Size and Texture-based Classification of Lung Tumors with 3D CNNs

ZhiHao Luo
Columbia University
zhihao.luo@columbia.edu

Marcus A. Brubaker
York University
mab@eecs.yorku.ca

Michael Brudno
University of Toronto
brudno@cs.toronto.edu

Abstract

In this paper, we explore the use of current deep learning methods in the field of computer-aided diagnosis (CAD). Specifically we propose the use of 3D convolutional neural nets (CNN) in classifying lung nodules based off of their appearance in CT scans. We explore the choices of network architectures, learning parameters and problem formulations. Comparing these results to other methods we show that the proposed method has close to perfect performance on the publicly available LIDC dataset, achieving an AUC of 0.9685 *and a false positive rate of* 0.46% *with a true positive rate of* 90% *where the ground truth is the expert opinion of a radiologist.*

1. Introduction

Lung cancer is a deadly disease. In 2015 alone it is estimated that it will be responsible for over 150,000 deaths in the United States, accounting for 27% of all cancer deaths: more than any other single cancer. [18]. Early identification of cancerous lung tumors is a key factor in improving the survival rate of lung cancer patients [22].

Thoracic CT scans are a common, non-intrusive imaging modality which is often used to identify cancerous lung tumors. However, identification from CT scans is not straightforward. At early stages, lung cancer may present only as a small nodule which can be hard to differentiate from healthy lung tissue or otherwise benign nodules. Without biopsy, it may take several screenings, spanning months or even years, to arrive at a diagnosis. This long timeline increases costs, adds strain to the healthcare system, and can be detrimental to the patient due to added stress and radiation exposure. While biopsies can provide definitive answers, they have both higher costs and a non-trivial risk of complications, making diagnosis from imaging preferable.

Currently CT scans are assessed by a radiologist who will identify potential lung nodules and attempt to determine whether they are benign or malignant. While computer-aided diagnosis (CAD) systems have been considered in the past [4], for lung nodule diagnosis they have rarely been shown to be useful or effective and in some cases have even been detrimental [6, 15]. To help address this problem, the LIDC database was created to serve as a benchmark dataset for CAD systems and to encourage the development of new methods [1].

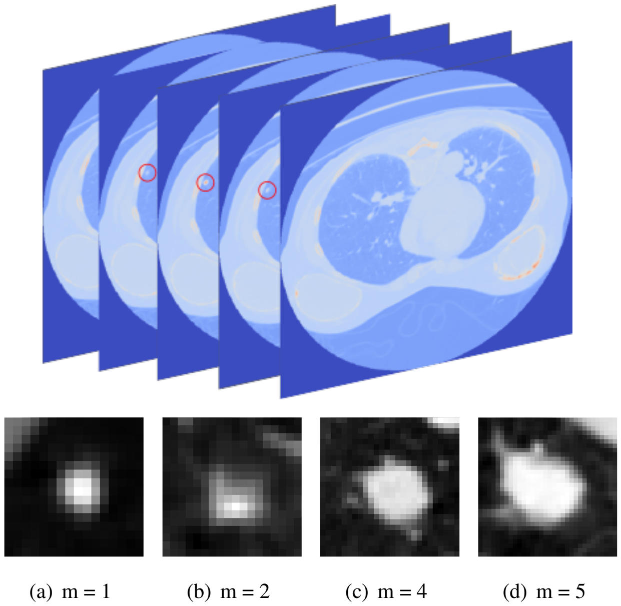

(a) m = 1
(b) m = 2
(c) m = 4
(d) m = 5

Figure 1. A sample of the data in the LIDC dataset. Shown Above is a set of 2D slices from a 3D CT scan with an identified lung nodule circled in red. Shown below are slices through sample nodules from four different malignancy levels in order of increasing likelihood of malginancy; the two on the left are likely benign while the two on the right are likely malignant.

In this paper we consider the problem of predicting the malignancy of identified lung nodules in CT scans. One of the most obvious identifying features of malignancy is size, as larger tumors are rarely benign. However, the identification of small tumors is critical to early diagnosis and successful treatments. For these cases, radiologists rely on aspects of nodule texture. As such, we propose an approach which explicitly combines size information with textural information to achieve accurate classification of nodules.

We propose the use of convolutional neural networks to recognize nodules based on texture. However, CNNs have primarily been applied to 2D images while CT scans are most naturally represented as a 3D volume. Thus, we extend the convolutional aspect of these networks to 3D and explore a variety of network architectures, objective functions and training parameters. Together with size information, we show that CNNs are extremely effective for this problem, achieving close to perfect results on the LIDC dataset.

1.1. Background and Related Work

The problem of identifying benign and malignant nodules from CT scans can be separated into two distinct subproblems: detection of all nodules, which is then followed by classification of each nodule as benign or malignant. Approaches for nodule detection (e.g., [23]) typically begin by separating the lung from other parts of the anatomy and extracting nodules through the use of standard segmentation algorithms like MRFs. More advanced techniques, e.g., using neural networks [20] or vector quantization [9], have also been considered.

