Lung Nodule Detection in CT Images using Deep Convolutional Neural Networks

Rotem Golan
Dept. of Computer Science
University of Calgary
Calgary, Alberta, Canada T2N 1N4
Email: grotem@ucalgary.ca

Christian Jacob
Dept. of Computer Science
Dept. of Biochemistry & Molecular Biology
University of Calgary
Calgary, Alberta, Canada T2N 1N4
Email: cjacob@ucalgary.ca

Jörg Denzinger
Dept. of Computer Science
University of Calgary
Calgary, Alberta, Canada T2N 1N4
Email: denzinge@cpsc.ucalgary.ca

***Abstract*—Early detection of lung nodules in thoracic Computed Tomography (CT) scans is of great importance for the successful diagnosis and treatment of lung cancer. Due to improvements in screening technologies, and an increased demand for their use, radiologists are required to analyze an ever increasing amount of image data, which can affect the quality of their diagnoses. Computer-Aided Detection (CADe) systems are designed to assist radiologists in this endeavor. Here, we present a CADe system for the detection of lung nodules in thoracic CT images. Our system is based on (1) the publicly available Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) database, which contains 1018 thoracic CT scans with nodules of different shape and size, and (2) a deep Convolutional Neural Network (CNN), which is trained, using the back-propagation algorithm, to extract valuable volumetric features from the input data and detect lung nodules in sub-volumes of CT images. Considering only those test nodules that have been annotated by four radiologists, our CADe system achieves a sensitivity (true positive rate) of 78.9% with 20 false positives (FPs) per scan, or a sensitivity of 71.2% with 10 FPs per scan. This is achieved without using any segmentation or additional FP reduction procedures, both of which are commonly used in other CADe systems. Furthermore, our CADe system is validated on a larger number of lung nodules compared to other studies, which increases the variation in their appearance, and therefore, makes their detection by a CADe system more challenging.**

I. Introduction

According to the American Cancer Society, lung cancer is the leading cause of cancer related deaths in the United States [1]. It was estimated that lung and bronchus cancers alone will cause 158,040 deaths in the United States in the year 2015, which is slightly more than the total number of deaths caused by brain/nervous system, breast, prostate, colon/rectum, and liver cancers combined. In addition, the 5-year survival rate of lung cancer is one of the lowest compared to other types of cancer, and is estimated to be 18% for years 2004 to 2010. Individual prognosis heavily depends on the extent of disease at the time of diagnosis. So for example, if the tumor is detected while it is still small and localized, then the 5-year survival rate is 54%; but if it is detected at a later stage, when metastases have already developed and the tumor becomes regional or distant, then the survival rate drops to 27% and 4% accordingly.

Unfortunately, most diagnoses occur at later stages of the disease, mainly due to lack of symptoms in its early stages. This has raised the idea of instituting a widespread lung cancer screening as a matter of public health policy, which has been examined by a number of research institutions. One of these institutions is the U.S. National Cancer Institute (NCI), which sponsored the National Lung Screening Trial (NLST) and concluded that there is a statistically significant 20.3% relative reduction in lung cancer mortality when using low-dose helical computed tomography (LDCT) scans as a screening modality compared to using chest x-ray [2]. Therefore, it has been suggested to make LDCT, and CT in general, the preferred screening modality for early detection and diagnosis of lung cancer.

A thoracic CT scan combines a series of X-ray images taken from different angles, and uses computer processing to create cross-sectional images, or slices, of the bones, blood vessels and soft tissues inside the chest. This results in a 3 dimensional image of the chest, where each volumetric pixel (voxel) has an attenuation value that is indicative to the type of material (tissue) present in its location.

As the resolution of CT screening technologies increases, and their demand, especially in the developing world, is on the rise, radiologists are overwhelmed with the amount of data they are required to analyze [3]. This has the potential to cause fatigue among radiologists, and therefore, affect the quality of their diagnoses. Computer-Aided Detection (CADe) systems have been developed in recent years to assist radiologists with this challenge [4], [5], [6], [7]. Their goal is to provide radiologists with a second opinion, and support them in their interpretation of medical images. For example, a successful CADe system might detect lung nodules, which would otherwise be overlooked by the radiologist, and bring these to the attention of the radiologist for further examination. However, if a CADe system produces too many false positives (FPs), the radiologist's trust in the system can be undermined.

In this paper we propose a CADe system for the detection of lung nodules in thoracic CT images. Our system is based on (1) the publicly available Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) database, which contains 1018 thoracic CT scans of

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

individuals at different stages of their disease, and (2) a deep Convolutional Neural Network (CNN), which is trained, using the back-propagation algorithm [8], to detect lung nodules in sub-volumes of CT images. The deep CNN is composed of two parts. Its first part is designed to extract valuable volumetric features from the input data, and is composed of multiple volumetric convolution, rectified linear units (ReLU), and max pooling layers. The second part of the CNN is the classifier. It is composed of multiple fully connected and threshold layers, followed by a softmax layer, and is expected to perform the high-level reasoning of the neural network. A simplified version of this CNN is illustrated in Figure 1.

