Pattern Recognition 63 (2017) 476–486

Contents lists available at ScienceDirect

Pattern Recognition

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

Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: MTANNs vs. CNNs

Nima Tajbakhsh*, Kenji Suzuki

Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3300 S Federal St, Chicago, IL60616, USA

ARTICLE INFO

Keywords:
Deep learning
Patch-based machine learning
Image-based machine learning
Massive-training artificial neural network
Convolution neural network
Focal lesions
Classification
Computer-aided diagnosis
Lung nodules

ABSTRACT

End-to-end learning machines enable a direct mapping from the raw input data to the desired outputs, eliminating the need for hand-crafted features. Despite less engineering effort than the hand-crafted counterparts, these learning machines achieve extremely good results for many computer vision and medical image analysis tasks. Two dominant classes of end-to-end learning machines are massive-training artificial neural networks (MTANNs) and convolutional neural networks (CNNs). Although MTANNs have been actively used for a number of medical image analysis tasks over the past two decades, CNNs have recently gained popularity in the field of medical imaging. In this study, we have compared these two successful learning machines both experimentally and theoretically. For that purpose, we considered two well-studied topics in the field of medical image analysis: detection of lung nodules and distinction between benign and malignant lung nodules in computed tomography (CT). For a thorough analysis, we used 2 optimized MTANN architectures and 4 distinct CNN architectures that have different depths. Our experiments demonstrated that the performance of MTANNs was substantially higher than that of CNN when using only limited training data. With a larger training dataset, the performance gap became less evident even though the margin was still significant. Specifically, for nodule detection, MTANNs generated 2.7 false positives per patient at 100% sensitivity, which was significantly ($p < 0.05$) lower than the best performing CNN model with 22.7 false positives per patient at the same level of sensitivity. For nodule classification, MTANNs yielded an area under the receiver-operating-characteristic curve (AUC) of 0.8806 (95% CI: 0.8389–0.9223), which was significantly ($p < 0.05$) greater than the best performing CNN model with an AUC of 0.7755 (95% CI: 0.7120–0.8270). Thus, with limited training data, MTANNs would be a suitable end-to-end machine-learning model for detection and classification of focal lesions that do not require high-level semantic features.

1. Introduction

End-to-end learning machines are particular machine learning models that seek a direct mapping from the raw input image data to the target output, eliminating the need for the design of an intermediate feature space. As a result, such learning machines require less engineering effort and fewer user interventions to produce the desired outputs. Yet, they have achieved extremely good results for many computer vision tasks, breaking state-of-the-art performance records previously held by the heavily engineered and hand-crafted approaches such as part-based models [1] and bag of visual words [2]. Although originally developed in the computer vision community, end-to-end machine-learning models have now found their ways to a variety of disciplines including natural language processing [3–5], drug discovery [6–8], and medical image analysis [9–13].

In this paper, we consider two classes of end-to-end learning machines, namely, massive-training artificial neural networks (MTANNs) and convolutional neural networks (CNNs). Although MTANNs have been actively used for a number of medical image analysis tasks over the past two decades, CNNs have recently emerged in the field of medical imaging, as a promising technique. It would be interesting to study how these two successful learning machines, which both stem from artificial neural networks (ANNs), compare to each other theoretically and experimentally. To that end, we investigated the performance of MTANNs and CNNs in lung nodule detection and classification, two well-studied topics in the field of medical image analysis. To our knowledge, no prior research has compared the effectiveness of MTANNs and CNNs in neither computer vision nor medical image analysis fields.

* Corresponding author.
E-mail addresses: ntajbakh@iit.edu (N. Tajbakhsh), ksuzuki@iit.edu (K. Suzuki).

http://dx.doi.org/10.1016/j.patcog.2016.09.029
Received 30 January 2016; Received in revised form 4 September 2016; Accepted 21 September 2016
Available online 07 October 2016

2. Material and methods

2.1. Massive-Training Artificial Neural Networks (MTANNs)

As the extension of neural filters [14,15], MTANNs can accommodate various pattern-recognition tasks [16–18] such as detection of focal lesions and classification of lesion types. MTANNs come at 2 major models: (1) 2D MTANNs, which are designed for processing 2D images, and (2) 3D MTANNs, which are the generalized form of 2D MTANNs and are designed for processing volumetric data. The first appearance of 2D MTANNs dates back in 2002 when they were developed for the reduction of false positives in computerized detection of lung nodules in low-dose computed tomography (CT) in a slice-by-slice fashion. The use of 2D MTANNs was further extended to a number of applications including the separation of bones from soft tissue in chest x-ray (CXR) [16] and the distinction between benign and malignant lung nodules on 2D CT slices [17]. The 3D MTANNs were first developed in 2006 for removing a particular source of false positives (i.e., rectal tubes) in computer-aided detection of polyps in CT colonography [18]. The success of 3D MTANNs in removing rectal tubes set the foundation for the subsequent 3D MTANN-based systems [19–22] for computer-aided detection of polyps.

An MTANN employs an ANN regression model that is capable of operating on pixel data directly. In the applications to focal lesion detection and classification, an MTANN adopted a shallow network [18–21] because low-level and mid-level representations of patterns were sufficient for those tasks, though it is capable of having a deeper network. A mixture of expert MTANNs utilizes an ensemble of multiple MTANNs with a combiner. This is because one single ANN regression model has a limited learning capacity and thus may not learn all the essential features needed to distinguish a lesion with a large appearance variability from non-lesion structures. Hence, the first step in the design of multiple MTANNs is to divide the non-lesion class into a number of sub-classes and then train each of the MTANNs to distinguish between the lesion class and each of the non-lesion subclasses. During the testing stage, each case receives a score from each of the MTANNs in the ensemble. To produce the final score, the outputs of individual MTANNs for each case are combined by the combiner such as averaging, a logical AND operator, or an additional ANN called an integration ANN. The architecture of each ANN in the MTANNs consists of hidden layers (typically one layer for focal lesion applications) with sigmoid activation functions and 1 output layer with a linear activation function. In the following, we explain how an ANN regression model is trained in the 2D MTANN framework.

Fig. 1 illustrates the training process. The inputs to the MTANN in the training process are a set of regions of interest (ROIs) and the corresponding "teaching" (desired) images that have the same size as the ROIs. Each pixel in the teaching image indicates the likelihood of the corresponding pixel in the input ROI to be the pattern of interest (e.g., a lesion). A negative ROI is extracted away from a lesion, and thus the corresponding teaching image is black. A positive ROI is centered on a lesion's location; and thus, the corresponding teaching image contains a certain distribution such as a 2D Gaussian function in its center. Mathematically, a teaching image is defined as follows:

$$T(v) = \begin{cases} \exp\{(v-\mu)^T \Sigma^{-1}(v-\mu)\}, & \text{for a lesion} \\ 0, & \text{otherwise,} \end{cases}$$

where $v = (x, y)$ indicates the location of a pixel with respect to the origin, μ denotes the center of the ROI, and Σ denotes the covariance matrix. Because lesions can generally appear in arbitrary shapes in medical images, it is common to use a diagonal covariance matrix with equal variance in each direction $\Sigma = diag(\sigma)$, where σ determines the pace of decay in the lesion likelihood as we deviate from the center of the ROI.

Once the training ROIs and the corresponding teaching images are constructed, the actual pairs of training samples and outputs are extracted. Training samples are overlapping or non-overlapping subregions (patches) that are extracted from the input ROI. If subregions are extracted at all locations in the input ROI, the centers of consecutive (overlapping) subregions differ by just one pixel. Because the input and output of the MTANN are an image patch and a single pixel, respectively, training patches from the image boarder (edge) areas cannot be extracted without zero padding. To avoid the use of image boarder zero padding, the training samples are extracted from the locations that are away from the image boarders by at least half of the image patch size. All pixel values in each of the subregions are entered as input to the ANN, whereas one pixel from the teacher image is entered into the output unit in the ANN as the teaching value. This single pixel is chosen at the location in the teacher image that corresponds to the center of the input subregion. Therefore, the training set from the ith input ROI, R_i, is represented by:

$$\mathcal{I}_{R_i} = \{\vec{I}_s(x, y) | x, y \in R_i\} = \{\vec{I}_1, \vec{I}_2, \ldots, \vec{I}_k\} \mathcal{T}_{R_i} = \{T_s(x, y) | x, y \in R_i\} = \{T_1, T_2, \ldots, T_k\}$$

where $\mathcal{I}_R$ and $\mathcal{T}_R$ denote the set of training image patches and the corresponding teaching outputs, respectively. Also, $\vec{I}_1$ denotes the first image patch extracted from the top left corner of the ROI that has been reshaped in the form of a vector, and T_1 denotes the corresponding

