**Noname manuscript No.**
(will be inserted by the editor)

# Pulmonary nodule classification with deep residual networks

**Aiden Nibali · Zhen He · Dennis Wollersheim**

Received: date / Accepted: date

**Abstract**
*Purpose* Lung cancer has the highest death rate amongst all cancers in the US. In this work we focus on improving the ability of computer-aided diagnosis (CAD) systems to predict the malignancy of nodules from cropped CT images of lung nodules.
*Methods* We evaluate the effectiveness of very deep convolutional neural networks at the task of expert-level lung nodule malignancy classification. Using the state-of-the-art ResNet architecture as our basis, we explore the effect of curriculum learning, transfer learning, and varying network depth on the accuracy of malignancy classification.
*Results* Due to a lack of public datasets with standardized problem definitions and train/test splits, studies in this area tend to not compare directly against other existing work. This makes it hard to know the relative improvement of the new solution. In contrast, we directly compare our system against two state-of-the-art deep learning systems for nodule classification on the LIDC/IDRI dataset using the same experimental setup and data set. The results show that our system achieves the highest performance in terms of all metrics measured including sensitivity, specificity, precision, AUROC, and accuracy.
*Conclusions* The proposed method of combining deep residual learning, curriculum learning, and transfer learning translates to high nodule classification accuracy. This reveals a promising new direction for effective pulmonary nodule CAD systems that mirrors the success of recent deep learning advances in other image-based application domains.

**Keywords** Lung nodule · Convolutional neural network · CT images

## 1 Introduction

Lung cancer has the highest death rate amongst all cancers for both men and women in the US [1], and is one of the leading causes of human mortality worldwide [8]. Computer-aided diagnosis (CAD) systems have the potential to offer a significant boost to the feasibility of computed tomography (CT) based screening programs by helping radiologists make correct classification decisions and reducing costs incurred by manually reading scans. In this paper we address the specific task of pulmonary nodule classification according to subjective human expert consensus as a means of improving the capabilities of CAD systems for lung cancer screening support.

This paper focuses on classifying lung nodules directly from cropped CT images without segmentation or hand-crafted features. In contrast, a common approach taken by existing automatic nodule classification systems is to 1) segment the nodule, 2) extract hand-crafted morphological and/or statistical features, and 3) classify the nodule based on these features. The exact details of this procedure varies somewhat.

There are many existing works that extract hand engineered features from segmented images [15,29,13] or non segmented images [16] and then feed them into some kind of classifier like support vector machines [16], decision trees [29], fully connected neural networks [13], or a classifier ensemble [15]. In contrast, Kumar et al. [12] proposed using a fully connected autoencoder to

A. Nibali
Department of Computer Science and Computer Engineering,
La Trobe University, Melbourne, Australia
E-mail: anibali@students.latrobe.edu.au

learn features automatically from nodule images. Unfortunately, the results were not favorable in light of many of the other related works that use hand-crafted features.

Another approach to nodule classification is the use of convolutional neural networks (CNNs) [14]. CNNs have state-of-the-art results in a wide variety of machine learning tasks such as estimating human poses [27], processing natural language [10], and playing Go [21]. A recent trend in CNN-related research has been to increase the *depth* of models (the number of weighted layers), with the increased depth yielding more accurate results in such models as VGG [22] and ResNet [5]. Deeper neural networks are thought to have increased representational power [25], which we explore in the context of nodule classification.

A significant advantage of using CNNs is that they remove the need for any kind of hand-crafted feature engineering from images, and instead learn discriminative features from the data directly. A few attempts have been made to classify pulmonary nodules using shallow CNN architectures [20, 7, 3, 19]. Shen et al. [20] were able to successfully classify malignant lung nodules using a 2-layer convolutional neural network on multiple crops of the nodule at different scales. Hua et al. [7] also use a shallow CNN with only 2 convolutional layers to perform classification, but include 2 fully connected layers before the network output. Ciompi et al. [3] perform classification of peri-fissural nodules with an existing pretrained CNN called OverFeat [18] (with 8 weighted layers). Multiple views of the nodule (axial, coronal and sagittal) are evaluated with OverFeat, and the posterior distributions combined to produce a final prediction. Setio et al. [19] perform classification using 9 separate CNNs (each with 3 convolutional layers) on different nodule views (axial, coronal, sagittal, and 6 diagonal planes) to determine nodule presence. The final classification result is obtained by fusing the CNN outputs with fully connected layers.

There are a couple of major difficulties associated with using CNNs to perform lung nodule classification. Firstly, the publicly available datasets for the task are small (hundreds or thousands of examples) when compared to other image classification datasets (up to millions of examples). Secondly, the differences between examples in the nodule classification task are subtle, whereas most existing CNN classifiers deal with classes that are much more visually distinct (eg. dogs and cars). These challenges lead us to three research questions explored in this paper:

*1) Does increasing the depth of neural networks help with the task of lung nodule classification?* In most conventional object recognition tasks there exists a well-defined hierarchy of concepts (eg. edge → wheel → car) which can intuitively benefit from network depth. It is not clear whether CT scans contain a hierarchy that is sufficiently rich to benefit from very deep networks. To explore this idea, our architecture is much deeper than existing lung nodule classification systems using CNNs [20, 7, 3, 19]. This increased depth is made possible by recent advancements such as batch normalization [9] and residual learning with ResNets [5].