This paper is organized as follows. Section II examines the properties of the LIDC-IDRI dataset which, in addition to CT images, includes valuable annotations about the various lung nodules it contains. Section III describes the architecture of our CNN, and how it is used to compute a 3D voting grid with which we predict the location and boundary of lung nodules. Section IV describes the results of our CADe system using a Free-response Receiver Operating Characteristic (FROC) analysis, and compares them to those obtained in previous work where the LIDC-IDRI dataset has also been used for validation. Finally, Section V concludes our work and discusses some future work.

A. Classification of CADe systems

In a comprehensive survey conducted by Suzuki et al. [9], three classes of classification techniques, into which the various CADe systems can be categorized, were identified. These are (1) Feature-based Machine Learning (FML), (2) Pixel/voxel-based Machine Learning (PML), and (3) Non-ML-based methods. Our CADe system belongs to the PML class of classification techniques since its volumetric features are trained in a supervised manner from the input data (i.e. voxel values of CT images). This is different from other CADe systems [4], [5], [6], [7], which belong to the FML class of classification techniques since their features are predetermined and are set manually by experts in the field. Non-ML-based methods are defined as methods that do not use ML techniques. This includes all methods that do not have a "learning from examples" component in them.

The detection of lung nodules in CT scans is no easy task, and one has to overcome the significant variability in the input data when approaching this task. First, since CT scanners are manufactured by different companies and are deployed with a wide range of radiation doses, they can vary in the way image reconstruction is performed and in the amount of image noise being generated [10]. Furthermore, CT scanners can be deployed with different configurations, so the images they produce can have values, such as slice thickness, pixel spacing, and image orientation, that are different from one another. Second, the size and shape of normal anatomical structures in the scans varies among different patients, and the CT images might contain artificial artifacts such as pace makers and artificial valves. Finally, lung nodules can vary in their appearance, from round solid objects to flat and liquid-like objects.

Modeling objects in CT images, such as lesions and organs, based on a simple model with a relatively small number of parameters, is unlikely to be sufficient to represent their complex structures. This means that Non-ML-based classifiers are probably not the right approach when tackling this task. However, FML and PML techniques have the potential of producing more complex models based on training examples, and therefore, have a better chance of overcoming this challenge of variability.

B. Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a type of feed-forward neural network where the individual neurons are tiled in such a way that they respond to overlapping regions in its receptive field. It is inspired by models of the biological visual system, proposed by [11], and continues to be consistent with the modern understanding of the physiology of the visual system [12]. The first computational models based on these local connectivities between neurons are found in Fukushima's Neocognitron [13]. Fukushima found that when neurons with the same parameters are applied on overlapping regions of the previous layer, at different locations, a form of translational invariance is obtained. This allows CNNs to detect objects in their receptive field in a way that is invariant to their size, location, orientation, and other visual properties. In addition, this limited connectivity of CNNs reduces the computational requirements necessary for their training compared to fully-connected neural networks.

CNNs were first trained using the back-propagation algorithm in [14], and ever since they have obtained state-of-the-art performance on several pattern recognition tasks. For example, large-scale CNNs were used to recognize objects in natural images as part of the ImageNet challenge [15], [16], [17], for which they have shown significant improvement in performance compared to other approaches. Another notable work is by Ciresan et al. [18], who demonstrated improved records on MNIST [19], [20], Latin letters [21], Chinese characters [22], traffic signs [23], NORB [24] and CIFAR10 [25] benchmarks using deep CNNs. A key question to our work is whether the success deep CNNs have had in other computer vision tasks also applies to the detection of lung nodules in CT scans and, in general, to the analysis of medical images.

II. Material

The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) database is a publicly available reference for the medical imaging research community [26], [27], [28]. Its aim is to support the development of computer-aided diagnostic methods for lung nodule detection and classification, and is the result of a collaboration between seven academic centers and eight medical imaging companies.

The LIDC-IDRI dataset contains 1018 thoracic CT images that originated from a total of 1010 patients. The im-

ages comply with the Digital Imaging and Communications in Medicine (DICOM) standard, and have a resolution of $[65, 764]\times512\times512$ voxels per scan, where $[65, 764]$ is the range of values for the number of slices in the 3D images, and 512×512 is the in-plane pixel resolution of each of the 2D slices. The average number of slices per scan in the dataset is 240, and the range of values for the slice thickness parameter is $[0.45, 5]$ mm with an average slice thickness of 1.73 mm. The range of values for the pixel spacing in each of the 2D slices is $[0.46, 0.97]$ mm with an average of 0.68 mm. These values help us understand the relationship between the image space (measured in voxels) and the real-world space (measured in mm).