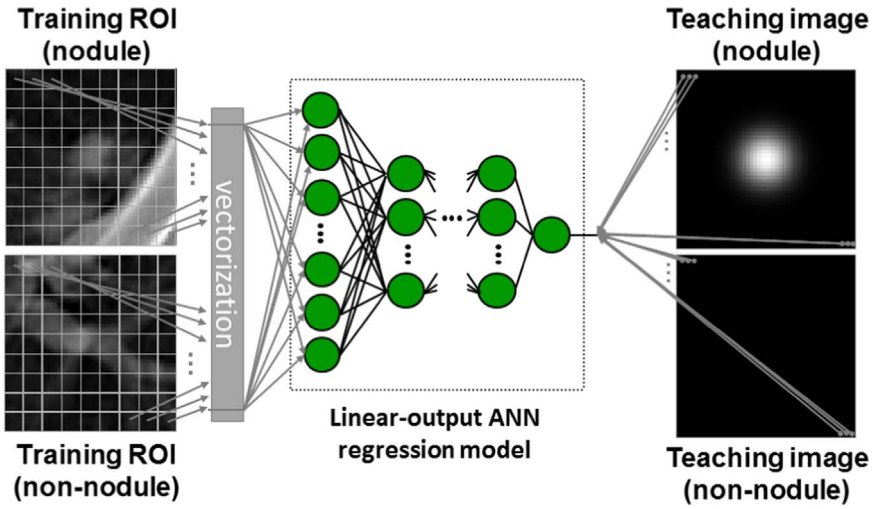


Fig. 1. Schematic overview of MTANN training. To avoid clutter in the figure, non-overlapping patches are depicted in the region of interest (ROI). In practice, the image patches are densely extracted from each ROI, resulting in a massive set of training patches.

teaching value for $\vec{I}_l$. An ROI of size $N \times N$ allows for the extraction of $k = (N - n + 1)^2$ subimages where n is the size of each subimage along x and y directions. The final massive training set is formed as the union of the training samples and the corresponding teaching values collected from each training ROI. Mathematically,

$$\mathcal{I} = \bigcup_i \mathcal{I}_{R_i}, \quad \mathcal{T} = \bigcup_i \mathcal{T}_{R_i}$$

During the testing stage, each MTANN in the ensemble is applied to an ROI in a convolutional fashion, producing a confidence map with the same size as the input ROI, if appropriate image boarder padding is applied; otherwise; smaller by at half of the image patch size on each side. To convert the confidence map into a single score for each ROI, the confidence map is multiplied with the same 2D Gaussian function used during the training stage, and then the resulting confidence values are summed. The choice of a 2D Gaussian function is motivated by the fact that a lesion is a focal object; and thus, pixels that are located farther away from the center pixel should contribute less to the lesion likelihood than do the nearer pixels. Next, the scores generated by different MTANNs for an ROI are combined by using averaging, the logical AND, or the integration ANN in order to produce the final score for the given ROI.

2.2. Convolutional Neural Networks (CNNs)

A CNN can be viewed as a simplified version of the Neocognitron model [23–25], which was proposed to simulate the human visual system in 1980 [23]. CNNs initially appeared in the early 1990s [26,27], but they did not enjoy much popularity at the time due to limited computational resources. However, with the advent of powerful graphics processing unit (GPU) computing and abundance of labeled training data, CNNs have once again emerged as a powerful feature extraction and classification tool, yielding record-breaking results in major computer vision challenges. The success of CNNs in computer vision has widely inspired investigators in the medical imaging community, resulting in a number of publications in a short period of time [28–32,10,13], which collectively demonstrates the effectiveness of CNNs for a variety of medical imaging tasks.

CNNs are so-named due to the convolutional layers in their architectures. Convolutional layers are responsible for detecting certain local features in all locations of their input images. To detect local structures, each node in a convolutional layer is connected to only a small subset of spatially connected neurons in the input image channels. To enable the search for the same local feature all over the input channels, the connection weights are shared between the nodes in the convolutional layers. Each set of shared weights is called a *kernel* or a *convolution kernel*. Thus, a convolutional layer with n kernels learns to detect n local features whose strength across the input images is visible in the resulting n feature maps. To reduce computational complexity and achieve a hierarchical set of image features, each sequence of convolution layers is followed by a *pooling layer*. The max pooling layer reduces the size of feature maps by selecting the maximum feature response in overlapping or non-overlapping local neighborhoods, discarding the exact location of such maximum responses. As a result, max pooling can further improve translation invariance. CNNs typically consist of several pairs of convolutional and pooling layers, followed by a number of consecutive 1×1 convolutional layers (a.k.a., fully connected), and finally a *softmax layer*, or a *regression layer*, to generate the desired outputs. In more modern CNN architectures, to achieve more computational efficiency, the pooling layer is replaced with a convolution layer with a stride larger than 1. A CNN typically has a large number of convolutional and fully connected layers; therefore, it is not uncommon for a CNN to contain millions or billions of weights in its architecture.

The inputs for training a CNN are a set of images and the corresponding labels. As with ANNs, weights in a CNN are first randomly initialized using a Gaussian distribution or initialized using smarter techniques [33,34] and are then updated using the back-propagation algorithm. However, because CNNs are parameter-rich models, they may over-fit to the training data. This can be a critical issue for the applications where only labeled training data are available. Common techniques to tackle the over-fitting problem are data augmentation [35], dropout regularization [36], and fine-tuning [37]. In data augmentation, a set of label-preserving image transformations is applied to each ROI, generating a large number of new yet correlated training samples. The common transformations are image scaling, translation, and rotation. Although the samples resulted from data augmentation are correlated, they have proved effective in reducing over-fitting. Dropout is a regularization technique, which, in each iteration, excludes a random subset of parameters from the weight update process. This simple technique can hinder over-fitting to the training data. Fine-tuning is also a very effective technique where the weights in a CNN are not trained from randomly initialized values, but rather from the weights of a CNN that is pre-trained on a large set of labeled training dataset from a different application. The above techniques have made it possible to obtain a high-performance CNN even for the vision applications where only limited training data are available.

2.3. Databases

2.3.1. Lung nodule detection

We used a database of low-dose thoracic helical CT (LDCT) [38,39] acquired from 31 patients, who participated voluntarily in a lung cancer screening program between 1996 and 1999 in Nagano, Japan. This database consists of 38 scans with a total of 1057 sections (slices) of size 512×512. The scans were acquired under a low-dose protocol of 120 kVp with 25 mA or 50 mA. Each section has 10 mm thickness and the pixel sizes within the sections vary between 0.586 and 0.684 mm. An experienced chest radiologist annotated 50 lung nodules in the scans, of which 38 nodules were confirmed lung cancers that had been "missed" by reporting expert radiologists during the initial clinical interpretation. Thus, this database contained very "difficult" nodules. The remaining 12 nodules in the scans were classified as "confirmed benign" (n=8), "suspected benign" (n=3), or "suspected malignant" (n=1). The above nodule classification was made through biopsy or by follow-up over a period of at least 2 years. Fig. 2 shows examples of nodules from this database. We used this database in this study, because it is a database with histopathological confirmations of the lesions, and because it has very challenging nodules that had been missed by expert radiologists.

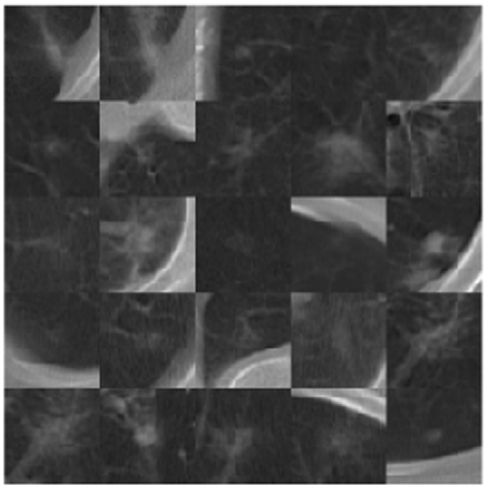


Fig. 2. Examples of nodules in our database. Non-solid (ground-glass) nodules and part-solid nodules that are major sources of false negatives are seen.

