinds of perturbations required to maximize the prediction error would be found in a naturally generated image is unknown. This property highlights the tremendous importance of abnormality localization in the clinical adoption of any computer-aided detection algorithm. Convolutional neural networks can indeed propose regions of interest corresponding to their classification predictions, a method that has been dubbed R-CNN.[10] Our tests indicate that the advantage gained by including both PA and AP images by having a larger data set likely offsets the potential drawbacks from differences in acquisition technique.

Having a balanced data set of categories is also an important consideration that has been discussed in the literature.[11] This becomes an issue when the goal is to detect rare pathology (ie, pneumothorax) as there are often not enough examples available to form a robust training set to match other categories. When there is a class imbalance, there is risk for the network to be biased toward the classes to which it has had more exposure. Receiver operating curve analysis helps to offset this by determining independent optimal thresholds for a positive result.

Another challenge in applying machine learning to medical data sets is training for rare pathology as there are few examples for the network to learn from. They also pose a challenge for validation due to the large sample size requirements. Our sample size is underpowered to validate the performance of the network to detect consolidation, edema, and pneumothorax. The required sample size based on a 1% prevalence of pneumothorax and a sensitivity of 78% is over 26,000, more than half the size of our entire data set. Manually reviewing this many studies would also be costly. Finally, the impact of rare diseases on positive predictive value is readily apparent in our results. That is, for diseases with low prevalence, the false-positive rate increases.[12]

The largest limitation of applying deep CNNs to radiographs and medical images in general is image resolution. Typical chest