Each scan has been examined by four experienced thoracic radiologists in a two-phase image annotation process. In the initial blinded-read phase, each radiologist was asked to independently review the images in the dataset, and mark lesions they identified as (1) nodules $\geq$ 3 mm, (2) nodules $<$ 3 mm, and (3) non-nodule $\geq$ 3 mm. In this paper, we only used the annotations for the first category of nodules, namely nodules $\geq$ 3 mm, which include all those nodules in the dataset with greatest in-plane dimension in the range $[3, 30]$ mm regardless of presumed histology, and for which the complete contour of each nodule was marked. In the second subsequent unblinded-read phase, each radiologist was asked to review their own marks, along with the anonymized marks of the three other radiologists, and make a final decision. The annotation files that are published as part of the LIDC-IDRI dataset contain only those markings from the second phase of the annotation process.

The goal of this two-phase annotation process was to identify as many lung nodules in each CT scan without forcing a consensus. But since no consensus was forced, one has to devise a grouping procedure which will determine which nodule markings represent the same lung nodule and which are not. Failure to do so properly can impact the validation of any CADe system. For example, let us examine the case where a wrong grouping procedure considers two markings, that are taken from the same nodule and are made by two different radiologists, to represent two separate nodules. Then, if a correct nodule prediction is made which "hit" both markings, then the number of true positives will falsely increase by 2 instead of 1; if no nodule prediction is made to "hit" these two markings, then the number of false negatives will falsely increase by 2 instead of 1.

Here, we propose a simple and effective grouping procedure which leads to a one-to-one correspondence with the nodule-count-by-patient file that was recently released by the LIDC-IDRI team [28]. This file contains ground-truth information regarding the number of nodules $\geq$ 3 mm that are found in each CT image. Our grouping procedure can be described as follows. Given the ground-truth number of nodules in each CT scan, and the number of nodule markings made by the four radiologists in each CT scan, we repeatedly group together the closest pair of nodules, in terms of the distance between their 3D centers, until the above two values are equal. During this procedure, we make sure that no two nodule markings of the same radiologist are grouped together, since we assume that a single radiologist will not mark the same nodule more than once.

TABLE I: Properties of the LIDC-IDRI dataset at four agreement levels.

Agreement level	totalNod	maxNodPerScan	avgNodPerScan
1	2670	23	2.62
2	1886	13	1.85
3	1395	12	1.37
4	908	8	0.89

The authors in [4], [5], [6], [7] did not have this file available to them at the time their work was conducted, and therefore, they had no ground-truth information with which they can validate the quality of their grouping procedures. Their grouping procedures are based on either the distance between centers of nodules or the overlap between nodules, but in any case, they had to incorporate a distance or overlap threshold which will determine which nodule markings are grouped together and which are not. This variability in grouping procedures leads to a less accurate comparison between the various CADe systems.

Once the grouping of nodule markings is complete, we can associate each nodule with one of four agreement levels. Agreement level j, where $1\leq j\leq4$, includes all those nodules which were marked by at least j radiologists. We expect nodules at a higher agreement level to be more easily detectable by a CADe system since they were identified by more radiologists. Table I describes, for each agreement level, the total number of nodules in the dataset (totalNod), maximum number of nodules per scan (maxNodPerScan), and average number of nodules per scan (avgNodPerScan). The minimum number of nodules per scan for all agreement levels is zero.

III. Our Computer-Aided Detection System

In this section, we describe our CADe system and discuss the steps required for its construction. First, we use back-propagation [8] to train the weights of a deep CNN. The input of the CNN is (1) a CT image sub-volume of size $5\times20\times20$, (2) positional information of the sub-volume in relation to the entire CT image, and (3) some parameters of the DICOM image. The output of the CNN is a value in the range $[0, 1]$, representing its estimate to whether the sub-volume contains a lung nodule.

Then, given a previously unprocessed three dimensional CT image of size $D\times512\times512$ ($D\in[65, 764]$), we apply the CNN multiple times throughout the CT image using a $5\times20\times20$ sliding window, and compute a three dimensional voting grid of size $D\times512\times512$ by averaging the outputs of the CNN in the various sliding window positions. Each entry in this voting grid provides us with an estimate, in the range $[0, 1]$, to whether its corresponding voxel is part of a lung nodule. We use the voting grid, together with two thresholds (threshold A and B), to predict the location and boundary of lung nodules.

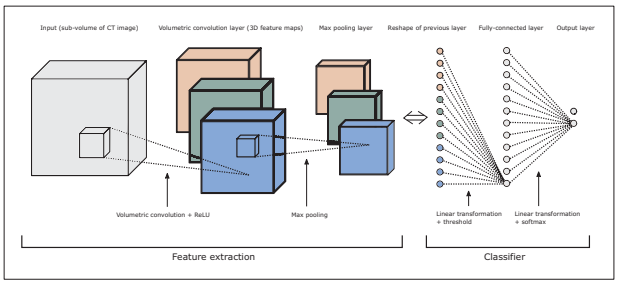

Fig. 1: A simple CNN which has similar layers to those used in our CNN. Each color represents the use of a different 3D feature kernel.