*2) How can we leverage transfer learning to achieve higher accuracy?* As humans we bring a wealth of prior knowledge and experience to every new task that we encounter. In contrast, neural networks are often expected to learn tasks entirely from scratch, based on a completely random initialization of weights. One approach to bridging this initial knowledge gap is *transfer learning*, in which the network is pretrained on a completely different task. The weights learned during pretraining then become a starting point for learning the desired task. We experiment with this idea by pretraining on CIFAR-10, a well-known and large image classification dataset, before training on the smaller nodule dataset.

*3) Can accuracy be improved by using a training curriculum?* The work of Bengio et al. [2] suggests that gradually increasing the difficulty of examples as training progresses can be beneficial. In this paper we describe a way to quantify the difficulty of an example for the task of nodule classification, and use this definition to create a training curriculum. Our results show that this approach improved the accuracy of nodule classification.

In addition to the above research questions, we also observe that there is currently a lack of comparison between different neural network architectures for lung CT analysis. Existing papers that use CNNs for analyzing lung nodules [7, 20, 3, 19] do not compare their results against each other on the same dataset and experimental setup. It is important to address this omission due to significant variations in CNN architectures and training strategies. Unless the algorithms are compared under the same experimental conditions we can not objectively conclude certain architectural choices and training algorithms are better than others.

We present two main sets of results. Firstly, we report on the positive influences that deeper networks, transfer learning, and curriculum learning have on the classification accuracy of our system. Secondly, we objectively compare our system against implementations of existing models under the same training and test conditions: 1) a simple baseline fully connected neural network, 2) a strong state-of-the-art rival convolutional neural network [19], and 3) an OverFeat network [18] which has proven performance in other image classification tasks. The best variant of our system achieved a classification accuracy of 89.90% on test examples taken from the LIDC/IDRI dataset [23]. In contrast, the best existing neural network architecture compared against (a shallow convolutional neural network from [19]) achieved an accuracy of 86.23%. In addition the results show that our system achieves the highest performance in terms of all metrics measured including, sensitivity, specificity, precision, AUROC and accuracy.

## 2 Materials and methods

### 2.1 The dataset

Our experimental dataset was derived from the publicly available LIDC/IDRI dataset [23]. LIDC contains CT scans for 1,010 patients, with each patient assessed by four radiologists to produce four sets of subjective nodule readings. Readings for the same nodule were grouped according to the Cornell LIDC nodule size report [17]. We averaged the malignancy ratings provided by at least three radiologists to produce a median rating from 1 to 5, which was treated as ground truth. Binary malignancy labels were derived by treating any example with malignancy rating above 3 as a positive example. Nodules with borderline median malignancy (rating = 3) were discarded.

We use the subjective LIDC malignancy ratings due to the lack of large-scale, publicly available, objective labels. However, should such a dataset become available, we anticipate that the difficulty of predicting objective labels would be similar since human experts tend to use the features most indicative of malignancy to arrive at their subjective ratings.

After preprocessing we were left with 831 examples of nodules, 50.66% of which were positive examples. The examples ranged in diameter from 3 mm to 42 mm according to nodule boundaries annotated by the radiologists. The original CT slice images were combined to form a three-dimensional volume of voxels from which arbitrary two-dimensional planes could be sliced. The voxel values were normalized using a fixed linear transformation, such that -1 and 1 correspond to -2048 and 4096 in Hounsfield units respectively. This normalization procedure preserves the significance of the Hounsfield scale whilst providing the network with a more easily digestible numeric range. Following the lead of Ciompi et al. [3], three orientations of extracted planes were considered: axial, sagittal and coronal. Each two-dimensional plane covered a 45 mm × 45 mm area centered on the nodule in question, and was represented as a 64 pixel × 64 pixel greyscale image. Since the CT scans did not have consistent spacing between pixels and slices, resampling with trilinear filtering was performed where necessary to keep scale of a voxel consistent across examples. Figure 1 depicts several cropped nodule CT images, which are representative of typical inputs provided to the neural network. Note that nodule segmentation (such as a contour around the nodule) was not used as part of the input to the network.

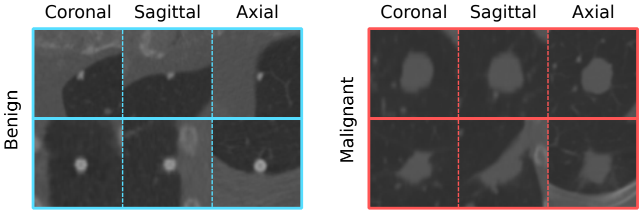

**Figure 1** Examples of two benign and two malignant nodules from the dataset. Each nodule volume was preprocessed to create a coronal, sagittal, and axial planar view, as shown.