The problem of nodule classification has not been as widely studied due to the lack of suitably large datasets. Most methods consist of manually crafting a low-dimensional feature vector describing aspects of the nodules. In [7] 3D gradients and ellipsoidal shape descriptors were used to construct a feature vector which was used with a linear classifier, trained using LDA. Similarly, [23] used a collection of hand-crafted features but with an SVM for classification after a set of manually specified rule-based classifiers. Beyond hand specified features, [14] used an autoencoder and unsupervised learning to extract a 200 dimensional feature vector which was then classified using a decision tree.

In contrast, we propose to directly learn the 3D features needed for classification through the use of a convolutional neural network (CNN). CNNs have recently been successful in 2D image recognition tasks [13] as well as in other problems. 3D CNNs have been considered before for action recognition [10], video analysis and classification [21, 12], and even for detection of cerebral micro-bleeds and brain lesions [5, 11]. Here we consider their application to the analysis of nodules in CT scans.

The Lung Image Database Consortium (LIDC) dataset is a publicly available database that contains 2,434 identified lung nodules with useful labels in 1,010 thoracic CT scans [1]. It is the result of the joint efforts of seven academic centers and eight medical imaging companies to create a larger scale dataset which could then be used to test and develop new CAD methods. Each CT scan is accompanied by detailed annotations from four radiologists which identify the locations of lung nodules and rate them on a range of aspects including subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy.

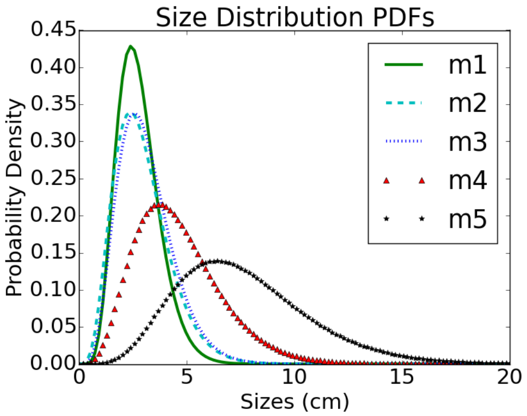


Figure 2. Nodule size distributions for different levels of malignancy. Note the distributions of the first three malignancy levels are very similar.

In the LIDC dataset, the malignancy of each nodule is rated with a value of 1 to 5 with 1 indicating definitely benign, 5 indicating definitely malignant and values in between indicating a degree of uncertainty by the radiologist. An example of a such CT scan and four nodules of different levels of malignancy can be seen in Figure 1. In this work we focus on predicting the malignancy of a nodule as both a multi-class classification problem and a binary classification problem as both are of interest. In the binary case, we combine levels 1 through 3 into a single “benign” class and levels 4 through 5 into a single “malignant” class.

2. Methods

Below we describe our proposed method for lung nodule classification. Our approach consists of probabilistic, size-based classifier and a convolutional neural network which operates on a size-normalized representation of the nodule. These two pieces are combined into a single probabilistic model.

2.1. Nodule Size

For lung nodules a strongly indicative characteristic of malignancy is size. To see this, we plot the distribution of nodule size for each malignancy level in the LIDC dataset. These distributions, plotted in Figure 2, clearly show that malignant nodules are more likely to be larger. It is also clear that there is significant overlap of the distributions leading to significant uncertainty based on size alone. Thus, while size can be an important cue, it is also very limited and a classifier which uses size information must appropriately handle this uncertainty. As such, we adopt a probabilistic formulation.

Formally, let s denote the size of a lung nodule, mea-

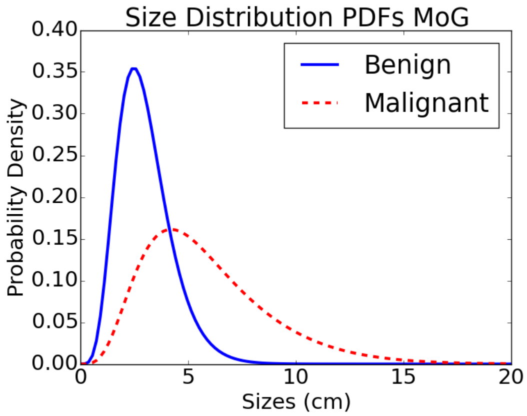


Figure 3. Nodule size distributions benign and malignant classes.

sured as the radius of a sphere which bounds the nodule and M denote the malignancy of a nodule. We use a Gamma distribution to represent the distribution of sizes for nodules with malignancy level i, i.e., $p(s|M = i) = \mathcal{G}(s|\alpha_i, \beta_i)$ where $\mathcal{G}(s|\alpha, \beta)$ denotes the probability density function of a Gamma distribution with shape parameter α and rate parameter β. The maximum likelihood estimate is used to find the parameters α_i and β_i for each malignancy level $i = 1, \dots, 5$.