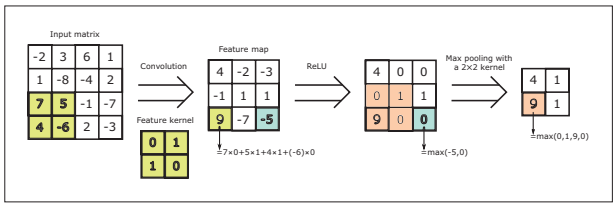

Fig. 2: A sequence of 2D convolution, ReLU, and max pooling operations. The stride value of both the convolution and max pooling operations, along the two axes, is 1.

We do so by considering each adjacent set of entries, which have values greater than threshold B, and have at least one value greater than threshold A, to represent a predicted lung nodule.

A. Training a deep CNN

Our deep CNN is trained, using the back-propagation algorithm, to detect lung nodules in CT image sub-volumes of size $5\times20\times20$. It is composed of two parts. The first part is designed to extract valuable volumetric features from the input data, and is composed of multiple volumetric convolution, rectified linear units (ReLU), and max pooling layers. The second part of the CNN is the classifier. It is composed of multiple fully connected and threshold layers, followed by a softmax layer, and is expected to perform the high-level reasoning of the neural network. Figure 1 illustrates a simple CNN which has similar layers to those used in our CNN. Figure 2 illustrates a sequence of two dimensional convolution, ReLU, and max pooling operations.

Our deep CNN can be described by listing its layers in sequence, from input to output, as follows: $5\times20\times20 - 96C3\times9\times9 - MP1\times2\times2 - 256C2\times4\times4 - MP2\times2\times2 - 384C1\times3\times3 - 384C1\times3\times3 - 256C1\times3\times3 - MP1\times2\times2 - 4096FC - 4096FC - 2FC$, where, for example, $5\times20\times20$ is the input layer (the receptive field) of the CNN, $96C3\times9\times9$ represents a volumetric convolution layer with 96 feature kernels of size $3\times9\times9$, $MP1\times2\times2$ represents a max pooling layer with a kernel of size $1\times2\times2$, and $4096FC$ represents a fully-connected layer with 4096 units.

The stride value of both the volumetric convolution and max-pooling layers, along the three axes, is set to 1. We use a rectified linear activation function (ReLU) for the convolutional layers, a threshold activation function (threshold = 1e-6) for the fully connected layers, and a softmax function for the output layer.

Finally, in addition to the activations of its previous layer, the first fully-connected layer of the CNN receives 7 additional values. The first 3 values represent positional information of the receptive field in relation to the entire CT image and for each of the three axes; the last 4 values represent information regarding the DICOM image, namely slice thickness (in mm), pixel spacing in each of the two in-plane axes (in mm), and the image orientation.

We chose the receptive field (input layer) of the CNN to be $5\times20\times20$ in the following manner. First, we started with a receptive field of size $1\times76\times76$ since the maximum in-plane diameter of a lung nodule in the dataset is 76. After training the CNN with this receptive field, and examining its results, we continued to examine smaller input sizes, and noticed that doing so leads to a better performance of the CNN. This improvement in accuracy stopped when we reached a receptive field of size $1\times20\times20$. Then, we increased the depth (i.e. number of slices) of the receptive field and noticed that this again leads to better performance, which has stopped when a depth value of 5 was reached. We hypothesis that this improvement in results is due to the fact that lung nodules are three dimensional objects, and having a receptive field that captures multiple image slices at a time allows the CNN to extract volumetric features that are essential for the detection of lung nodules. Consequently, we chose $5\times20\times20$ to be the receptive field of our CNN.

In order to understand the relationship between the receptive field of our CNN, of size $5\times20\times20$, and the size of the various nodules in the dataset, we examine the following statistics. The maximum/minimum/average in-plane diameter of a nodule in the dataset is 76/1/15 pixels, respectively, and the maximum/minimum/average depth (i.e. number of slices) of a nodule in the dataset is 56/1/6 voxels, respectively. This means that the receptive field of our CNN can either be contained inside a nodule or it can contain an entire nodule. Consequently, our CNN is expected to learn volumetric features that are present both inside lung nodules and in their surroundings. It is important to note that in order to preserve the original values of the DICOM images as much as possible, no scaling was applied to the CT images of the dataset.

Our CNN is trained using the back-propagation algorithm, where the LIDC-IDRI dataset is used to generate the training and test sets. We generate these sets by using a 80/20 ratio to split the CT images of the LIDC-IDRI dataset. This leads to a training set of size 814, and a test set of size 204, out of the 1018 images of the dataset. Having an independent test set ensures that the test results of our CADe system are good indicators for its generality and robustness.