2.3.2. Lung nodule classification

We used a nodule database [38] consisting of 76 histopathologically confirmed lung cancers in 73 patients and 413 benign nodules in 342 patients. The nodule size ranged from 3 mm to 29 mm. Of the 76 primary lung cancers, 22 (28.9%) nodules were identifiable in a single section, 37 (48.7%) nodules in two sections, and 17 (22.3%) nodules in three sections. The 413 benign nodules consisted of 265 (64.2%) nodules in a single section, 133 (32.2%) in two sections, and 15 (3.6%) nodules in three sections. An experienced chest radiologist determined the center of each nodule in the section wherein the nodule appeared the largest (if the nodule appeared in more than 1 section). Fig. 3 shows examples of benign and malignant nodules from this database. We used this database in this study, because it is a database with histopathological confirmations of all lesions, and because classification of nodules in LDCT is a very challenging task even for expert radiologists.

3. Experiments

The architectures of MTANNs were chosen according to the corresponding publications [16,17]. For lung nodule detection, we used the architecture suggested in [16], which consisted of 9 MTANNs: 5 MTANNs were trained to distinguish nodules from various-sized vessels; and 4 MTANNs were applied to eliminate some other opacities. Each MTANN in the ensemble consisted of 1 hidden layer with 25 neurons, and it was trained with subimages of size 9×9 extracted from 10 nodule ROIs and 10 non-nodule ROIs of size 50×50. In the testing stage, the output of each MTANN was binarized by a pre-specified threshold; and then, the resulting binary outputs were combined using a logical AND operator. For lung nodule classification, we used the architecture suggested in [17], which consisted of 6 MTANNs with 20 neurons in their single hidden layers. Each MTANN was trained with subimages of size 9×9 extracted from 10 malignant nodule ROIs and 10 benign ROIs of size 50×50. In the testing stage, the output of each MTANN was fed to an integration ANN to produce the final confidence score. The integration ANN had 1 hidden layer with 4 neurons in it.

We used 4 distinct CNN architectures in our experiments: a shallow CNN (sh-CNN), a LeNet architecture, a relatively deep CNN (rd-CNN) whose deviations are commonly used in medical imaging applications, and a deep CNN called AlexNet that is popular in the computer vision community. These 4 architectures are shown in Fig. 5 and are further detailed in Table 1. We trained the above CNN architectures from scratch by minimizing the logistic cost function $\mathcal{L} = -\sum_n \log(p_n) + \lambda W^2$ where W denotes the network weights, p_n is the probability that the nth training sample belongs to the true class,

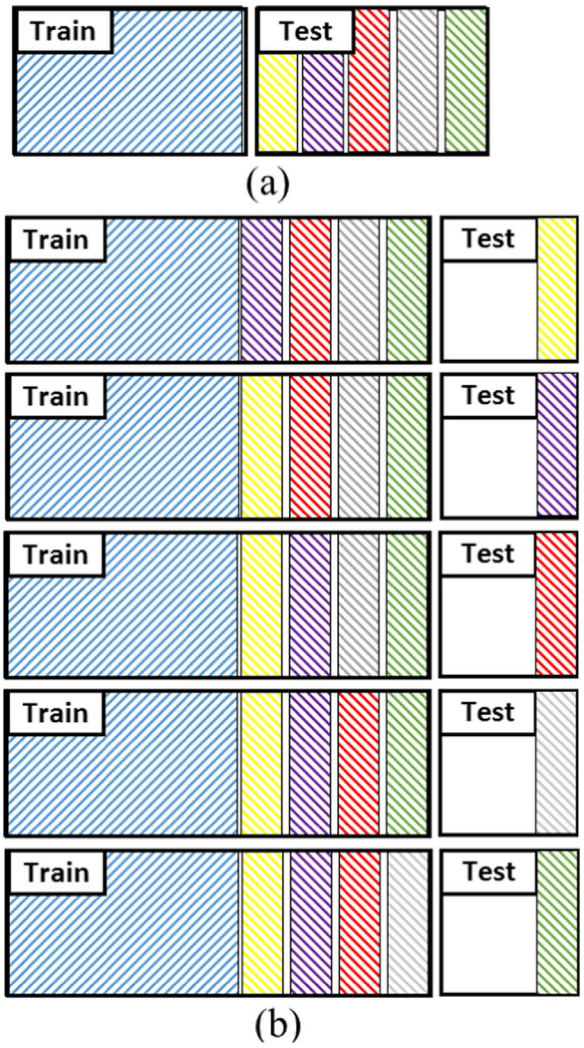


Fig. 4. Training and evaluation protocols used in our experiments. (a) Division protocol where the database is split into disjoint training and testing sets; (b) 5-fold protocol wherein the training set is the union of the training set of division protocol and 4/5 of the testing set of the division protocol. The testing set consists of the remaining 1/5 of the testing set of the division protocol.

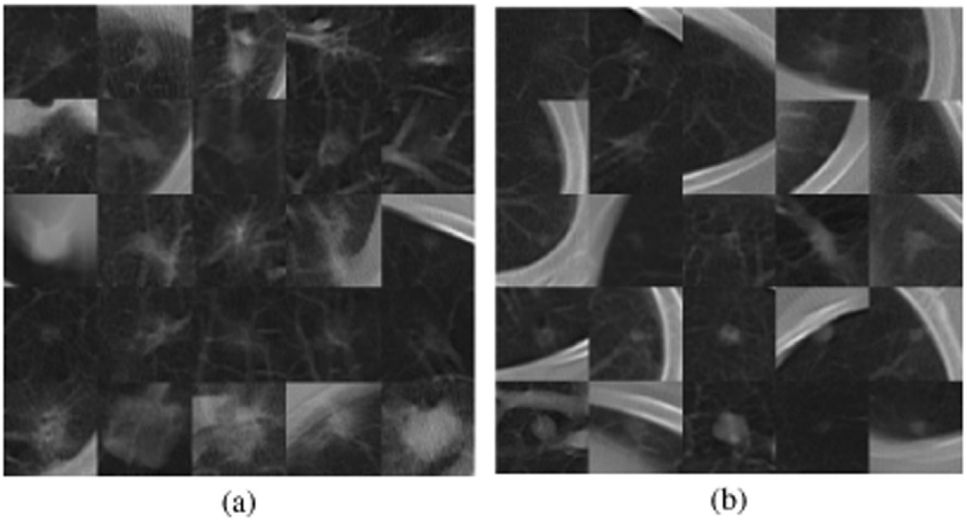


Fig. 3. Examples of (a) malignant nodules and (b) benign nodules in our database. Large variations of nodule patterns and size are seen, including non-solid, part-solid, and solid nodules of different sizes.

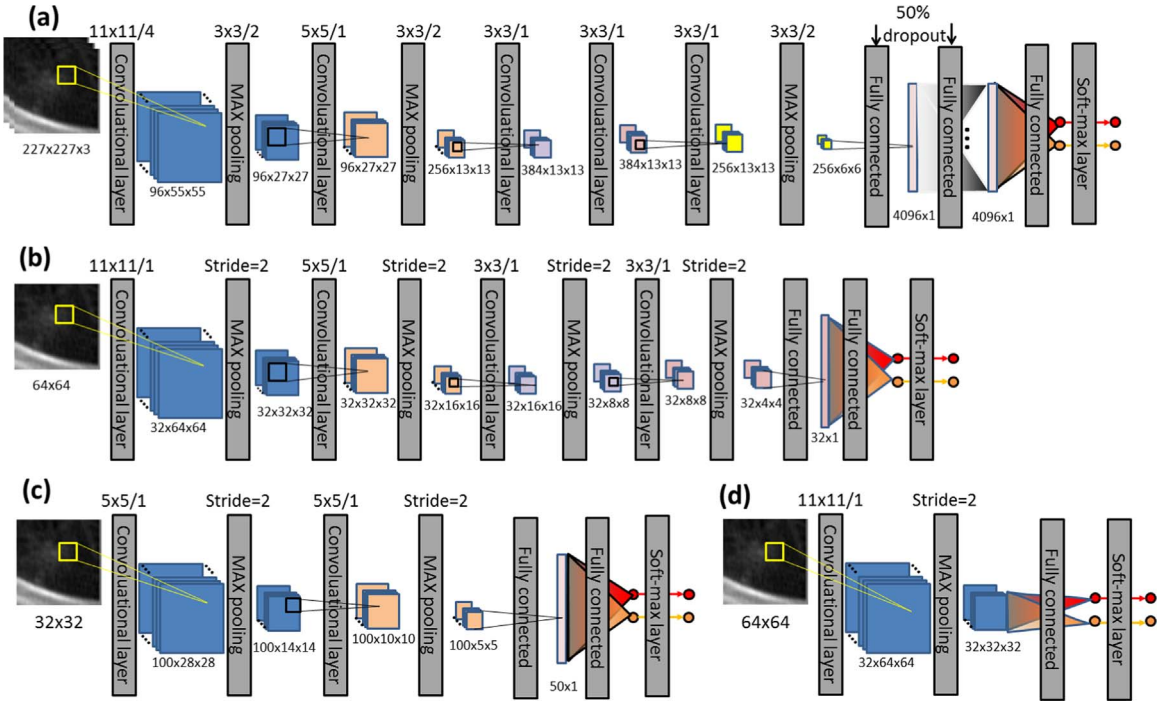


Fig. 5. Schematic overview of the CNN architectures used in our experiments. (a) A deep CNN (AlexNet), (b) a relatively deep (rd-CNN), (c) the LeNet architecture, and (d) a shallow CNN (sh-CNN).

and λ is the regularization parameter that helps prevent over-fitting when the number of weights is larger than the number of training samples. For this purpose, we used the Caffe library [40], which is arguably one of the most reliable and popular open source implementations of CNNs. For the AlexNet, in addition to training from scratch, we considered fine-tuning of the pre-trained AlexNet model that was available in Caffe. This pre-trained model had been trained using 1.2 million images labeled with 1000 semantic classes.