A size-based classifier can then be constructed by considering the posterior probability of the $M = i$ given the observed nodule size s. Using Bayes rule, we have that

$$p(M = i|s) = \frac{p(s|M = i)p(M = i)}{\sum_1^5 p(s|M = j)p(M = j)}$$

where $p(M = i)$ is the prior probability of malignancy class i which can be easily estimated from data. This equation can be evaluated for $i = 1, \dots, 5$ and the malignancy level which achieves the highest probability is selected.

As might be expected, this classifier struggles to accurately differentiate malignancy levels. A confusion matrix for such a classifier is shown in Figure 6, where it can be seen that even though many larger nodules are correctly identified as level 5, most nodules of levels 1 through 4 are classified as level 3 due to a higher prior probability, $p(M = 3)$, clearly indicating the need for a more detailed, texture based analysis.

We also consider the problem of binary classification, where, instead of predicting one of five levels, a binary determination of benign or malignant is made. To do this, we combine together levels 1, 2 and 3 into a "benign" class and levels 4 and 5 into a "malignant" class. The size distributions of the benign and malignant classes are then, respectively,

$$p(s|M = \mathrm{b}) = Z_b^{-1} \sum_{i=1}^{3} p(s|M = i)p(M = i) \quad (1)$$

$$p(s|M = \mathrm{m}) = Z_m^{-1} \sum_{i=4,5} p(s|M = i)p(M = i) \quad (2)$$

where $Z_b = \sum_{i=1}^3 p(M = i)$ and $Z_m = \sum_{i=4}^5 p(M = i)$. These combined distributions can be seen in Figure 3.

2.2. Texture

Size information alone is insufficient to predict the malignancy of lung nodules. Texture, shape, density and other aspects of nodule appearance are also critical and must be incorporated. While previous methods have attempted to manually calculate some kind of feature descriptor which includes these characteristics [14], here we propose to use a convolutional neural network (CNN).

A neural network consists of a set of layers of simple computational units referred to as neurons. The first layer consists simply of the input with the output of each layer feeding into the next, producing a cascading form of computation until the final layer which is used as the output. Different network architectures are defined then by the number of layers, the number and type of neurons in each layer and the connectivity between layers. CNNs are differentiated from traditional neural networks through the use of convolutional layers which restrict the form of connectivity to be "local" and share neuron weight parameter between units in a layer. This significantly reduces the number of parameters while providing for certain types of input invariance and still allowing the network to learn very complex functions. CNNs have recently been highly successful in computer vision [13].

There are several hurdles which must be considered when applying CNNs to lung nodules. Nodules are irregularly shaped and sized while neural networks require a fixed sized input. To address this we extract a bounding box around the nodule and then interpolate this at a fixed set of points, resulting in a nodule volume of 20 × 20 × 10 where the z-dimension is coarser because CT scans typically have lower resolution along the vertical axis. Note that this also results in the network input being relatively invariant to the overall size of the nodule, ensuring that the network is only considering characteristics of appearance. In addition, CNNs have traditionally been used on 2D image data and hence the network architectures are not directly applicable. Thus we will explore a range of network architectures and learning parameters. Finally, while the LIDC dataset is large by some standards, it is too small for the direct application of CNNs. Thus we augment our dataset by including randomly rotated versions of each nodule, since we expect the malignancy of the nodule to be unaffected by

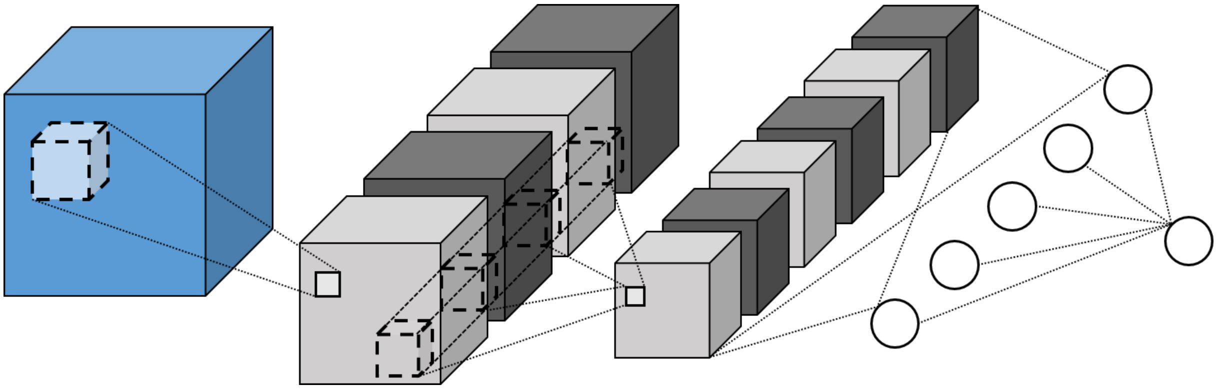


Figure 4. An example architecture of a 3D Convolutional Neural Network used here. On the left is the input 3D volume, followed by two convolutional layers, a fully connected layers and an output layer. In the convolutional layers, each filter (or channel) is represented by a volume.

its orientation in the CT scan. In the remainder of this section, we describe the technical details of the neural network architecture we used and how it was trained.