During training, the sub-volumes are randomly extracted from the CT images of the training set, and are normalized according to an estimate of the normal distribution of the voxel values in the dataset. A sub-volume of size $5\times20\times20$

is considered to be a nodule instance only if at least one of its slices contains, or is fully contained in, a complete 2D region of interest of a nodule. Alternatively, a sub-volume is considered to be a non-nodule instance only if none of its voxels is part of a nodule. The sampling from both the training and test set is done in a balanced fashion. This means that, on average, there is a similar number of non-nodule and nodule instances being extracted from these sets, and consequently, that our CNN processes a similar number of non-nodule and nodule instances during training and validation.

Training ends after 70 epochs of 4000 batches each, where the size of each batch is 128. The batch size represents the number of input instances that are being processed by the CNN in each iteration of the back-propagation algorithm. Having a batch size of 1 leads to a pure stochastic gradient decent algorithm since the weights of the neural network will be updated after each processing of a single training instance. Alternatively, a batch size that is greater than 1, but is less than the size of the training set, leads to a mini-batch gradient decent algorithm. The learning rate is determined according to a pre-defined scheme which starts with a value of 0.01 and then gradually decreases until reaching a value of 0.0005. Testing ends after randomly extracting and processing 1024 sub-volumes from each image in the test set.

Finally, there are two issues that commonly arise when training Deep Neural Networks (DNNs). These are overfitting and computation time. Overfitting occurs when a statistical model captures patterns and regularities that are present in the training set but are not found outside of it. As a result, once such a model is applied to a previously unprocessed instance of the problem, it would not perform as well as it would on an instance from the training set. It generally happens when a model is excessively complex, such as having too many parameters relative to the number of instances in the training set.

Training a DNN is computationally expensive, and one has to take this into consideration when deciding which size of network to use and which hardware to run it on. We executed our experiments on a Tesla K80 GPU, which improved the running-time of our system by a factor of 33.3% compared to only utilizing the CPU. More specifically, training a single CNN, including the time required for validating the CNN's performance on the test set every 10 epochs, takes about 40 hours to complete. Alternatively, performing the same computation with CPU only takes about 60 hours to complete. To make this possible, we used Torch [29], a scientific computing framework with wide support for machine learning algorithms, and Nvidia's cuDNN [30], a GPU-accelerated library of primitives for DNNs.

B. Predicting the location and boundary of lung nodules

Once the training of a CNN is complete, and given a previously unprocessed three dimensional CT image of size $D\times512\times512$ ($D\in[65, 764]$), we apply the CNN multiple times throughout the CT image using a $5\times20\times20$ sliding window, and compute a three dimensional voting grid of size $D\times512\times512$ by averaging the outputs of the CNN in the various sliding window positions. More specifically, the receptive field of the CNN is moved throughout the CT image in an ordered fashion so that it is applied to all the voxels in the image. For the in-plane axes, the receptive field is moved 10 voxels at a time, and for the depth axis, it is moved 1 voxel at a time. Whenever a CNN is applied, its output value is added to all the entries in the voting grid that correspond to its receptive field. Then, each entry in the voting grid is divided by the number of times the CNN was applied to its corresponding voxel.

Each entry in the resulting voting grid provides us with an estimate, in the range $[0, 1]$, to whether its corresponding voxel is part of a lung nodule. We use this voting grid, together with two thresholds (threshold A and B), to predict the location and boundary of lung nodules. We do so by considering each adjacent set of entries, which have values greater than threshold B, and have at least one value greater than threshold A, to represent a predicted lung nodule.

In section IV, we describe how different values of threshold A are examined in order to compute the FROC curve of a CADe system. Once the FROC curve is ready, it can be used to determine the final value of threshold A by taking into consideration the tradeoff between the sensitivity and false positive per scan values of the CADe system. Then, this final value of threshold A can be used in real time when the CADe system is deployed on previously unprocessed CT images.

Furthermore, threshold B has been fixed to a value of 0.7, and all experiments in this work assume this value. Decreasing the value of threshold B leads to an increased size of nodule predictions made by our CADe system. This needs to be carefully controlled and limited so that our CADe system does not produce nodule prediction that are too big. Such nodule predictions will not be of much help to radiologists since one of the primary goals of any CADe system is to provide its users with a minimal approximation to where lung nodules might be present. Failure to so can undermine the radiologists' trust in the system. Consequently, the value of threshold B was chosen so that it results in an average size of nodule predictions that is similar to the average size of lung nodules in the dataset. For example, given that threshold B equals to 0.7, the average size of nodule predictions made by our CADe system is 2558 voxels, while the average size of lung nodules at agreement level 4 in the test set is 7060 voxels.

Figure 3 illustrates a cross-sectional image of a lung nodule that has been annotated by all four radiologists. Its boundary is marked in green, and the nodule prediction boundary, made by our CADe system, is marked in red. It demonstrates that our CADe system is able to successfully identify the center area of this lung nodule without marking too many excessive pixels that are outside of it.