We conducted our experiments in 2 scenarios to rigorously evaluate the two machine learning models. We refer to the first scenario as "division" wherein the object proposal or lesion candidates were split into disjoint training and testing sets. In the division scenario, candidates were split according to the schemes suggested in [16,17], which reserved the majority of samples for testing. This evaluation scenario allows for a stringent evaluation of the two machine-learning models given limited training data. We refer to the second scenario as "5-fold" wherein the testing set of the division scenario was first divided into 5 disjoint subsets in the lesion-level; and then, the resulting subsets were added to the training set of the division scenario in a 5-fold cross validation manner. That is, 4 subsets were added to the training set; and the remaining subset was used for testing. This process was repeated 5 times, resulting in 5 classification models, each of which generating predictions for the corresponding testing subset. By collecting the predictions generated for each of the 5 test subsets, a performance curve was generated for the whole test set. Through this evaluation scenario, we can study how the increase in the size of training set impacts the performance of the classification models (both scenarios have the same testing set). Fig. 4 illustrates how the image data were divided into training and testing sets for the division and 5-fold scenarios.

For lung nodule detection, we first applied a base computer-aided detection (CADe) scheme [16], consisting of gray-level-based lung segmentation, feature extraction and analysis, and linear-discriminant-analysis-based classification, to the entire database. The CADe scheme generated 1128 candidates with 50 true positives and 1078 false positives. We collected patches from each candidate with data augmentation. Specifically, we extracted square patches at 6 scales, 40 translations from the candidate locations, and 8 orientations, resulting in a total of 1920 variations for each nodule candidate. Note that, due to relatively large thickness of slices, we chose to use gray-scale 2D patches for our experiments. The collected patches were then divided according to the division and 5-fold validation schemes for training and testing CNNs. The training set in the division scenario consisted of 10 true positives and 90 false positives (because of the 9 MTANNs in the ensemble), and the testing set consisted of 40 true positives and 988 false positives (non-nodules). For training a CNN in each scenario, we formed a stratified set of training patches by down-sampling the majority class (non-nodule). During the testing stage, the probability of each candidate being a nodule was computed as the average of probabilities assigned to the data-augmented patches. For performance comparison, we used free-response ROC (FROC) analysis.

Fig. 6 shows FROC curves for nodule detection. The confidence intervals for the FROC plots were computed according to the method suggested in [42]. Fig. 6(a) compares the performance of the MTANNs and CNNs in the division scenario. As can be seen, the performance of the MTANNs are higher than that of CNNs in most of the operating points with a significant margin ($p < 0.05$). As indicated by the overlapping error bars, the difference between the three CNN architectures was not significant at any of the operating points, which suggests that deep architectures become ineffective given limited training data. However, as shown in Fig. 6(b), the performance gap between MTANNs and CNNs becomes less evident when CNNs were trained and evaluated in the 5-fold cross validation scenario. The improved performance is attributed to the use of larger training sets available in the 5-fold cross validation scenario. Noteworthy, at 100% sensitivity, MTANNs generate 2.7 false positives per patient, which is significantly lower than the best performing CNN with 22.7 false positives per patient ($p < 0.05$). Fig. 6(c) compares the performance of each CNN architecture in the division and 5-fold scenarios. Clearly, the performance improvement is more substantial for deeper architectures, which suggests that deeper architectures can more effectively leverage the additional training instances than the shallower architectures.

Fig. 7 compares the top 30 "difficult" false positives generated by the MTANNs and the best performing CNN (fine-tuned AlexNet) for

Table 1
The CNN architectures used in our experiments. (a) A deep CNN (AlexNet), (b) a relatively deep CNN (rd-CNN), (c) the LeNet architecture, and (d) a shallow CNN (sh-CNN).

Layer	Type	Input	Kernel	Stride	Pad	Output
(a) AlexNet						
0	Input	227 ×227 ×3	N/A	N/A	N/A	227 ×227 ×3
1	Convolution	227 ×227 ×3	11 ×11	4	0	96 ×55 ×55
2	Max pooling	96 ×55 ×55	3 ×3	2	0	96 ×27 ×27
3	Convolution	96 ×27 ×27	3 ×3	1	2	256 ×27 ×27
4	Max pooling	256 ×27 ×27	3 ×3	2	0	256 ×13 ×13
5	Convolution	256 ×13 ×13	3 ×3	1	2	384 ×13 ×13
6	Convolution	384 ×13 ×13	3 ×3	1	2	384 ×13 ×13
7	Convolution	384 ×13 ×13	3 ×3	1	2	256 ×13 ×13
8	Max pooling	256 ×13 ×13	3 ×3	2	0	256 ×6 ×6
9	Fully connected	256 ×6 ×6	6 ×6	1	0	4096 ×1
10	Fully connected	4096 ×1	1 ×1	1	0	4096 ×1
11	Fully connected	4096 ×1	1 ×1	1	0	2 ×1
(b) rd-CNN						
0	Input	64 ×64	N/A	N/A	N/A	64 ×64
1	Convolution	64 ×64	11 ×11	1	5	32 ×64 ×64
2	Max pooling	32 ×64 ×64	2 ×2	2	0	32 ×32 ×32
3	Convolution	32 ×32 ×32	5 ×5	1	2	32 ×32 ×32
4	Max pooling	32 ×32 ×32	2 ×2	2	0	32 ×16 ×16
5	Convolution	32 ×16 ×16	3 ×3	1	2	32 ×16 ×16
6	Max pooling	32 ×16 ×16	2 ×2	2	0	32 ×8 ×8
7	Convolution	32 ×8 ×8	3 ×3	1	2	32 ×8 ×8
8	Max pooling	32 ×8 ×8	2 ×2	2	0	32 ×4 ×4
9	Fully connected	32 ×4 ×4	4 ×4	1	0	32 ×1
10	Fully connected	32 ×1	1 ×1	1	0	2 ×1
(c) LeNet						
0	Input	32 ×32	N/A	N/A	N/A	32 ×32
1	Convolution	32 ×32	5 ×5	1	5	100 ×28 ×28
2	Max pooling	100 ×28 ×28	2 ×2	2	0	100 ×14 ×14
3	Convolution	100 ×14 ×14	5 ×5	1	2	100×10 ×10
4	Max pooling	100 ×10 ×10	2 ×2	2	0	100 ×5 ×5
5	Fully connected	100 ×5 ×5	5 ×5	1	2	50 ×1
6	Fully connected	50 ×1	1 ×1	1	0	2 ×1
(d) sh-CNN						
0	Input	64 ×64	N/A	N/A	N/A	64 ×64
1	Convolution	64 ×64	11 ×11	1	5	32 ×64 ×64
2	Max pooling	32 ×64 ×64	2 ×2	2	0	32 ×32 ×32
3	Fully connected	32 ×1	32 ×32	1	0	2 ×1

lung nodule detection. The ROIs are shown in an increasing level of difficulty, namely, from the top left to bottom right ROI, the level of difficulty changes from easier to the most difficult to distinguish from nodules. The false positives of the MTANNs were collected at an operating point where 100% sensitivity was achieved. At this operating point, the MTANNs generated 85 false positives or equivalently 2.7 false positives per patient. Similarly, the operating point of the CNN was chosen as the point at which the CNN achieved 100% sensitivity with 713 false positives (22.7 false positives per patient). As seen from Fig. 7, the majority of the top false positives generated by the MTANNs and CNN include the chest wall, indicating that false positive sources tend to be near the chest wall. Some of the false positives look like lung nodules to our eyes, although they were not confirmed as lung nodules by radiologists in their reviews.