### 2.2 Deep residual networks

In this section we provide a more detailed description of the deep residual network (ResNet) [5], as it is fundamental to our proposed solution. The main benefit of using ResNets is the ability to train deep networks with dozens of weighted layers. As the depth of a network increases it becomes increasingly difficult for the gradients to backpropagate from the loss function to the various layers without either diminishing to zero or exploding. ResNets allow gradients to pass unattenuated through layers by using an identity (skip) connection. Figure 2 shows a basic residual network block with the inclusion of the identity connection. The residual block will learn the following function.

$$H(x) = F(x) + x \tag{1}$$

The residual function $F(x)$ is learnt by training the weight layers using labeled data. The weight layers can consist of any type of neural network layer including convolutional or fully connected layers. The residual block allows the forward pass through the network to selectively skip over certain parts of the network by setting $F(x)$ equal to zero for those parts. This makes it possible to build a really deep network consisting of many different layers of feature extractors, each capturing a different possible characteristic of the data. For any given input instance, only the parts of the network

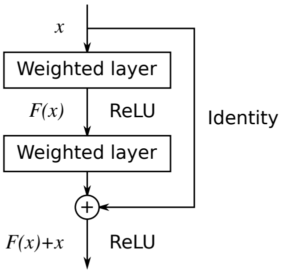

**Figure 2** A basic residual network block [5].

that are relevant to classifying that particular instance are turned on.

## 2.3 Our model

We apply the basic principles of the residual network (ResNet) to the problem of lung nodule classification. However, our ResNet implementation has several differences to the original. Firstly, it accepts a 64 pixel × 64 pixel greyscale input, and the initial convolution layer has been adjusted to accommodate this. Secondly, the number of feature maps has been lowered throughout the network, which allows us to fit the 3-column network in GPU memory without sacrificing minibatch size or network depth. ~~Thirdly, the network is~~Thirdly, the ResNet is modified to be "fully convolutional" and does not contain any fully connected layers. In general, this reduces the number of learnable parameters and increases translational invariance.

We tested a number of different model configurations. Figure 3 shows a specific example of our model with 3 columns (sequences of network layers extending horizontally from the input ). However, in general $n$ columns can be used. Each column is a modified instance of ResNet, denoted $f_k(x_k, \theta_k)$, where $x_k$ is the $k^{th}$ planar view of the nodule and $\theta_k$ denotes the column's parameters. The final nodule malignancy prediction, $y$, is obtained via a weighted sum of column outputs (Equation 2). $c_k$ is the learnt importance weighting for column $k$. A sigmoid function is used to squash the final output between 0 and 1, which will always be a valid probability.

$$y = \sigma(\sum_k c_k f_k(x_k, \theta_k)) \tag{2}$$

One way to consider the arrangement of multiple columns is as an ensemble of $n$ models, where each model operates on a different planar view of the nodule.

The main structural element used in our ResNet implementation is the revised residual block proposed by He et al. [6] as a follow-up to the original ResNet paper (refer to the "zoomed-in" portion of Figure 3). A residual block contains two stacked 3 × 3 convolutional layers which learn to produce "residuals" that are added to the block input. Each convolution is preceded by a preactivation which consists of spatial batch normalization and rectified linear units (ReLUs).

### *2.3.1 Three-column configuration*

Using three 2D planar views (axial, sagittal, and coronal; Figure 4) instead of the full 3D volume allowed us to significantly reduce the size of the input while still retaining important features for accurate classification. This corresponds to a three-column configuration of our model. Let $f_c(x_c, \theta_c)$, $f_s(x_s, \theta_s)$, and $f_a(x_a, \theta_a)$ be the columns for the coronal, sagittal, and axial planar views respectively. This gives us Equation 3, which is an instance of Equation 2.

$$y = \sigma(c_c f_c(x_c, \theta_c) + c_s f_s(x_s, \theta_s) + c_a f_a(x_a, \theta_a)) \tag{3}$$

We found that $c_c$, $c_s$, and $c_a$ had similar values after training, meaning that the three columns made similar contributions to the result.

## 2.4 Deep learning techniques

In addition to residual learning, many other deep learning techniques were investigated in order to maximize classification accuracy.

### *2.4.1 Batch normalization*

Batch normalization [9] regularizes and hastens the training of neural networks. This is achieved by gathering minibatch statistics during training to normalize each layer's input distribution. Equation 4 shows how normalized values $\hat{x}^{(k)}$ are calculated from $d$-dimensional input $x = (x^{(1)}...x^{(d)})$ for a fully connected layer. Applying batch normalization to convolutional layers involves a similar set of calculations.

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}} \tag{4}$$

Batch normalization has quickly been adopted by the deep learning community as it makes training deep networks easier and reduces the impact of selecting poor initial weights. As an additional bonus, batch normalization introduces a form of regularization due to the noisy approximation of minibatch statistics. This helps prevent the network from simply "memorizing" the labels of training nodules, which would not generalize to unseen examples. As is shown in Figure 3, batch normalization is placed before the ReLU non-linearities in the ResNet architecture.