IV. Experimental Results

In this section, we will first describe the results of training the above CNN in four different configurations, one for each of the four agreement levels. Each CNN has been trained to detect

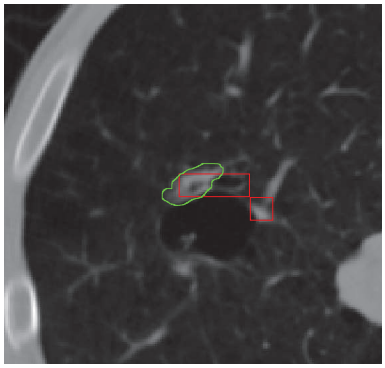

Fig. 3: A cross-sectional image of a lung nodule. The nodule boundary is marked in green, and the nodule prediction boundary, made by our CADe system, is marked in red.

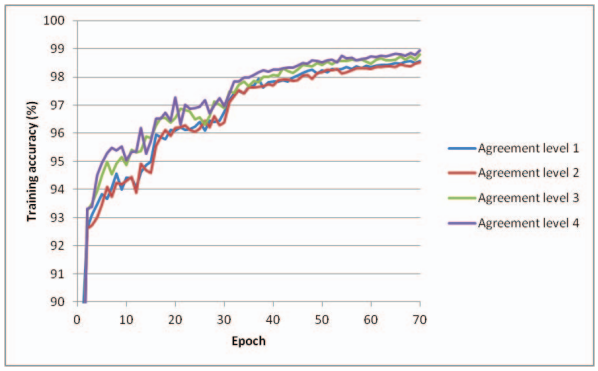

Fig. 4: Training accuracy of four CNNs, one for each agreement level, during 70 epochs of training.

lung nodules, at the appropriate agreement level, in CT image sub-volumes of size $5\times20\times20$. Then, we present a Free-response Receiver Operating Characteristic (FROC) analysis for our CADe system, and compare it to some previous work where the LIDC-IDRI dataset has also been used for validation.

The training accuracy of these four CNNs during the 70 epochs of training is shown in Figure 4. The test accuracy of these four CNNs during training is shown in Figure 5. These figures show a high level of accuracy for all agreement levels. More specifically, it shows that the higher the agreement level, the better the accuracy for both the training and test set. This makes sense since lung nodules that are at higher agreement levels were identified by more radiologists, and are often times larger compared to nodules at lower agreement levels, and therefore, are more easily detectable by our CADe system.

Furthermore, Figures 4 and 5 show that there is not a significant difference between the training accuracy and test accuracy during the training of any of the CNNs. This is an indication that our CNNs do not suffer from overfitting, or in other words, that the size of our CNN and the number of epochs used for its training are set reasonably for the number of training instances we have available.

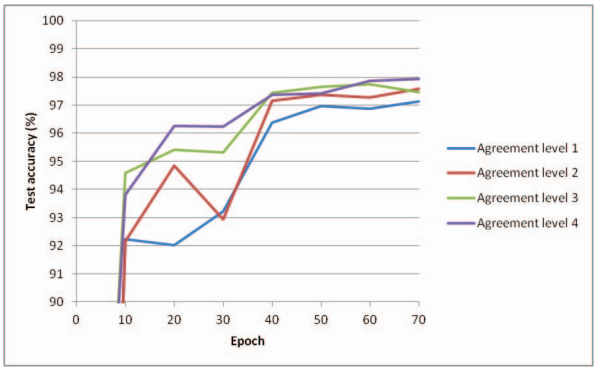

Fig. 5: Test accuracy of four CNNs, one for each agreement level, during 70 epochs of training.

A Free-response Receiver Operating Characteristic (FROC) curve is a tool for characterizing the performance of a free-response system at all decision thresholds simultaneously [31]. A CADe system is considered to be a free-response system since its aim is not just to predict whether a medical image contains an abnormality or not, but to predict the exact location and boundary of an undefined number of abnormalities in the image. If the former was true, a conventional ROC curve would suffice to characterize the performance of the system. The FROC curve was first introduced in [32], where it was used to visualize the performance of a free-response task from the auditory domain. Its importance for radiology applications was first recognized by [33], and ever since it has been widely used to characterize the performance of CADe systems and other localization tasks.

Figure 6 illustrates the FROC curves of our CADe system in four different configurations, one for each of the four agreement levels, based on the images in the test set. It shows that our CADe system is able to achieve a sensitivity (true positive rate) of 78.9% with 20 false positives (FPs) per scan, or a sensitivity of 71.2% with 10 FPs per scan, on lung nodules that have been annotated by all four radiologists (i.e. nodules at agreement level 4). Furthermore, it shows that the performance of our system on lung nodules at agreement level 1 is lower compared to higher agreement levels, which is reasonable since these nodules have been annotated by only one radiologist, and therefore, are expected to be harder to detect.