2.2.1 Convolutional Neural Networks

A convolutional neural network consists of some number of convolutional layers, followed by one or more fully connected layers and finally an output layer. An example of this architecture is illustrated in Figure 4. Formally, we denote the input to layer m of the network by $I^{(m)}$. The input to a 3D convolutional layer m of a neural network is a $n_1^{(m-1)} \times n_2^{(m-1)} \times n_3^{(m-1)}$ 3D object with $n_c^{(m-1)}$ channels, so $I^{(m-1)} \in \mathbb{R}^{n_1^{(m-1)} \times n_2^{(m-1)} \times n_3^{(m-1)} \times n_c^{(m-1)}}$ and its elements are denoted by $I^{(m,\ell)}_{i,j,k}$ where i, j, and k index the 3D volume and ℓ selects the channel. The output of a convolutional layer m is defined by its dimensions, *i.e.*, $n_1^{(m)} \times n_2^{(m)} \times n_3^{(m)}$ as well as the number of filters or channels it produces $n_c^{(m)}$. The output of layer m is a convolution of its input with a filter and is computed as

$$I^{(m,\ell)}_{i,j,k} = f_{\tanh}(b^{(m,\ell)} + \sum_{i',j',k',\ell'} I^{(m-1,\ell')}_{i',j',k'} W^{(m,\ell)}_{i-i',j-j',k-k',\ell'}) \tag{3}$$

where $W^{(m,\ell)}$ and $b^{(m,\ell)}$ are the parameters which define the ℓth filter in layer m The locations where the filters are evaluated (*i.e.*, the values of i, j, k for which $I^{(m,\ell)}_{i,j,k}$ is computed) and the size of the filters (*i.e.*, the values of $W^{(m,\ell)}$ which are non-zero) are parameters of the network architecture. Finally, we use a hyperbolic tangent activation function with $f_{\tanh}(a) = \tanh(a)$.

Convolutional layers preserve the spatial structure of the inputs, and as more layers are used, build up more and more complex representations of the input. The output of the convolutional layers is then used as input to a fully connected network layer. To do this, the spatial and channel structure is ignored and the output of the convolutional layer is treated as a single vector. The output of a fully connected is a 1D vector $I^{(m)}$ whose dimension is a parameter of the network architecture. The output of neuron i in layer m is given by

$$I_i^{(m)} = f_{\mathrm{ReLU}}\left(b^{(m,i)} + \sum_j I_j^{(m-1)} W_j^{(m,i)}\right) \tag{4}$$

where $W^{(m,i)}$ and $b^{(m,i)}$ are the parameters of neuron i in layer m and the sum over j is a sum over all dimensions of the input. The activation function $f_{\mathrm{ReLU}}(\cdot)$ here is chosen to be a Rectified Linear Unit (ReLU) with $f_{\mathrm{ReLU}}(a) = \max(0, a)$. This activation function has been widely used in a number of domains [24, 16] and is believed to be particularly helpful in classification tasks as the sparsity it induces in the outputs helps create separation between classes during learning [17, 3].

The last fully connected layer is used as input to the output layer. The structure and form of the output layer depends on the particular task. Here we consider two different types of output functions. In classification problems with K classes, a common output function is the softmax function

$$f_i = \frac{\exp(I_i^{(o)})}{\sum_j \exp(I_j^{(o)})} \tag{5}$$

$$I_i^{(o)} = b^{(o,i)} + \sum_{k=1}^{K} W_k^{(o,i)} I_k^{(N)} \tag{6}$$

where N is the index of the last fully connected layer, $b^{(o,i)}$ and $W^{(o,i)}$ are the parameters of the ith output unit and $f_i \in [0, 1]$ is the output for class i which can be interpreted as the probability of that class given the inputs. We also consider a variation on the logistic output function

$$f = a + (b-a)\left(1 + \exp(b^{(o)} + \sum_j W_j^{(o)} I_j^{(N)})\right)^{-1} \tag{7}$$

which provides a continuous output f which is restricted to lie in the range (a, b) with parameters $b^{(o)}$ and $W^{(o)}$. We call this the *scaled logistic* output function. We note that when considering a ranking-type multi-class classification problem like predicting the malignancy level this output function might be expected to perform better.