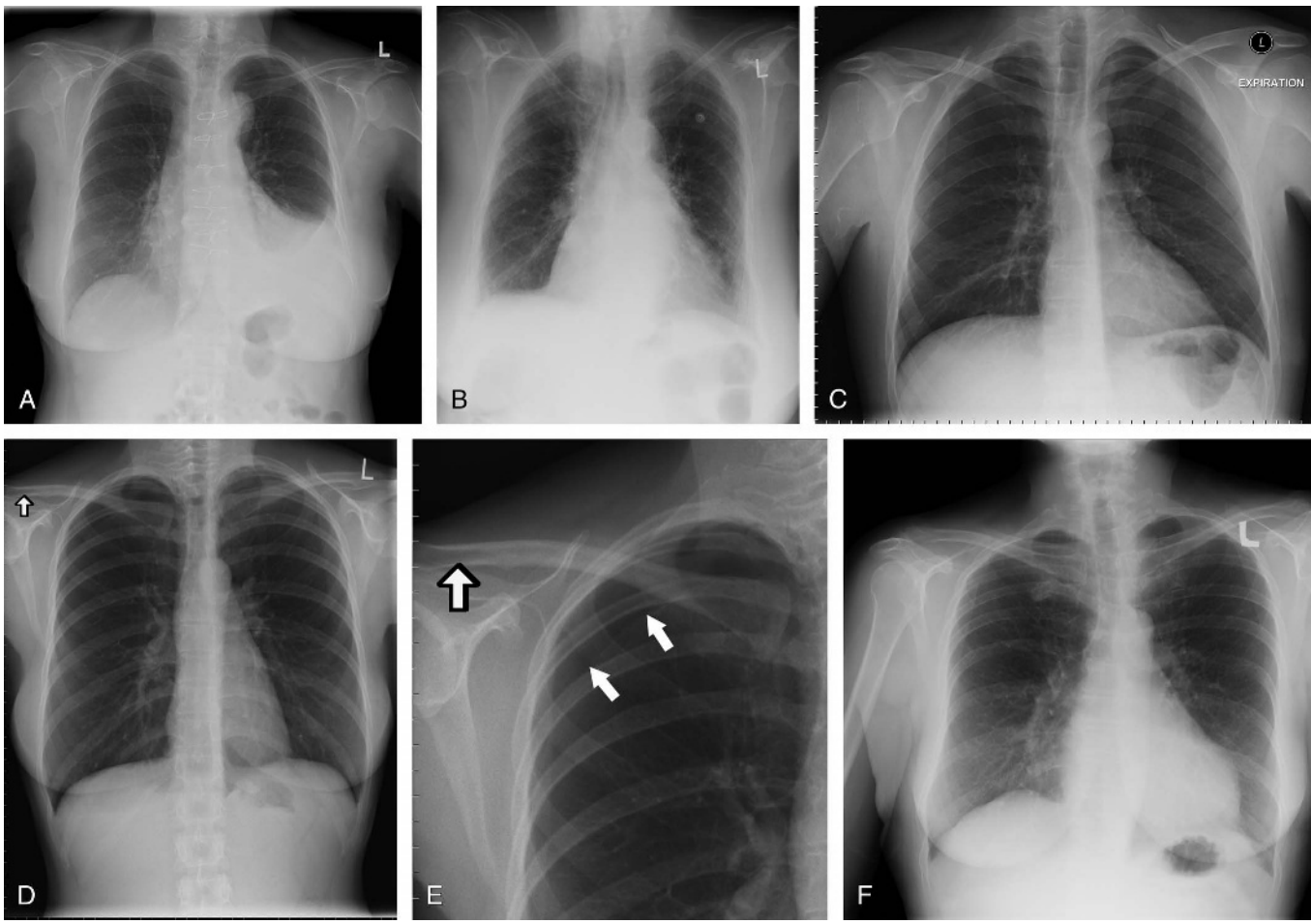

FIGURE 4. Selected false-negative and false-positive examples. Positive results according to ROC analysis are provided in parentheses. A, Effusion, 77.74% (>8.0%); cardiomegaly, 9.59% (>23.0%); consolidation, 9.16% (>15.5%); pneumothorax, 3.14% (>3.8%); normal, 0.19% (>20.0%). B, Cardiomegaly, 75.19% (>23.0%); effusion, 13.53% (>8.0%); normal, 5.15% (>20.0%); consolidation, 3.85% (>15.5%); pneumothorax, 1.70% (>3.8%). C, Normal, 97.34% (>20.0%); consolidation, 1.00% (>15.5%); pneumothorax, 0.69% (>3.8%); effusion, 0.52% (>8.0%); cardiomegaly, 0.41% (>23.0%). D, Normal, 98.51% (>20.0%); consolidation, 0.69% (>15.5%); pneumothorax, 0.45% (>3.8%); effusion, 0.19% (>8.0%); cardiomegaly, 0.13% (>23.0%). E, Pneumothorax in panel D magnified. F, Normal, 31.48% (>20.0%); consolidation, 30.87% (>15.5%); effusion, 18.57% (>8.0%); cardiomegaly, 11.75% (>23.0%); pneumothorax, 6.55% (>3.8%).

radiographs at our institution are acquired as 16-bit grayscale images up to 2500 pixels in height. The requirement to downsample images comes from the enormous number of parameters at each layer of the network. Even at 256 pixels, a 7-layer CNN can result in 60 million parameters.[1] We hypothesize that the detection of certain pathology (ie, pneumothorax, consolidation, and edema) would improve with increasing image resolution; however, further experimentation is needed to confirm this. Other pathologies, less reliant on image resolution, such as pleural effusion or cardiomegaly, are reasonably detectable at 256 × 256 pixels. A solution to this limitation is to develop more memory efficient neural networks, similar to that seen in GoogLeNet, which is an active area of research in machine learning.[13] The development of faster and larger memory GPUs or distributing computation on large-scale computing clusters will also allow for full-resolution medical images to be used as input. It is difficult to accurately predict the amount of computational power required to accommodate full-resolution images and achieve successful learning; however, we hypothesize it may be possible with the state-of-the-art dedicated deep learning systems currently emerging on the market.

Although CNNs show promise in image classification, their utility as decision support tools and adoption into clinical practice will depend on the rationalization of their decisions.[14] This will bring much needed comfort and confidence allowing physicians to verify predictions made by the network and ensure predictions are not due to extraneous factors such as technique, position, or a host of other preprocessing factors. While our current network does not perform this task, the possibility exists and is an active area of research.[15]

Our most important finding, despite the limitations discussed previously, is that CNNs have potential to differentiate between normal and grossly abnormal studies with high confidence. We predict promising results could be achieved in the classification of findings on other types of radiographic studies such as extremity x-rays and abdominal x-rays if enough training examples are made available. Similar techniques can be applied to 3D volume sets such as CT and MRI; however, the memory requirements and amount of training examples required to achieve successful results are difficult to predict.[16]

CONCLUSIONS

This study demonstrates that the state of today's CNNs can be trained with a modestly sized medical data set to meaningfully detect pathology on chest radiographs. Similar to how physicians receive an

immediate computer interpretation when they order an electrocardiogram, we predict that in the future, CNNs will enable immediate screening interpretations of radiographs for clinicians. Although there is no evidence that CNNs can make complex problem solving decisions typical of radiology reporting, this technology may aid in triaging work lists and providing a "second look" for radiologists, helping to reduce human error.

ACKNOWLEDGMENTS

The authors thank the support of NVIDIA Corporation with the donation of the Titan X GPUs used for this research.

REFERENCES

1. Krishevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012;25:1106–1114.
2. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *arXiv preprint arXiv*. 2015; 1512.03385.
3. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015:1–9.
4. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Cogn Model*. 1988;5:1.
5. Groves P, Kayyali B, Knott D, et al. The 'big data' revolution in healthcare—accelerating value and innovation. *Center for US Health System Reform Business Technology Office*. 2013:1–17.
6. Bar Y, Diamant I, Wolf L, et al Deep learning with non-medical training used for chest pathology identification. In:*SPIE Medical Imaging*. *International Society for Optics and Photonics*. 2015;94140V.
7. Depeursinge A, Chin AS, Leung AN, et al. Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution computed tomography. *Invest Radiol*. 2015;50:261–267.
8. Naing L, Winn T, Rusli BN. Practical issues in calculating the sample size for prevalence studies. *Med Stat Arch Orofacial Sci*. 2006;1:9–14.
9. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. *arXiv*. 2014; 1312.6199v4.
10. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv*. 2013; 1311.2524.
11. Huang C, Li Y, Change Loy C, et al. Learning deep representation for imbalanced classification. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016:5375–5384.
12. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ*. 1994;309:102.
13. Zhou X, Li S, Qin K, et al. Deep adaptive network: an efficient deep neural network with sparse binary connections. *arXiv preprint arXiv*. 2016; 1604.06154.
14. Zintgraf L, Cohen T, Welling M. A new method to visualize deep neural networks. *arXiv*. 2016; 1603.02518v2.
15. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2016; 1506.01497v3.
16. Dou Q, Chen H, Yu L, et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Trans Med Imaging*. 2016; 35:1182–1195.