For lung nodule classification, we extracted square patches with data augmentation from the ROIs selected by an experienced chest radiologist. Specifically, we extracted patches at 3 scales, at 40 translations from the center of the ROI, and at 8 orientations, resulting in a total of 960 patches for each ROI. We divided the collected patches at the ROI-level according to the division and 5-fold cross validation schemes. That is, samples collected from a particular ROI were not distributed between the training and testing sets, rather, they were all assigned to either the training set or the testing set. The training set in the division scenario consisted of patches from 10 malignant nodules and 60 benign nodules (because of the 6 MTANNs in the ensemble), and the testing set consisted of the patches from 66 malignant nodules and 353 benign nodules. To avoid a bias towards the majority class (benign nodules), we formed a stratified set of training patches by down-sampling the majority class (benign). During the testing stage, the probability of each ROI being a malignant nodule was computed as the average of probabilities assigned to the patches that were extracted from the ROI with data augmentation. For performance comparison, we used ROC analysis.

Fig. 8 shows ROC curves for nodule classification. The confidence intervals for the ROC plots were computed according to the method suggested in [42]. Fig. 8(a) shows the ROC curves for the MTANNs and each of the CNN architectures. As can be seen, CNNs with varying depths performed comparably, yielding no significant performance improvement compared to each other. However, the MTANNs achieved a substantial improvement over the CNN-based systems, particularly at the optimal operating points located around the elbow of the ROC plots. Fig. 8(b) shows the ROC curves for the 5-fold scenario. As with the division scenario, CNN-based systems perform closely yet inferior to that of the MTANNs. For a quantitative comparison, we tabulate the area under the curve (AUC) with 95% confidence intervals for each model in Table 2. As can be seen, the MTANNs yield a significantly higher AUC than do the listed CNN-based models ($p < 0.5$), but as indicated by overlapping intervals, different variants of CNNs perform comparably. Fig. 8(c) compares the performance of each CNN architecture in the division and 5-fold scenarios. The increased number of training samples in the 5-fold scenario did not improve nodule classification performance significantly.

Fig. 9 shows the top 20 “easy-to-classify” and “hard-to-classify” malignant and benign nodules according to the malignancy scores produced by the MTANNs and the fine-tuned AlexNet in the 5-fold cross validation scenario. Fig. 9(a) and (b) shows easy-to-classify malignant nodules. As shown, both models, and to a larger degree the MTANNs, tend to assign a larger malignancy score to a larger malignant nodule. Fig. 9(c) and (d) shows easy-to-classify benign nodules. In this case, both models assign lower malignancy scores to smaller nodules. These observations suggest that the size of nodules has been considered by both models as a distinguishing feature for distinction between malignant and benign nodules. That makes sense, because malignant nodules tend to be larger than benign nodules in nature. However, as shown in Fig. 9(e) and (f), heavy reliance on the size of nodules has led to the misclassification of small malignant nodules as benign. Similarly, as shown in Fig. 9(e) and (f), large benign nodules were misclassified as malignant by both machine-learning models.

4. Discussion

In this study, we based the comparison between CNNs and MTANNs on the applications for which MTANNs had previously shown promising performance. This is because our study was to investigate whether the state-of-the-art deep learning machines such as the CNNs, which are well-established in the computer-vision field, could outperform the MTANNs as a well-established machine-learning model for medical vision tasks. This is an interesting research question because the CNNs are currently considered as a panacea that can outperform the previously suggested solutions for a variety of medical imaging applications. However, in this study, we demonstrated that the use of the CNNs was not as effective as the MTANNs for lung nodule detection and classification. Given the relatively similar characteristics of focal lesions, our conclusion may also generalize to similar tasks such as polyp detection, breast mass detection, and liver tumor detection, where the low-level and mid-level features captured by MTANNs are adequate for accurate detection. However, we would also like to emphasize that the conclusions reached in this paper may not generalize to more complex medical vision tasks such as image plane

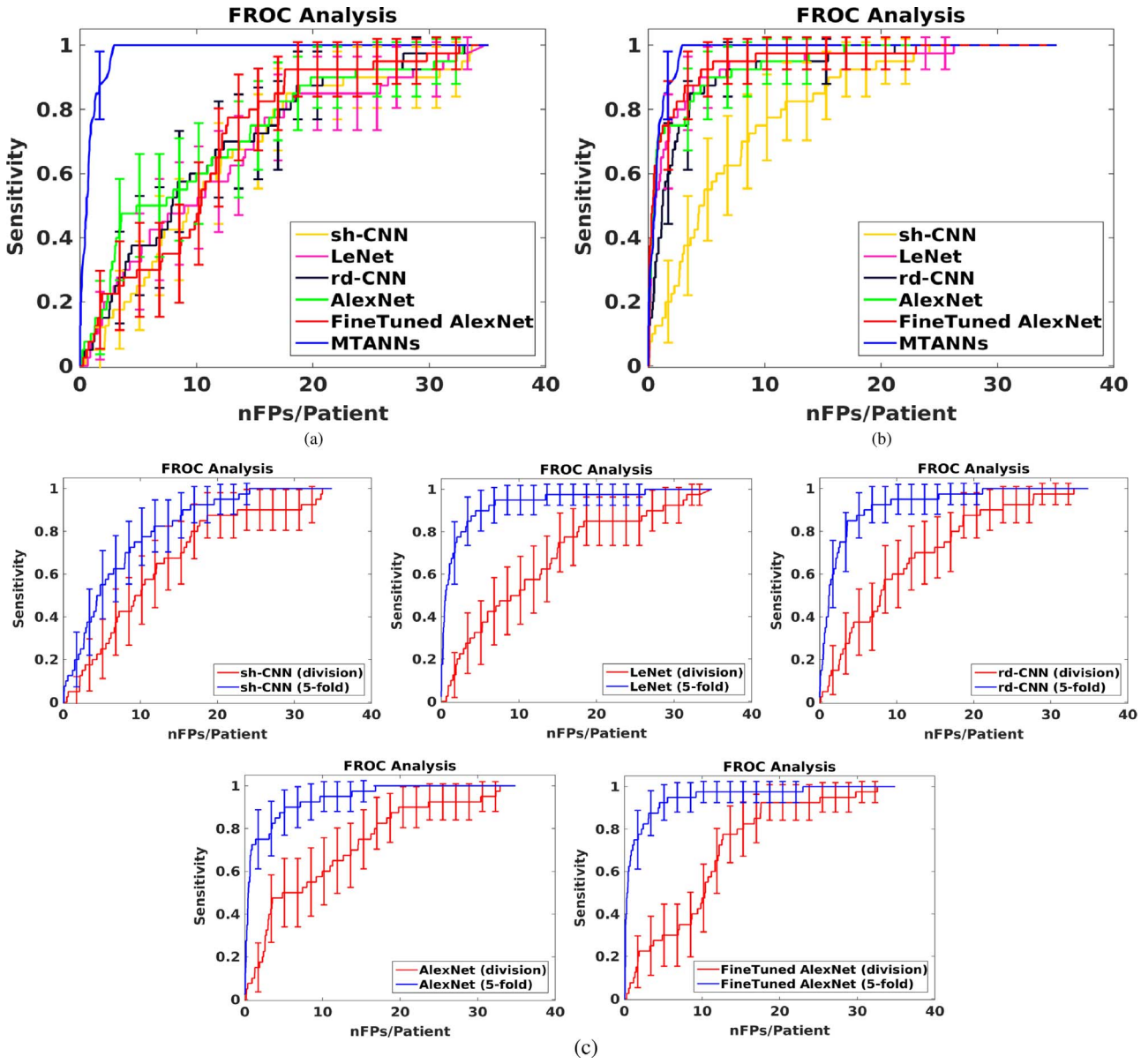


Fig. 6. Evaluation of lung nodule detection. Comparison between MTANNs and 3 CNN architectures in (a) the division protocol and (b) 5-fold protocol. (c) Performance comparison for each CNN architecture using the division and 5-fold protocols.

recognition [43] or pathology identification [44], where high-level semantic features extracted by deep CNNs are indispensable.