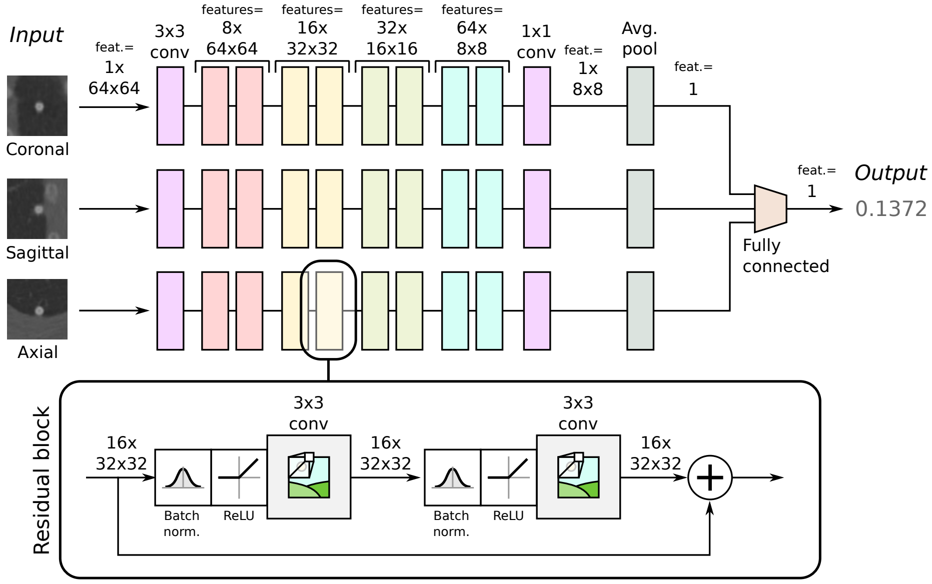

**Figure 3** A configuration of our model using three ResNet-18 columns (drawn horizontally). The magnified inset shows how an individual block is constructed.

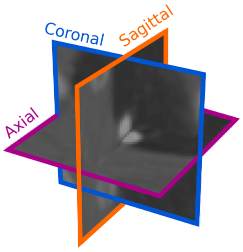

**Figure 4** Coronal, sagittal, and axial planar views of a nodule.

### *2.4.2 Pretraining*

Every neural network must be initially trained from randomly initialized starting weights. Given a small dataset it is often important to pretrain the model on a larger dataset first in order to give the early convolutional layers better initial values. This can significantly improve the generalization ability of the model. We generated pretrained column parameters using the CIFAR-10 dataset, which contains 60,000 images separated into ten classes. After pretraining, the model was trained as per usual on the nodule data, with the notable difference of parameters starting with values learned from another task as opposed to purely random initialization. Our motivation for pretraining on CIFAR-10 is that the amount of nodule data we have is small, and often the features learned close to the input of neural networks are linear edge-detecting filters that tend to develop regardless of the specific task.

### *2.4.3 Curriculum learning*

Curriculum learning [2] is a general term used to describe the technique of somehow increasing the difficulty of training examples as the model learns. There is an intuition here that, like humans, artificial neural networks will learn better by starting with easier problems. For the nodule classification problem we define an *easy example* to be one which is clearly malignant (median rating = 5) or clearly benign (median rating = 1), and an *easy minibatch* to be a minibatch containing only easy examples. When loading inputs during training, we stochastically decide whether the current minibatch should be easy or not using the following probability:

$$\Pr(\textit{easy minibatch}) = \frac{k}{N_{epochs}} \tag{5}$$

Where $k$ = the current epoch and $N_{epochs}$ = the total number of epochs. Therefore the model will be exposed to an increasing number of difficult examples as it learns, which may improve parameter optimization as the complexity of the loss surface is gradually increased with time. We choose a gradual approach instead of a drastic switch, as suddenly changing the surface being optimized would likely have harmful effects on learning.

### *2.4.4 Test data augmentation*

Augmentations involving variations in scale, rotation, and translation were applied to the nodules at test time. Transformations to the nodule examples were applied in three-dimensional space using randomly sampled values. Scale was increased or decreased by up to

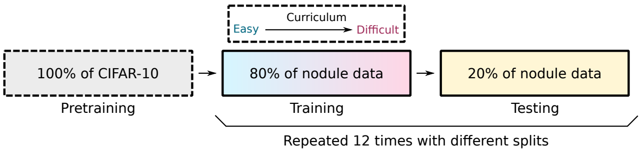

**Figure 5** Flowchart of the training and evaluation process. “Optional” phases are shown with dashed borders.

2%, translation was applied by up to 1 mm in any direction, and rotation was applied by up to 2 degrees about a random axis. By passing multiple variations of the input through the network with small transformations, we can gather multiple malignancy predictions. These predictions are then combined by taking the average, which should improve robustness at test time and yield better accuracy. Various forms of test data augmentation have been successfully utilized in many well-established CNN architectures [11,22,18,24,5].

We also tried applying similar augmentations to the training data, but found that doing so did not have a noticeable positive effect on performance. Hence for clarity in the results we do not include experiments with training augmentation.