2.2.2 Training

Given a collection of data and a network architecture, our goal is to fit the parameters of the network to that data. To do this we will define an objective function and use gradient based optimization to search for the network parameters which minimize the objective function. Let $\mathcal{D} = \{\mathbf{n}_i, y_i\}_{i=1}^{D}$ be the set of D (potentially augmented) training examples where $\mathbf{n}$ is an input (a portion of a CT scan) and y is the output (the malignancy level or a binary class indicating benign or malignant) and Θ denote the collection of all weights W and biases b for all layers of the network. The objective function has the form

$$E(\Theta) = \sum_{i=1}^{D} L(y_i, f(\mathbf{n}_i, \Theta)) + \lambda E_{prior}(\Theta) \tag{8}$$

where $f(\mathbf{n}, \Theta)$ is the output of the network evaluated on input $\mathbf{n}$ with parameters Θ, $L(y, \hat{y})$ is a loss function which penalizes differences between the desired output of the network y and the prediction of the network $\hat{y}$. The function $E_{prior}(\Theta) = \|W\|^2$ is a weight decay prior which helps prevent over-fitting by penalizing the norm of the weights and λ controls the strength of the prior.

We consider two different objective functions in this paper depending on the choice of output function. For the softmax output function we use the standard cross-entropy loss function $L(y, \hat{y}) = -\sum_{k=1}^{K} y_k \log(\hat{y}_k)$ where y is assumed to be a binary indicator vector and $\hat{y}$ is assumed to be a vector of probabilities for each of the K classes. A limitation of a cross-entropy loss is that all class errors are considered equal, hence mislabeling a malignancy level 1 as a level 2 is considered just as bad as mislabeling it a 5. This is clearly problematic, hence for the scaled logistic function we use the squared error loss function to capture this. Formally, $L(y, \hat{y}) = (y - \hat{y})^2$ where we assume y and $\hat{y}$ to be real valued.

Given the objective function $E(\Theta)$, the parameters Θ are learned using stochastic gradient descent (SGD) [2]. SGD operates by randomly selecting a subset of training examples and updating the values of the parameters using the gradient of the objective function evaluated on the selected examples. To accelerate progress and reduce noise due to the random sampling of training examples we use a variant of SGD with momentum [19]. Specifically, at iteration t, the parameters are updated as

$$\Theta_{t+1} = \Theta_t + \Delta\Theta_{t+1} \tag{9}$$

$$\Delta\Theta_{t+1} = \rho\Delta\Theta_t - \epsilon\nabla E_t(\Theta_t) \tag{10}$$

where $\rho = 0.9$ is the momentum parameter, $\Delta\Theta_{t+1}$ is the momentum vector, ϵ_t is the learning rate and $\nabla E_t(\Theta)$ is the gradient of the objective function evaluated using only the training examples selected at iteration t. At iteration 0, all biases are set to 0 and the values of the filters and weights are initialized by uniformly sampling from the interval $[-\sqrt{\frac{6}{fan\_in+fan\_out}}, \sqrt{\frac{6}{fan\_in+fan\_out}}]$ as suggested by [8] where fan_in and fan_out respectively denote the number of nodes in the previous hidden layer and in the current layer. Given this initialization and setting $\epsilon_0 = 0.01$, SGD is run for 2000 epochs, during which ϵ_t is decreased by 10% every 25 epochs to ensure convergence.

2.3. Combining Size and Texture

So far we have constructed separate classification methods based on nodule size s and size-normalized appearance $\mathbf{n}$. To combine these we consider the posterior distribution over a class M. Applying Bayes rule and assuming that size and appearance information are independent given the class, we have

$$p(M|s, \mathbf{n}) = \frac{p(M|\mathbf{n})p(s|M)}{\sum_j p(M = j|\mathbf{n})p(s|M = j)} \tag{11}$$

where $p(M|\mathbf{n})$ is the output of the softmax convolutional neural network and $p(s|M)$ is the size-based classifier.

3. Experiments & Results

To evaluate the proposed approach and explore some of the decisions made, we used the previously described LIDC dataset [1]. For simplicity in training and testing we selected the ratings of a single radiologist. All experiments were done using 10-fold cross validation. To evaluate the results we considered a variety of testing metrics. For the binary classification problem, where a nodule is classified as either benign or malignant a natural statistic might be classifier accuracy, however since a significant majority of nodules are benign, this statistic can be misleading. Instead, we consider ROC curves and the area under the curve (AUC) metric as well as the false positive rate (FPR) at different true positive rates (TPR). For the multi-class classification problem, where a nodule is classified as being in one of five malignancy levels, we utilize per-class accuracy and average accuracy. We can also use the multi-class classifiers to perform binary classifications, and thus also use AUC and FPR at TPR measures.