We conducted our evaluations in the division scenario where limited training data were used for training the CNNs and MTANNs,

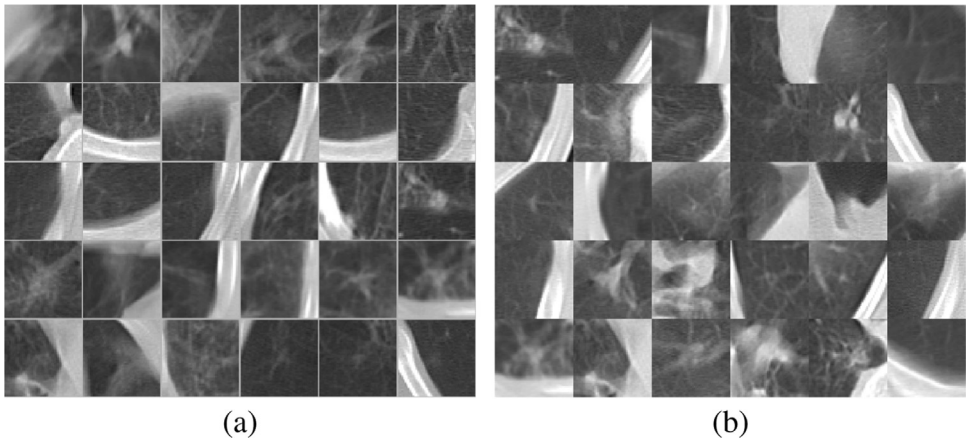


Fig. 7. Top 30 "difficult" false positives generated by the (a) MTANNs and (b) fine-tuned AlexNet for nodule detection in the 5-fold cross validation scenario.

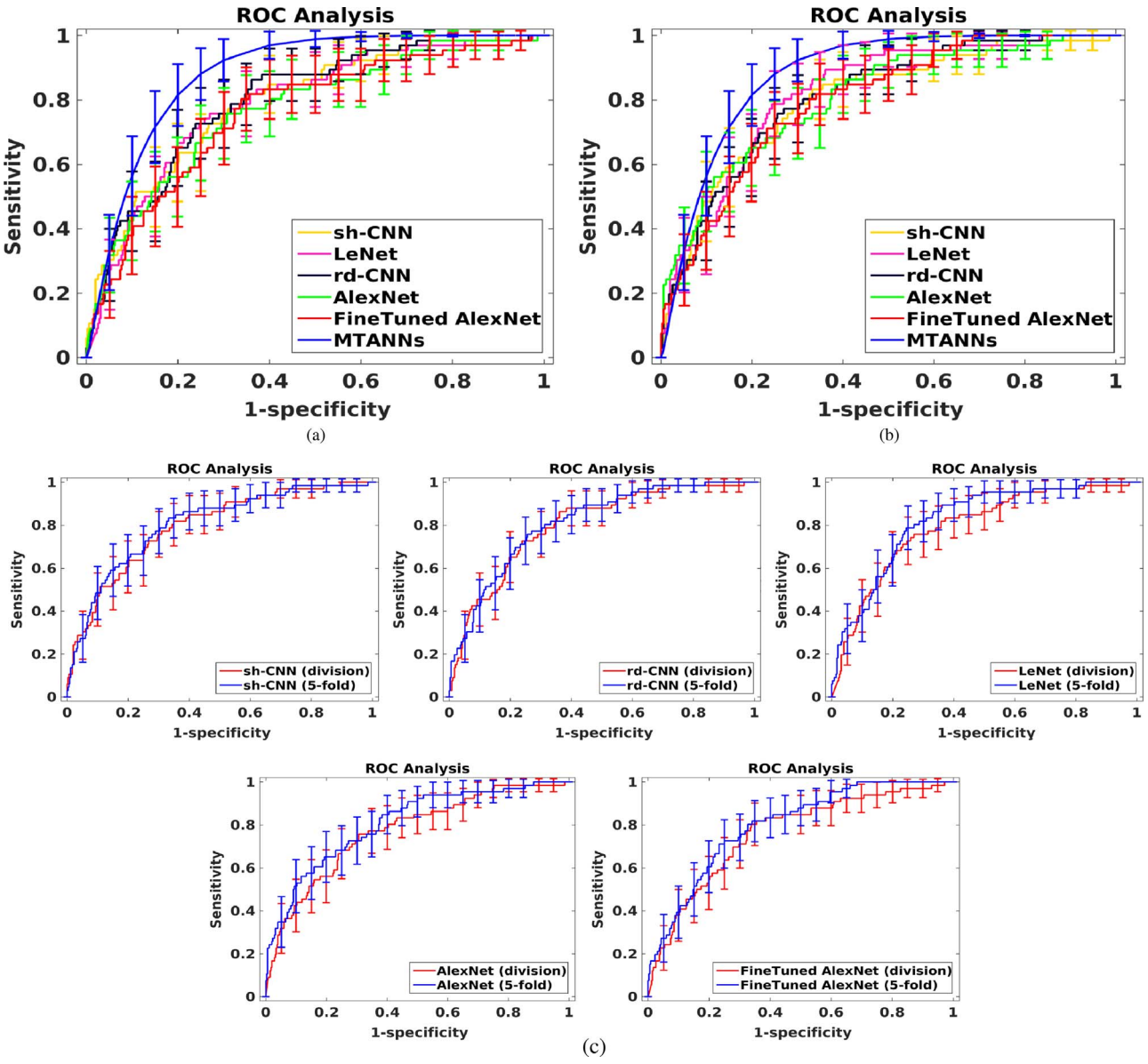


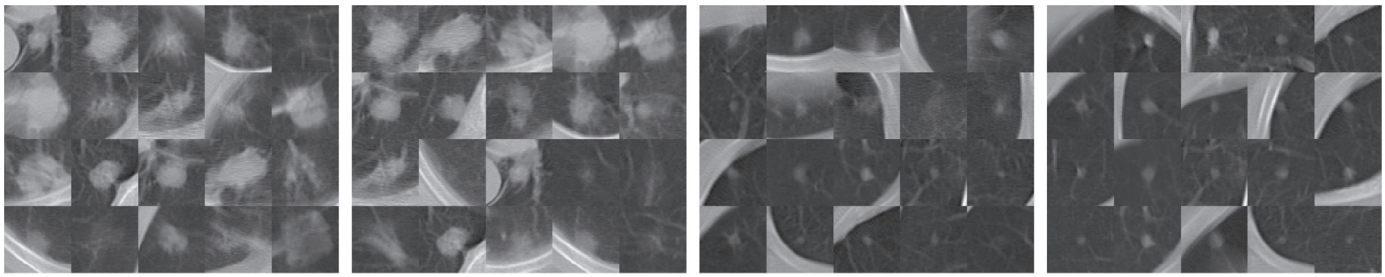
Fig. 8. Evaluation of lung nodule classification. Comparison between the MTANNs and 3 CNN architectures in (a) the division protocol and (b) 5-fold protocol. (c) Performance comparison for each CNN architecture using the division and 5-fold protocols.

Table 2
AUCs with 95% confidence intervals for the MTANNs and CNNs trained for nodule classification using the 5-fold cross validation.

Learning machine	AUC
MTANNs	0.8806 (95% CI: 0.8389–0.9223)
sh-CNN	0.7709 (95% CI: 0.7079–0.8272)
LeNet	0.7586 (95% CI: 0.6843–0.8140)
rd-CNN	0.7813 (95% CI: 0.7189–0.8306)
AlexNet	0.7685 (95% CI: 0.7025–0.8311)
FT AlexNet	0.7755 (95% CI: 0.7120–0.8270)

and also in the 5-fold cross validation scenario where a larger training set was employed for training. Our experiments demonstrated that the performance of the MTANNs was higher than that of the CNNs in the division scenario and to less extent in the 5-fold scenario. We also observed that the CNNs trained using the 5-fold cross validation substantially outperformed the CNNs trained in the division scenario for the nodule detection task. This observation was consistent with the common knowledge in the field where training deep learning machines such as CNNs require a large amount of labeled training data. For nodule classification, however, a change from the division to 5-fold scenario led to an insignificant performance gain for the CNNs. We attribute this phenomenon to the challenging nature of nodule classification. In fact, distinction between malignant and benign nodules using visual characteristics is a difficult task even for human experts. This is confirmed by an observer performance study [45] based on the same nodule database, which reported a low AUC of 0.56 for the average performance of five radiologists and an AUC of 0.63 when the radiologists were shown relevant (similar) examples of benign and malignant nodules provided by a content-based image retrieval system. Referring to Table 2, we find it interesting that both the CNNs and MTANNs achieved significantly higher performance than that of the average human observers.