### 2.4.5 Training

Generally speaking, the goal of training a neural network for classification is to optimize the parameters ($\boldsymbol{\theta}$) of the model ($f(\boldsymbol{x},\boldsymbol{\theta})$) such that it produces the intended output label ($y$) for each input ($\boldsymbol{x}$). Our training set consists of input/output pairs, $\langle \boldsymbol{x}_i, y_i \rangle$, where $\boldsymbol{x}_i$ is a CT image of a nodule and $y_i \in \{0,1\}$ is the associated malignancy label. It follows that our optimization goal is to find $\boldsymbol{\theta}^*$, the parameters which maximize the probability of the network predicting the correct label for each training example.

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}}(\textstyle\sum_i \Pr(y = y_i|\boldsymbol{x}_i,\boldsymbol{\theta})y_i) \\ &= \arg\min_{\boldsymbol{\theta}}(\textstyle\sum_i[-\log(f(\boldsymbol{x}_i,\boldsymbol{\theta}))y_i + \\ &\qquad\qquad -\log(1-f(\boldsymbol{x}_i,\boldsymbol{\theta}))(1-y_i)])\end{aligned} \tag{6}$$

We use the Adadelta [28] optimizer to find $\boldsymbol{\theta}^*$, since it adaptively sets learning rates.

Figure 5 depicts a high-level view of the training process in flowchart form. Initially the model begins with random parameters. If pretraining is to be undertaken, the column model is then trained on all of the CIFAR-10 dataset. The model is then trained on 80% of the nodule data, which has been designated as the training set. If curriculum learning is enabled, easier nodule examples are selected towards the beginning, gradually increasing in difficulty as training progresses. Once training is completed, the system is evaluated on the remaining 20% of the nodule data which comprise the test set.

During training we used minibatches of 128 examples per iteration. We stopped training each model after 200 epochs, since the parameters had converged by this point. An epoch is defined as being one complete pass through all training examples. Early stopping with a validation set was not employed since such an approach would require the small dataset to be split further to create a validation set. This would either result in a smaller training set or an even smaller test set.

*Multi-column network training* When the network has multiple columns, the training process is extended slightly. We begin by training a single column as described above on axial images only. We then clone this column to construct the final multi-column network and train for another 50 epochs with curriculum learning disabled. This additional training fine-tunes the other columns to better recognize sagittal and coronal planes.

### 2.4.6 Implementation of comparison networks

We compare our system against a simple fully connected network as a baseline and two strong rivals consisting of a state-of-the-art shallow CNN [19] and a pretrained OverFeat network [18].

For each of the reference implementations we again use Adadelta [28] as the optimization algorithm. One exception is the fully connected network which failed to learn using Adadelta - this particular network was trained using RMSProp [26] with a learning rate of $1 \times 10^{-5}$.

Below we present the comparison models in more detail.

*Multilayer perceptron (MLP)* A multilayer perceptron (MLP) is a feed-forward neural network consisting of stacked fully connected layers. By flattening the input image into a one-dimensional vector, it is possible to use a MLP to perform classification of nodules. This is a simple baseline model. The specific network we implemented consists of three fully connected (FC) layers configured as shown in Figure 6.

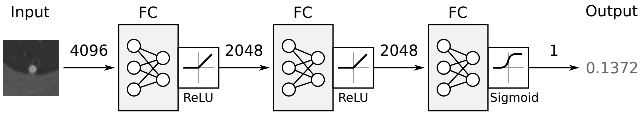

**Figure 6** The baseline MLP model.

*Shallow CNN (setio-cnn)* To represent the shallow CNNs used in many state-of-the-art nodule classification systems we implemented the "false positive reduction" CNN described by Setio et al. [19]. This is the deepest custom CNN we could find amongst the related works. The model uses 50% dropout for regularization and has three convolutional layers (Figure 7).

*OverFeat* OverFeat [18] is a well-known CNN based on AlexNet [11] that contains 6 convolutional layers. Ciompi et al. [3] make use of a publicly available version of OverFeat which has been pretrained on ImageNet data, adapting it to perform nodule classification. We followed a similar process to compute 4096-element feature vectors from nodule CT slices, then used an MLP to produce the final output prediction.

### *2.4.7 Performance metrics*

In the context of nodule classification we consider malignant nodules to be positive examples. *Accuracy* is the percentage of examples which were correctly classified. *Sensitivity* is true positive rate. *Specificity* is true negative rate. *Precision* is the ratio of true positives to predicted positives.

*Area under receiver operating characteristic curve.* To convert the malignancy probability output by the classifier to a binary response, we must set some threshold (eg. $\tau = 0.5$). However, decreasing or increasing $\tau$ will cause the classifier to produce more positive or negative predictions respectively. The classifier can be characterized with a receiver operating characteristic (ROC) curve as $\tau$ is varied from 0 to 1. The area under the ROC curve (AUROC) expresses the probability that the classifier will rank a random positive example higher than a random negative example [4].