Method	AUC	FPR (%) at Specific TPR		
		at 85%	at 90%	at 95%
Size	0.8145	83.11	89.23	94.28
C2H2	0.9631	0.30	0.56	35.31
C2H1	**0.9685**	**0.20**	**0.46**	**12.85**
C1H1	0.9603	1.32	2.38	23.42

Table 1. Effects of CNN architectures on binary classification performance. SC is the size-based classifier for reference. CnHm indicates a network with n convolutional layers followed by m fully connected layers

Method	AUC	FPR (%) at Specific TPR		
		at 85%	at 90%	at 95%
C2H1 b1	0.9684	0.20	0.76	14.26
C2H1 b10	**0.9685**	**0.20**	**0.46**	**12.85**
C2H1 b50	0.8754	29.99	39.10	60.85
C2H1 b100	0.8131	40.52	53.52	69.20

Table 2. Training with smaller mini-batches converge to better optima as the batch-size 1 and batch-size 10 experiments yield better AUC values as well as FPR's at specific TPR's than the batch-size 50 and batch-size 100 experiments

Network Architectures: In our first set of experiments we considered a range of CNN architectures for the binary classification task. Early experimentation suggested that the number of filters and neurons per layer were less significant than the number of layers. Thus, to simplify analysis the first convolutional layer used 20 filters with size $5 \times 5 \times 3$, the second convolutional layer (if present) used 10 filters with $4 \times 4 \times 2$ and all fully connected layers used 50 neurons. These were found to generally perform well and we considered the impact of one or two convolutional layers followed by one or two fully connected layers. The networks were trained as described above and the results of these experiments can be found in Table 1. Our results suggest that two convolutional layers followed by a single hidden layer is the optimal network architecture for this dataset. The addition of a second layer slightly decreases performance, suggesting that the added capacity of the network may be beginning to overfit. Regardless, all architectures significantly outperformed the size-based classifier, suggesting that there is significant information in nodule appearance. We also explored max-pooling in the convolutional layers, however this caused no significant change in performance.

Method	AUC	FPR (%) at Specific TPR		
		at 85%	at 90%	at 95%
C2H1	**0.9685**	**0.20**	**0.46**	**12.85**
C2H1 NA	0.7176	57.66	68.59	79.56

Table 3. Using augmented data yields significantly better results than using no augmentation (NA).

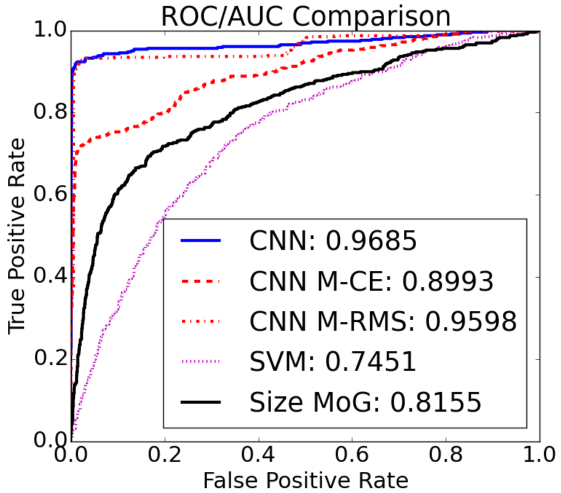


Figure 5. ROC curves of the several different nodules classifiers along with their AUC scores. The CNN and CNN M-RMS achieves best results

Batch Size: Another important parameter in the training of neural networks is the number of observations that are sampled at each iteration, the size of the so-called minibatch. The use of minibatches is often driven in part by computational considerations but can impact the ability of SGD to find a good solution. Indeed, we found that choosing the proper minibatch size was critical for learning to be effective. We tried minibatches of size 1, 10, 50 and 100, and show the results in Table 2. While the nature of SGD suggests that larger batch sizes should produce better gradient estimates and therefor work better, our results here show that the opposite is true. Smaller batch sizes, even as small as 1, produce the best results. We suspect that the added noise of smaller batch sizes allows SGD to better escape poor local optima and thus perform better overall.

Augmentation: As we noted above, the size of the dataset here is relatively small and as such, we opted to use data augmentation. Specifically, each nodule is randomly rotated about the center by an angle that is drawn uniformly from the SO(3). Two random rotations of each nodule plus the original is included in the augmented dataset. This turned out to have a significant impact on the accuracy of the resulting appearance based classifiers. The results, summarized in Table 3, show that the AUC drops from 0.9685 to 0.7176 when the augmented data is removed from the training set. This not only demonstrates the importance of data augmentation, but suggests that perhaps further augmentation (e.g., non-uniform scaling, changes in resolution, etc) may yield even further improvements.

Method	AUC	FPR (%) at Specific TPR		
		at 85%	at 90%	at 95%
C2H1 M-CE	0.8177	24.03	43.10	56.50
C2H1 M-RMS	**0.9598**	**0.61**	**0.96**	**46.53**

Table 4. Quantitative results for 2-class classification using binarized versions of a multi-class classifiers. The CNN trained using the RMS objective function outperforms the one trained with CE objective, however both are outperformed by the original normal binary classifier C2H1 in Table 1.

Baselines: We considered a variety of approaches to nodule classification. Specifically, we compared the binary CNN discussed above with a support vector machine (SVM) using an RBF kernel, a size-based classifier (Size MoG) based off of the distributions in Figure 3 and two binarized multi-class CNNs. The results are shown in Figure 5. For the binarized CNNs we trained two different multi-class classifiers and then used them to distinguish benign (classes 1 through 3) from malignant (classes 4 and 5). In one case, denoted ”CNN M-CE” in Figure 5, we used a softmax output layer with five output classes and the cross-entropy loss function. In the other case, denoted “CNN M-RMS”, we used the scaled logistic output function with the squared error loss function. The results show that the binary CNN outperforms all baselines and, except for “CNN M-RMS”, does so by a large margin.