To ensure that CNNs have been used to their full potentials when comparing against MTANNs, we explored 4 distinct CNN architectures and employed the well-established techniques to prevent the CNNs from over-fitting to the training data. Specifically, a shallow network was chosen for comparison because it required fewer training samples; and thus, it could be considered suitable, given the limited training

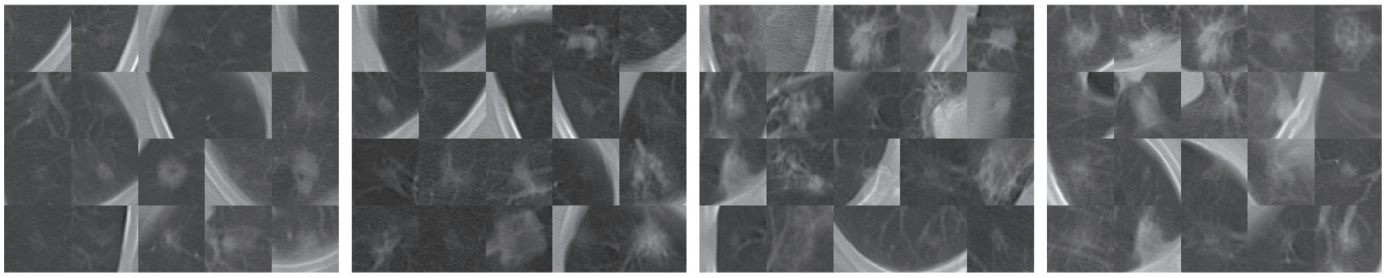


(a) Malignant nodules that received the highest to higher malignancy scores (true positives) by the MTANNs

(b) Malignant nodules that received the highest to higher malignancy scores (true positives) by the fine-tuned AlexNet

(c) Benign nodules that received the lowest to lower malignancy scores (true negatives) by the MTANNs

(d) Benign nodules that received the lowest to lower malignancy scores (true negatives) by the fine-tuned AlexNet



(e) Malignant nodules that received the lowest to lower malignancy scores ("difficult" false negatives) by the MTANNs

(f) Malignant nodules that received the lowest to lower malignancy scores ("difficult" false negatives) by the fine-tuned AlexNet

(g) Benign nodules that received the highest to higher malignancy scores ("difficult" false positives) by the MTANNs)

(h) Benign nodules that received the highest to higher malignancy scores ("difficult" false positives) by the fine-tuned AlexNet

Fig. 9. Compassion between easy- and hard-to-classify malignant and benign nodules. According to panels (a)–(d), both machine-learning models use the size of nodules as a characteristic feature to distinguish between malignant and benign nodules. However, heavy reliance on the size of nodules has led to misclassification of small malignant nodules as benign (see panels (e) and (f)) and misclassification of large benign nodules as malignant (see panels (g) and (h)).

datasets in medical imaging applications. The relatively deep CNN and LeNet were chosen for comparison because they are common CNN architectures in the medical imaging literature. The deep AlexNet architecture was also explored because it could demonstrate how the record-breaking AlexNet architecture would perform for focal lesion detection and classification. To avoid over-fitting, we used weight regularization in the cost function, extensively performed data augmentation to enrich the datasets, and adopted dropout regularization to hinder over-fitting in the parameter-rich fully connected layers. As an alternative to training from scratch, we also explored fine-tuning of a pre-trained model, which is arguably one of the best if not the best technique for training a CNN from a limited training set. Therefore, we conclude that the competitive performance evaluation, which we presented in this paper, may draw a fair comparison between MTANNs and CNNs.

Our experiments based on CNN architectures of varying depths provided insights into the effective depth of a CNN for focal lesion detection and classification. For nodule detection, we observed that the shallow CNN with 1 convolutional layer performed on a par with the deeper CNNs given a limited training set. However, the performance of the shallow CNN was significantly lower than that of the deeper architectures when a larger training set was used for training. Furthermore, comparing the performance of the two deep architectures (Alexnet and rd-CNN) revealed that the AlexNet offered no significant performance improvement over the rd-CNN, suggesting that rd-CNN with 4 convolutional layers may be adequate for nodule detection. For nodule classification, our experiments showed no significant performance improvement when using CNN architectures of varying depths. This performance trend is probably caused by similar visual characteristics of the benign and malignant nodules, which leaves little room for further improvement by means of deeper architectures. In contrast, nodule detection consists in distinguishing nodules from a large variety of normal structures, for which deeper architectures have more capability than do shallow architectures. These findings suggested that the use of deep architectures, which is fundamental to achieving high performance in the field of computer vision, might not be as effective for focal lesion detection and classification in medical images. This may call for a systematic study on the effective depth of CNNs for medical imaging applications.

To provide a deeper comparison between CNNs and MTANNs, we would like to further discuss these two learning machines in terms of learned feature hierarchies, handling of sample uncertainty, and reliance on the size of the training set:

- *Learned feature hierarchies*: Consider an MTANN that has a hidden layer with k nodes and an output layer with 1 output node. Further assume that the above MTANN is trained using $d \times d$ image patches. During the training phase, $k \times d \times d$ weights will be learned between the input and the hidden layer, which can be viewed as learning k filters of size $d \times d$; k weights will be learned between the hidden layer and the output layer, which can be viewed as learning a fusion rule for combining the k filter responses in order to produce the desired output. Therefore, each MTANN in an ensemble learns k low-level features and the corresponding combination rule. During the testing phase, the likelihood of being a lesion is computed for small $d \times d$ regions in an ROI; and then, the resulting likelihood values for lesion parts are aggregated using a Gaussian function. Therefore, the MTANNs never attempt to detect the entire lesion at once, rather, they learn to detect lesion parts using low-level features; and then, aggregate the predictions in a weighted manner to compute an overall lesion score for the whole ROI. CNNs, on the other hand, not only learn the low-level features but also extract the mid-level and high-level features in order to produce a likelihood value for a given ROI. The distinction between the above-explained featured hierarchies can contribute to the higher performance of the MTANNs for focal lesion detection. In Fig. 7, we showed that the

majority of the top false positives generated by the MTANNs and CNN appeared around the chest wall. However, examining the remaining top false positives generated by the two models reveals that the CNN-based model shows a higher degree of reliance in the presence of chest walls than does the classification model based on the MTANNs. This is because a CNN tends to extract semantic features from ROIs. This together with the presence of the chest wall in the training ROIs has led to the inclusion of chest walls as a characteristic feature of the lung nodules. On the other hand, the MTANNs tend to extract the low-level features and use a Gaussian function centered in the middle of each ROI for weighted averaging of the scores, which makes the MTANNs more agnostic to image information located in the border areas (chest wall).

- *Handling of sample uncertainty*: In Sections 2.1 and 2.2, we explained that CNNs take the original ROIs with the corresponding binary labels for training, but MTANNs receive small image patches from the ROIs and the corresponding continuous outputs ranged between 0 and 1. Therefore, the CNNs learn a classification model, but MTANNs learn a regression model. This seemingly subtle difference can contribute to the superior performance of the MTANNs. This is because a CNN treats all the training samples equally regardless of their levels of uncertainty; as a result, the decision boundary can be readily affected by the presence of hard-to-classify training samples [46]. One way to overcome this limitation is to score the ROIs according to their level of difficulty. However, a manual approach is both subjective and expensive, and an automatic approach is still an immature area of research [47]. In the MTANNs, however, the uncertainty associated with difficult training samples is embedded in the continuous teaching values. Basically, the positive patches that are selected farther from the lesion location are considered hard-to-classify samples; and thus, they receive smaller likelihoods of being a lesion. Therefore, the proper handling of uncertainty in the MTANNs could contribute to its higher performance.
- *Reliance on the size of the training set*: MTANNs differ from CNNs in that they require only a small number of training ROIs. This is indeed a major advantage of the MTANNs over CNNs, which makes them particularly suitable for medical imaging applications where it is difficult and expensive to obtain a large number of labeled training data. This advantage stems from the fact that the MTANNs do not learn directly from the ROIs, rather, from small image patches, which can be effortlessly collected from a few training ROIs [48]. Furthermore, modeling appearance variability in small image patches is relatively less challenging than that of large ROIs, further decreasing the number of samples required for training the MTANNs. In contrast, the CNNs directly learn from the training ROIs; and thus, one must obtain a large number of training ROIs through either acquiring new cases or performing data augmentation. The former is difficult and expensive because lesions are not found frequently in medical images, and because they require expert annotations. The latter is computationally cheap, but the resulting training samples are highly correlated. Therefore, it can only partially compensate for insufficient training ROIs. This can be seen in Fig. 6(b) where the inclusion of additional unique training ROIs in the 5-fold cross validation scenario substantially improved the performance of the CNNs trained in the division scenario, even though sufficient data augmentation had been performed in the division scenario.