### *2.4.8 Evaluation*

The accuracy of models varies slightly according to training/test splits and fluctuations between training iterations. Note that for our evaluation the terms "test set" and "validation set" are synonymous since parameters are selected solely on the training set. To compare different models as fairly as possible the impact of these variations was reduced by training and testing each model 24 times with ~~different~~randomized dataset splits. Each trial involved training for 200 epochs with minibatches of size 128, setting aside one fifth of the data for testing the ability of the model to generalize to previously unseen examples. Although slightly different to traditional k-fold cross-validation, this form of evaluation has the distinct advantage of enabling more trials without reducing the size of the test set (which would increase variance). The randomized splits serve the same fundamental purpose as the fixed splits in k-fold cross-validation, allowing us to set aside a portion of the data for testing the generalization of the model.

### *2.4.9 Hardware*

Each model was trained and evaluated using a Maxwell architecture NVIDIA Titan X GPU. The training time per model was typically in the order of a few hours to convergence, though we did not benchmark on this explicitly since certain aspects of the system (such as the data loader) were built for experimental flexibility rather than speed.

## 3 Results

### 3.1 Depth

In order to determine the optimal ResNet depth for nodule classification, we conducted a series of experiments on single column networks with pretraining and curriculum learning disabled. Depth was adjusted by changing the number of blocks within each of the four groupings shown in Figure 3. For example, a depth of 18 has 2 blocks per grouping (since $18 = 1+4\times2+1$). The results shown in Figure 8 indicate that increased depth does not necessarily correspond to an increase in accuracy (error bars mark the 80% confidence interval obtained via statistical bootstrapping). More specifically, there is a "sweet spot" at 18 weighted layers where accuracy is at a maximum. This seems to suggest that, for our particular task, increasing depth can only improve accuracy up to a certain point. We observe that since the issue is not with vanishing gradients, the dataset itself must not be especially conducive to extremely deep hierarchical representation. Since 18 weighted layers was found to be optimal, we used ResNet-18 for the next set of experiments.

### 3.2 Comparison of configurations

Different variants of our ResNet-based model were trained and evaluated. The number of training epochs was fixed at 200, and the results were averaged across 24 trials. Each of the 24 trials was a run of the exact same experiment, but with different training/test splits. The measured accuracy (with and without test set augmentation) for each tested configuration is shown in Table 1.

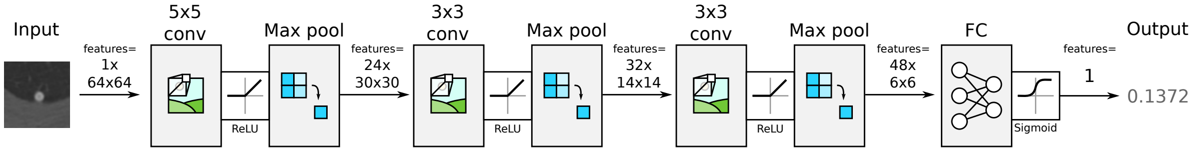

**Figure 7** The “setio-cnn” shallow convolutional neural network model.

**Table 1** Accuracies for deep residual model variants

| Columns |  | Pretrain | Curriculum | Test accuracy % |  |
|---|---|---|---|---|---|
| Type | Count |  |  | Plain | Augmented |
| ResNet-18 | 1 | ✗ | ✗ | 85.75 | 85.76 |
| ResNet-18 | 1 | ✓ | ✗ | 86.15 | 86.16 |
| ResNet-18 | 1 | ✗ | ✓ | 88.07 | 88.22 |
| ResNet-18 | 1 | ✓ | ✓ | 89.09 | 89.09 |
| ResNet-18 | 3 | ✗ | ✗ | 87.37 | 87.37 |
| ResNet-18 | 3 | ✓ | ✗ | 87.67 | 87.88 |
| ResNet-18 | 3 | ✗ | ✓ | 88.69 | 88.89 |
| ResNet-18 | 3 | ✓ | ✓ | **89.64** | **89.90** |

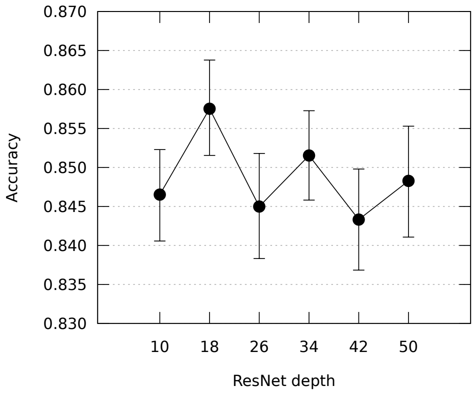

**Figure 8** Classification accuracy for different ResNet depths.

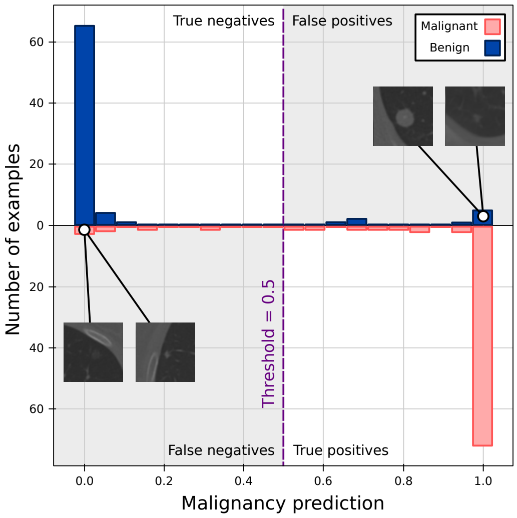

**Figure 9** Histogram of test set predictions from our best model. Examples are grouped by label (malignant/benign) and binned by malignancy probability. Instances of strongly misclassified examples are shown.