Assuming some threshold A, we define the number of false positives (FPs) in each CT scan to be the number of nodule predictions, made by our CADe system, that do not contain any voxel of an annotated lung nodule. Also, we define the true positive (TP) value to be the number of lung nodules that have been successfully detected by our system, and the false negative (FN) value to be the number of lung nodules that have not been detected by our system. Given the TP and FN values, the sensitivity (true positive rate) of a CADe system can be computed according to the following definition:

$$sensitivity = \frac{TP}{TP + FN}$$

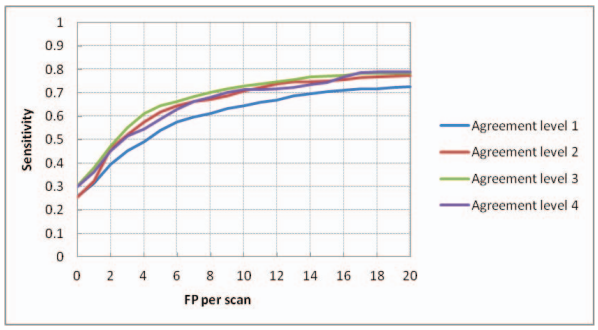

Fig. 6: FROC curves of our CADe system for nodules in the test set at four agreement levels.

A single FROC curve is plotted by computing, for a range of threshold A values, the sensitivity and FPs per scan values, and using linear interpolation to connect between the computed points. More specifically, for each CT scan, we initialize the value of threshold A with the maximal value in the corresponding voting grid. Then, after computing the sensitivity and FP values for this specific value of threshold A, we decrease threshold A by 0.01 which usually leads to more FPs being produced by the system. This process continues until the number of FPs that the system produces is greater than 20. Finally, for each FP per scan value, the sensitivity values are averaged among all the CT images in the test set.

The total number of CT images in the test set is 204, all of which are used to validate the performance of a single CNN during its training. However, when computing the FROC curve of the entire CADe system, some of these images are ignored. This is because we can only consider CT images that contain at least one lung nodule when performing a FROC analysis. CT images that contain no nodules are ignored since they will always have a sensitivity of zero, regardless of the number of FPs the system produces on them. Consequently, the number of CT images in the test set (and the number of nodules they contain), for agreement level 1 to 4, is 181 (507), 159 (361), 147 (295), and 119 (204), respectively.

We compare our results to those obtained in [4], [5], [6], [7], where the LIDC-IDRI dataset has also been used for validation. Table II describes the number of true positives in relation to the number of nodules in the test set, sensitivity, and FPs per scan values of our CADe system compared to four other studies. Our CADe system and the CADe systems described in [5], [7] were validated on an independent test set, while the CADe systems described in [4] and [6] were validated using a 2-fold and a 7-fold cross-validation test, respectively.

Our results show a comparable sensitivity value, at the expanse of a higher FP per scan value, in respect to other studies. However, in contrast to other CADe systems, our system does not utilize any lung segmentation or additional false positive reduction procedures. These procedures have shown to reduce the amount of FPs that are produced by a CADe system, and therefore, adding them to our CADe system can improve its results. Furthermore, our CADe system is validated on a larger number of lung nodules compared to other studies. More specifically, it is validated on 204 lung nodules, while other studies are validated on between 38 and 143 lung nodules. This increases the variation in the appearance of the various nodules in the test set, and therefore, makes their detection by a CADe system more challenging.

TABLE II: Summary of results for five CADe systems.

Study	True positives	Sensitivity	FP per scan	Agreement level
Our CADe system	161/204 | 145/204	78.9% | 71.2%	20 | 10	4
Riccardi et al. [4]	83/117	71%	6.5	4
Golosio et al. [5]	30/38	79%	4	4
Messay et al. [6]	118/143	82.7%	3	1
Tan et al. [7]	70/80	87.5%	4	4

V. Conclusion and Future Work

In this paper, we apply a deep learning approach for the problem of detecting lung nodules in CT scans. More specifically, we train a deep Convolutional Neural Network (CNN) to detect lung nodules in sub-volumes of CT images, and use it to predict the location and boundary of lung nodules in previously unprocessed CT images. The CNN is composed of a total of 20 layers, which can be partitioned into two parts. The first part of the CNN is designed to extract valuable volumetric features from the input data, and is composed of multiple volumetric convolution, ReLU, and max pooling layers. The second part of the CNN is the classifier. It is composed of multiple fully connected and threshold layers, followed by a softmax layer, and is expected to perform the high-level reasoning of the neural network.