5. Conclusion

In this paper, we have compared 2 classes of end-to-end machine-learning models, namely MTANNs and CNNs. For this purpose, we considered 2 well-studied topics in the field of medical imaging: detection of lung nodules and distinction between benign and malignant nodules in low-dose CT images. We conducted our experiments in 2 scenarios. In the first scenario, we compared the performance of the CNNs and MTANNs after being trained using limited training data. Our experiments showed that the performance of the MTANNs was higher than that of the CNNs for both lung nodule detection and classification. In the second scenario, we used large training datasets for training the CNNs. We observed a lower performance gap between the two models, but the difference was still significant. Specifically, for nodule detection, the MTANNs generated 2.7 false positives per patient at 100% sensitivity, which was significantly ($p < 0.05$) lower than the best performing CNN model with 22.7 false positives per patient at the same level of sensitivity. For nodule classification, the MTANNs yielded an AUC of 0.8806 (95% CI: 0.8389–0.9223), which was significantly ($p < 0.05$) higher than the best performing CNN model with an AUC of 0.7755 (95% CI: 0.7120–0.8270). We further theoretically compared the MTANNs and CNNs and discussed the possible reasons for the superiority of the MTANNs.

Acknowledgment

The authors are grateful to the members of the Suzuki Lab for their valuable discussions, and to Shusuke Sone, MD, and Feng Li, MD, for the use of the database.

References

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[2] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on Statistical Learning in Computer Vision, ECCV, 2004, pp. 1–22.
[3] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[4] Y. Kim, Y. Jernite, D. Sontag, A.M. Rush, Character-aware neural language models, arXiv:1508.06615.
[5] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Advances in Neural Information Processing Systems, 2014, pp. 2042–2050.
[6] T. Unterthiner, A. Mayr, G. Klambauer, S. Hochreiter, Toxicity prediction using deep learning, arXiv:1503.01445.
[7] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, V. Pande, Massively multitask networks for drug discovery,arXiv:1502.02072.
[8] I. Wallach, M. Dzamba, A. Heifets, Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery, arXiv:1510.02855.
[9] N. Tajbakhsh, M.B. Gotway, J. Liang, Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks, in: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, 2015.
[10] N. Tajbakhsh, S.R. Gurudu, J. Liang, A comprehensive computer-aided polyp detection system for colonoscopy videos, in: International Conference on Information Processing in Medical Imaging, Springer, 2015, pp. 327–338.
[11] H.R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, C. Kevin, L. Kim, R.M. Summers, Improving computer-aided detection using convolutional neural networks and random view aggregation,arXiv:1505.03046.
[12] H.R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, R. M. Summers, Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Springer, 2015, pp. 556–564.
[13] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: fine tuning or full training?, IEEE Trans. Med. Imaging 35 (5) (2016) 1299–1312.
[14] K. Suzuki, I. Horiba, N. Sugie, Efficient approximation of neural filters for removing quantum noise from images, IEEE Trans. Signal Process. 50 (7) (2002) 1787–1799.
[15] K. Suzuki, I. Horiba, N. Sugie, M. Nanki, Neural filter with selection of input features and its application to image quality improvement of medical image sequences, IEICE Trans. Inf. Syst. 85 (10) (2002) 1710–1718.
[16] K. Suzuki, S.G. Armato III, F. Li, S. Sone, K. Doi, Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography, Med. Phys. 30 (7) (2003) 1602–1617.
[17] K. Suzuki, F. Li, S. Sone, et al., Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network, IEEE Trans. Med. Imaging 24 (9) (2005) 1138–1150.
[18] K. Suzuki, H. Yoshida, J. Näppi, A.H. Dachman, Massive-training artificial neural network (MTANN) for reduction of false positives in computer-aided detection of polyps: suppression of rectal tubes, Med. Phys. 33 (10) (2006) 3814–3824.
[19] K. Suzuki, H. Yoshida, J. Näppi, S.G. Armato III, A.H. Dachman, Mixture of expert

3D massive-training ANNs for reduction of multiple types of false positives in CAD for detection of polyps in CT colonography, Med. Phys. 35 (2) (2008) 694–703.
[20] K. Suzuki, D.C. Rockey, A.H. Dachman, CT colonography: : advanced computer-aided detection scheme utilizing MTANNs for detection of "missed" polyps in a multicenter clinical trial, Med. Phys. 37 (1) (2010) 12–21.
[21] K. Suzuki, J. Zhang, J. Xu, Massive-training artificial neural network coupled with Laplacian-eigenfunction-based dimensionality reduction for computer-aided detection of polyps in CT colonography, IEEE Trans. Med. Imaging 29 (11) (2010) 1907–1917.
[22] J.-W. Xu, K. Suzuki, Massive-training support vector regression and gaussian process for false-positive reduction in computer-aided detection of polyps in CT colonography, Med. Phys. 38 (4) (2011) 1888–1902.
[23] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (4) (1980) 193–202.
[24] K. Fukushima, Neocognitron capable of incremental learning, Neural Netw. 17 (1) (2004) 37–46.
[25] S. Deutsch, A simplified version of Kunihiko Fukushima's neocognitron, Biol. Cybern. 42 (1) (1981) 17–21.
[26] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.
[27] B.B. Le Cun, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Handwritten digit recognition with a back-propagation network, in: Advances in Neural Information Processing Systems, Citeseer, 1990.
[28] N. Tajbakhsh, S. R. Gurudu, J. Liang, Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 79–83.
[29] W. Shen, M. Zhou, F. Yang, C. Yang, J. Tian, Multi-scale convolutional neural networks for lung nodule classification, in: S. Ourselin, D.C. Alexander, C.-F. Westin, M.J. Cardoso (Eds.), International Conference on Information Processing in Medical Imaging, Lecture Notes in Computer Science, vol. 9123, Springer International Publishing, 2015, pp. 588–599.
[30] H. Roth, L. Lu, A. Seff, K. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, R. Summers, A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations, in: P. Golland, N. Hata, C. Barillot, J. Hornegger, R. Howe (Eds.), International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2014, Lecture Notes in Computer Science, vol. 8673, Springer International Publishing, 2014, pp. 520–527.
[31] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, D. Shen, Deep convolutional neural networks for multi-modality isointense infant brain image segmentation, NeuroImage 108 (2015) 214–224.
[32] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks,arXiv:1505.03540.
[33] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.
[34] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, arXiv:1502.01852.
[35] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[37] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 580–587.
[38] S. Sone, S. Takashima, F. Li, Z. Yang, T. Honda, Y. Maruyama, M. Hasegawa, T. Yamanda, K. Kubo, K. Hanamura, et al., Mass screening for lung cancer with mobile spiral computed tomography scanner, Lancet 351 (9111) (1998) 1242–1245.
[39] F. Li, S. Sone, H. Abe, H. MacMahon, S.G. Armato, K. Doi, Lung cancers missed at low-dose helical CT screening in a general population: comparison of clinical, histopathologic, and imaging findings 1, Radiology 225 (3) (2002) 673–683.
[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, arXiv:1408.5093.
[42] D.C. Edwards, M.A. Kupinski, C.E. Metz, R.M. Nishikawa, Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model, Med. Phys. 29 (12) (2002) 2861–2870.
[43] J. Margeta, A. Criminisi, R. Cabrera Lozoya, D.C. Lee, N. Ayache, Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition, Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. (2015) 1–11.
[44] Y. Bar, I. Diamant, L. Wolf, H. Greenspan, Deep learning with non-medical training used for chest pathology identification, in: SPIE Medical Imaging, International Society for Optics and Photonics, 2015, p. 94140V.
[45] Q. Li, F. Li, J. Shiraishi, S. Katsuragawa, S. Sone, K. Doi, Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules, Med. Phys. 30 (10) (2003) 2584–2593.
[46] M. Kobetski, J. Sullivan, Improved boosting performance by explicit handling of ambiguous positive examples, in: Pattern Recognition Applications and Methods, Advances in Intelligent Systems and Computing, vol. 318, 2015, pp. 17–37.
[47] J. Ba, R. Caruana, Do deep nets really need to be deep?, in: Advances in Neural Information Processing Systems, 2014, pp. 2654–2662.
[48] K. Suzuki, K. Doi, How can a massive training artificial neural network (MTANN) be trained with a small number of cases in the distinction between nodules and vessels in thoracic CT?, Acad. Radiol. 12 (10) (2005) 1333–1341.

Nima Tajbakhsh is currently a Senior Research Associate at Illinois Institute of Technology. Prior to this, he earned Ph.D. in Biomedical Informatics from Arizona State University, in 2015. He has developed several computer-aided detection systems for various imaging modalities ranging from CT, and echocardiography to endoscopy videos. His current research interests lie at the applications of deep learning in medical imaging.

Kenji Suzuki worked at Hitachi Medical Corp., Aichi Prefectural Univ., and Univ. of Chicago. Since 2014, he has been an Associate Professor at Illinois Institute of Technology. He published 300 papers (including 110 journal papers) in machine learning in medical imaging and computer-aided diagnosis, served on Editorial of a number of journals, and edited 12 journal special issues.