The results show that using modern deep learning techniques improves accuracy at test time, with the best model involving pretraining and curriculum learning. The 3-column models always outperformed their 1-column counterparts.

Figure 9 depicts a histogram of output predictions from our best model. Most examples are correctly classified with high confidence. By examining the misclassified examples at either extreme, we can begin to understand why the output sometimes differed from ground truth. The false negatives contained small, faint nodules that were assessed as malignant by the radiologists. The false positives either contained large, pronounced nodules or faint images that strongly resemble the false negatives. Since large benign nodules are uncommon in the dataset, it is quite possible that with a larger dataset the system could learn to better classify examples like these.

### 3.3 Comparison with existing systems

In order to benchmark the performance of ResNet-based systems relative to existing work, we selected our best variant (single column ResNet-18 with pretraining and curriculum learning). This was then subjected to an in-depth comparison with MLP, shallow CNN (setio-cnn), and OverFeat models. We also tested an enhanced version of our strongest rival, setio-cnn+, which incorporated pretraining, curriculum learning, and a 3-column setup. As is shown in Table 2, our ResNet-based system outperforms everything else for each of the performance metrics gathered. The Wilcoxon signed-rank test was applied to the measured accuracies for each pair of models, revealing that the results are statisti-

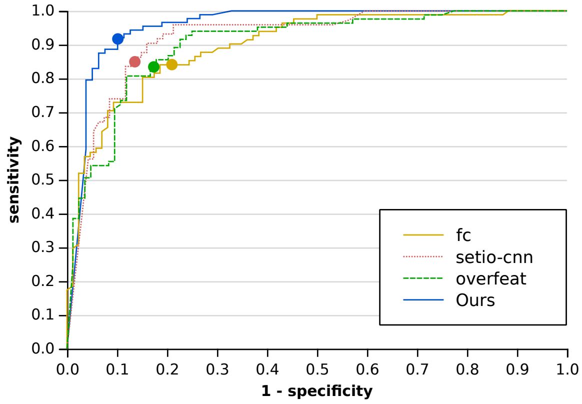

**Figure 10** ROC curves for each model. Points are marked where the accuracy threshold is 0.5.

**Table 2** Comparison of nodule classification systems

| Model | Sensitivity | Specificity | Precision | AUROC | Accuracy |
|---|---|---|---|---|---|
| MLP | 78.87% | 82.68% | 83.93% | 0.9041 | 80.59% |
| setio-cnn | 87.82% | 84.48% | 85.67% | 0.8950 | 86.23% |
| setio-cnn+ | 89.21% | 84.96% | 86.31% | 0.9342 | 87.18% |
| OverFeat | 82.92% | 81.01% | 82.21% | 0.9003 | 81.94% |
| Ours | **91.07%** | **88.64%** | **89.35%** | **0.9459** | **89.90%** |

cally significant with $p$-values under 0.0001 (excluding the setio-cnn/setio-cnn+ pair, which yielded a $p$-value of 0.02). The deep learning techniques which benefit our network were ~~also~~ found to benefit setio-cnn, but not to the same extent. Interestingly, our system also exhibits a very nice balance between sensitivity and specificity, which implies that it is equally good at recognizing positive and negative examples. The ROC curves in Figure 10 provide evidence that our system demonstrates the best behavior in terms of both sensitivity and specificity across a wide range of threshold values.

## 4 Discussion

Using a deep residual network with pretraining, curriculum learning, and a 3-column architecture appears to be a very strong recipe for success in the lung nodule classification task. An interesting observation is that it really seems to be the complete combination of these factors which yields superior results. For example, our results showed that a shallow convolutional network is better than ResNet-18 without any enhancements applied to either. However, after incorporating deep learning techniques we found that the enhanced ResNet-18 architecture was better than the enhanced shallow network (setio-cnn+) by a significant accuracy margin of over 2.5 percentage points. If we had simply assumed that ResNet-18 would always be inferior based on our initial results, we would not have been able to achieve classification accuracies that were as high as we observed.

We theorize that the key reason why ResNet received more of a benefit from curriculum and transfer learning than the shallow CNN is that its higher representational power is not being appropriately guided otherwise. That is to say, ResNet is more susceptible to overfitting than the shallow CNN when trained on a small dataset. Once we have overcome the overfitting discrepancy between the two architectures, the upside of having increased representational power shines through and we arrive at the favorable results presented in Table 2.