Combined Size and Texture: Combining together information on both the size and texture of a nodule, we expect that we should be able to improve the quality of predictions. To test this, we added size information to the two best performing CNN models, using the method described above. Perhaps surprisingly, the results of combining size and texture information is mixed. While perhaps disappointing, this can be understood by recognizing that the CNN models are already performing extremely well, with AUCs over 0.95 suggesting that classification performance on this task may already be nearly optimal. Further despite human intuition, the size-based classifier is in fact relatively poor, achieving an AUC of only 0.8145. Thus, our results suggest that contrary to conventional wisdom nodule size is at best only a weak indicator of malignancy.

Multi-class classification: So far we have focused on the binary classification task. We also considered the problem of identifying the specific malignancy level. Based on the experimentation performed above, we used a batch size of 10 and fix the network architecture to be the same as in the binary case, with two convolutional layers followed by a single fully connected layer which is then connected to an output layer.

Method	AUC	FPR (%) at Specific TPR		
		at 85%	at 90%	at 95%
Size Only	0.8145	83.11	89.23	94.28
C2H1	0.9685	**0.20**	**0.46**	12.85
C2H1 + Size	**0.9699**	2.28	2.63	**4.10**
C2H1 M-RMS	**0.9508**	**0.71**	**1.62**	48.41
C2H1 M-RMS + Size	0.9350	11.13	12.70	**16.54**

Table 5. Combined size and texture based classifiers. Our results suggest that texture alone is sufficient to identify malignancy and that adding size information can actually be detrimental to classifier performance.

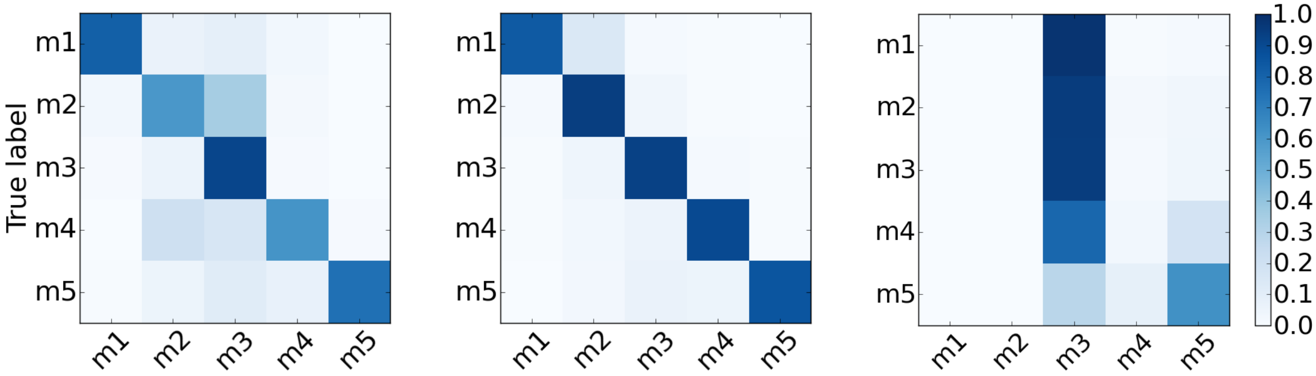
We considered two types of output layers and loss functions for this task. The first used a softmax output function with the cross-entropy loss function and the other using a scaled logistic output function with the squared error loss function. For the scaled logistic output function, class predictions are made by rounding the continuous output of the network.

The results of these two approaches and a size-based classifier, can be seen in Figure 6. Both the Softmax CNN (left) and Scaled Logistic CNN (middle) can be seen to perform quite well, with the Scaled Logistic CNN doing better overall. This is as expected, as the scaled logistic output function captures the fact that the output is a ranking rather than a strict classification problem, as is assumed with the softmax output function. The true weakness of the size-based classifier can be seen here, as nearly all nodules from classes 1 through 4 are classified as class 3 since class 3 has the highest prior probability of the classes and the similar size distribution. Only the largest nodules are able to be correctly classified using size information alone.

To quantitatively assess the performance of these methods, we also compute the per-class accuracy and average accuracy of these three methods. These results are shown in Table 6. As expected, we see that the size classifier does poorly on all classes except for class 3, as it has classified nearly every nodule as belonging to class 3. Overall, the Scaled Logistic CNN outperforms the other two methods by a significant margin.