We tried applying 3D convolutions to the problem as well, but found they were unwieldy due to the large memory requirements that they imposed. Furthermore, their use was complicated by inconsistency among CT slice depth spacings within the dataset.

## 5 Conclusions

We have investigated multiple variants of a pulmonary nodule malignancy classification system based on deep residual networks. In addition, this is the first paper to objectively compare results against two state-of-the-art rival convolutional neural network models (setio-cnn and OverFeat) under similar training and testing conditions. Our best system configuration was shown to outperform all alternatives through experiments which were designed to report real-world classification performance on previously unseen examples. Our results suggest that modern advancements in deep learning are also applicable to medical imaging, and could be used to increase the feasibility of lung cancer screening programs involving early detection with CT scans.

Interesting future work includes comparing the systems presented here with solutions based on hand-engineered features, and leveraging the deep residual architecture to perform automatic nodule detection. Furthermore, the models in this paper were trained to reproduce the subjective rating of an ensemble of radiologists, and it remains an open question as to how well this correlates with performance in real clinical practice.

**Acknowledgements** The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

**Ethical standards** This article does not contain any studies with human participants or animals performed by any of the authors. Informed consent was obtained from all individual participants included in the study. Aiden Nibali, Zhen He, and Dennis Wollersheim declare that they have no conflict of interest.

## References

1. American Cancer Society: Cancer facts & figures 2016 (2016). URL http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf. Last accessed 2016-08-23
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proc. 26th ICML, pp. 41–48 (2009)
3. Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., van Ginneken, B.: Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Medical Image Analysis **26**(1), 195–202 (2015)
4. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 29th IEEE CVPR (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Identity Mappings in Deep Residual Networks. ArXiv Preprint ArXiv:1603.05027 (2016)
7. Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H., Chen, Y.J.: Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets and therapy **8**, 2015–2022 (2015)
8. International Agency for Research on Cancer: Estimated incidence, mortality and prevalence worldwide in 2012 (2012). URL http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx?cancer=lung. Last accessed 2016-08-23
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. 32nd ICML, pp. 448–456 (2015)
10. Kim, Y.: Convolutional neural networks for sentence classification. In: Proc. EMNLP 2014 (2014)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Adv. Neural Inf. Process. Syst. 25, pp. 1097–1105 (2012)
12. Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classification using deep features in CT images. In: 12th Conf. Comput. Robot Vis., pp. 133–138 (2015)
13. Kuruvilla, J., Gunavathi, K.: Lung cancer classification using neural networks for CT images. Comput. Methods Programs Biomed. **113**(1), 202–209 (2014)
14. LeCun, Y., Boser, B., Denker, J.S., Howard, R.E., Habbard, W., Jackel, L.D., Henderson, D.: Handwritten digit recognition with a back-propagation network. In: Adv. Neural Inf. Process. Syst. 2, pp. 396–404 (1990)
15. Lee, M.C., Boroczky, L., Sungur-Stasik, K., Cann, A.D., Borczuk, A.C., Kawut, S.M., Powell, C.A.: Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. Artif. Intell. Med. **50**(1), 43–53 (2010)
16. Madero Orozco, H., Vergara Villegas, O.O., Cruz SÃ¡nchez, V.G., Ochoa DomÃnguez, H.d.J., Nandayapa Alfaro, M.d.J.: Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomed. Eng. Online **14**, 9 (2015)
17. Reeves, A.P., Biancardi, A.M.: The Lung Image Database Consortium (LIDC) Nodule Size Report (2011). URL http://www.via.cornell.edu/lidc/. Last accessed 2016-08-23
18. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: Integrated recognition, localization and detection using convolutional networks. In: Proc. ICLR 2014 (2014)
19. Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Riel, S.v., Wille, M.W., Naqibullah, M., Sanchez, C., Ginneken, B.v.: Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imag. **35**(5), 1160–1169 (2016)
20. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: Inf. Process. Med. Imaging, 9123, pp. 588–599 (2015)
21. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. ICLR 2015 (2015)
23. Smith, K., Clark, K., Bennett, W., Nolan, T., Kirby, J., Wolfsberger, M., Moulton, J., Vendt, B., Freymann, J.: Data from LIDC-IDRI (2015). URL https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI. Last accessed 2016-08-23
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. ArXiv Preprint ArXiv:1409.4842 (2014)
25. Telgarsky, M.: Benefits of depth in neural networks. ArXiv Preprint ArXiv:1602.04485 (2016)
26. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude (2012)
27. Toshev, A., Szegedy, C.: DeepPose: Human pose estimation via deep neural networks. In: Proc. CVPR 2014 (2014)
28. Zeiler, M.D.: ADADELTA: An adaptive learning rate method. ArXiv Preprint ArXiv:1212.5701 (2012)
29. Zinovev, D., Feigenbaum, J., Raicu, D., Furst, J.: Predicting Panel Ratings for Semantic Characteristics of Lung Nodules. Tech. Rep. (2010). URL http://via.library.depaul.edu/tr/18