4. Conclusion and Future Work

This paper presents an approach to classifying lung nodules from CT scans. The approach consists of using a 3D convolutional neural network (CNN) on a size normalized input to classify the nodule. We explore a range of network architectures and learning parameters in order to select the values which perform best. We also compare against several other baselines, including size-based classifiers. We also consider both binary and multi-class formulations of the prediction problem, showing that similar network architectures can be effective for both



(a) Softmax CNN

(b) Scaled Logistic CNN

(c) Size-based Classifier

Figure 6. Confusion matrices for multi-class classification of malignancy levels.

Method	Class Accuracy (%)					Overall (%)
	m1	m2	m3	m4	m5	
Softmax CNN	81.4	59.9	91.8	61.3	75.0	60.3
Scaled Logistic CNN	**83.7**	**94.5**	93.1	**90.0**	**85.6**	**89.3**
Size-based	0.0	0.0	**94.5**	3.0	62.2	31.9

Table 6. Multi-class accuracy for 3 different classifiers. The size-based classifier performs poorly due to the high degree of similarity of size distributions for the first 3 malignancy classes. As a result of this, and the fact that m3 is the most likely malignancy level, most nodules are classified as m3.

Our experimental results utilized the publicly available LIDC [1] dataset. On it we show that the proposed method achieves an extremely high AUC of over 0.96 and outperforms the considered baselines. Further, we show that the performance of the method is such that additional information based on nodule size has at best a mixed impact on classifier performance.

The only other paper that tackles a similar problem is [14]. The authors report a true positive rate of 83.35% at the cost of 0.39 false positives per scan. However, this paper used a slightly different problem definition, training and assessing their results using only a subset of 157 patients which had biopsy results. In comparison we considered the problem of matching an expert radiologists opinion which allowed us to utilize the entire dataset. For reference, our method achieved a 95% true positive rate with 0.096 false positives per scan

The problem of lung nodule classification is still not well solved. The approach proposed here is promising, but more work remains to be done. First, current datasets are somewhat limited due to their size. The LIDC dataset is one of the largest datasets available, but with only a few thousand identified nodules, it is still relatively small and results here suggest that predictive performance on that dataset may already have hit their limit. Beyond that, the use of 3D CNNs for nodule detection is a natural direction to explore. Given their performance for malignancy classification, one should expect them to perform well on the detection task as well. Finally, further study is warranted into the use of nodule size in classification. While our results here suggest there is only marginal value in including size information, there may be other ways to incorporate it and perhaps with larger and more diverse datasets it could play a more significant role.

’

References

[1] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931, 2011.

[2] L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.

[3] G. E. Dahl, T. N. Sainath, and G. E. Hinton. Ieee international conference on acoustics, speech and signal processing, icassp 2013, vancouver, bc, canada, may 26-31, 2013. In *ICASSP*. IEEE, 2013.

[4] K. Doi. Current status and future potential of computer-aided diagnosis in medical imaging. *The British Journal of Radiology*, 2014.

[5] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P.-A. Heng. Automatic detection of

cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195, 2016.

[6] L. H. Eadie, P. Taylor, and A. P. Gibson. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European Journal of Radiology*, 81(1):e70–e76, 2012.

[7] Z. Ge, B. Sahiner, H.-P. Chan, L. M. Hadjiiski, P. N. Cascade, N. Bogot, E. A. Kazerooni, J. Wei, and C. Zhou. Computer-aided detection of lung nodules: false positive reduction using a 3d gradient field method and 3d ellipsoid fitting. *Medical physics*, 32(8):2443–2454, 2005.

[8] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

[9] H. Han, L. Li, H. Wang, H. Zhang, W. Moore, and Z. Liang. A novel computer-aided detection system for pulmonary nodule identification in ct images. In *Proceedings of SPIE Medical Imaging Conference 2014*, volume 9035, 2014.

[10] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, 2013.

[11] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *arXiv preprint arXiv:1603.05959*, 2016.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] D. Kumar, A. Wong, and D. A. Clausi. Lung nodule classification using deep features in ct images. In *Computer and Robot Vision (CRV), 2015 12th Conference on*, pages 133–138. IEEE, 2015.

[15] H. Lee and Y.-P. P. Chen. Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications*, 42(12):5356–5365, 2015.

[16] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

[17] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress, 2010.

[18] A. C. Society. *Cancer Facts & Figures 2015*. Atlanta.

[19] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, pages 1139–1147, 2013.

[20] M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis. A novel computer-aided lung nodule detection system for ct images. *Medical Physics*, 38(5630), 2011.

[21] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[22] H. K. Weir, M. Thun, B. Hankey, L. Ries, H. Howe, P. Wingo, A. Jemal, E. Ward, R. Anderson, and B. Edwards. Annual report to the nation on the status of cancer, 1975-2000. *Journal of the National Cancer Institute*, 95(17):1276–1299, 2003.

[23] X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh, and G. Beddoe. Shape-based computer-aided detection of lung nodules in thoracic ct images. *Biomedical Engineering, IEEE Transactions on*, 56(7):1810–1820, 2009.

[24] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, et al. On rectified linear units for speech processing. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3517–3521. IEEE, 2013.