Our proposed voxel-based CADe system achieves a sensitivity of 78.9% with 20 false positives (FPs) per scan, or a sensitivity of 71.2% with 10 FPs per scan, on lung nodules that have been annotated by all four radiologists. These results are comparable to previous work in terms of sensitivity, but are not as good in terms of their FP per scan value. However, our CADe system does not include any lung segmentation or additional FP reduction procedures, both of which are commonly used in other CADe systems and have the potential to improve the results of our system. Furthermore, our CADe system is validated on a larger number of lung nodules compared to other studies, which increases the variation in their appearance, and therefore, makes their detection by a CADe system more challenging.

The question of how well can deep CNNs be applied to the detection of lung nodules in CT images and, in general, to the analysis of medical images, is key to this work. In other words, one of the goals of this work is to examine whether the success deep CNNs have had in other computer vision tasks also applies to the detection of lung nodules in CT scans. To answer these questions, further research needs to be conducted, but based on the results presented here, we conclude that the use of CNNs can indeed be a promising avenue of research in the field of CADe systems.

We propose a number of ways to improve the performance of our CADe system. First, using a larger dataset, which contains a greater number of lung nodules of different shape and size, can improve the robustness and accuracy of our system. Second, segmenting the boundaries of the lungs can decrease the number of FPs that our system predicts by allowing it to ignore those nodule predictions that are outside of the lungs. Third, implementing an additional FP reduction step can be of great value. This can be done by collecting all the TP and FP instances that our CADe system produces, and use a separate machine learning algorithm to distinguish between the two. Finally, since our CADe system uses a voting grid to predict the location and boundary of lung nodules, it is possible to use an ensemble of classifiers, in addition to our CNN, to determine the values of the voting grid.

ACKNOWLEDGMENTS

We thank Zebra Medical Vision for their support in writing this paper and providing us with the hardware required to execute our experiments. Also, we 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 [26], [27], [28]. Finally, we thank Alberta Innovates Technology Futures (AITF), who funded this research as part of their Graduate Student Scholarships program.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA: a cancer journal for clinicians*, vol. 65, no. 1, pp. 5–29, 2015.

[2] B. S. Kramer, C. D. Berg, D. R. Aberle, and P. C. Prorok, "Lung cancer screening with low-dose helical ct: results from the national lung screening trial (nlst)," *Journal of medical screening*, vol. 18, no. 3, pp. 109–111, 2011.

[3] R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, B. J. Erickson, and D. F. Kallmes, "The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload," *Academic radiology*, vol. 22, no. 9, pp. 1191–1198, 2015.

[4] A. Riccardi, T. S. Petkov, G. Ferri, M. Masotti, and R. Campanini, "Computer-aided detection of lung nodules via 3d fast radial transform, scale space representation, and zernike mip classification," *Medical physics*, vol. 38, no. 4, pp. 1962–1971, 2011.

[5] B. Golosio, G. L. Masala, A. Piccioli, P. Oliva, M. Carpinelli, R. Cataldo, P. Cerello, F. De Carlo, F. Falaschi, M. E. Fantacci *et al.*, "A novel multithreshold method for nodule detection in lung ct," *Medical physics*, vol. 36, no. 8, pp. 3607–3618, 2009.

[6] T. Messay, R. C. Hardie, and S. K. Rogers, "A new computationally efficient cad system for pulmonary nodule detection in ct imagery," *Medical Image Analysis*, vol. 14, no. 3, pp. 390–406, 2010.

[7] M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis, "A novel computer-aided lung nodule detection system for ct images," *Medical physics*, vol. 38, no. 10, pp. 5630–5645, 2011.

[8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, p. 3, 1988.

[9] K. Suzuki, "Machine learning in computer-aided diagnosis of the thorax and colon in ct: A survey," *IEICE transactions on information and systems*, vol. 96, no. 4, pp. 772–783, 2013.

[10] R. S. Maia, C. Jacob, A. K. Hara, A. C. Silva, W. Pavlicek, and M. J. Ross, "An algorithm for noise correction of dual-energy computed tomography material density images," *International journal of computer assisted radiology and surgery*, vol. 10, no. 1, pp. 87–100, 2015.

[11] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, p. 106, 1962.

[12] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, "A quantitative theory of immediate visual recognition," *Progress in brain research*, vol. 165, pp. 33–56, 2007.

[13] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

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

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42, 2014.

[18] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

[19] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] P. J. Grother, "Nist special database 19 handprinted forms and characters database," *National Institute of Standards and Technology*, 1995.

[22] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Chinese handwriting recognition contest 2010," in *Pattern Recognition (CCPR), 2010 Chinese Conference on*. IEEE, 2010, pp. 1–5.

[23] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 1453–1460.

[24] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–97.

[25] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[26] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *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*, vol. 38, no. 2, pp. 915–931, 2011.

[27] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[28] T. C. I. A. Team, "Data from lidc-idri," 2015. [Online]. Available: http://dx.doi.org/10.7937/K9/TCIA.2015.LO9QL9SX

[29] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: a modular machine learning software library," IDIAP, Tech. Rep., 2002.

[30] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cudnn: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.

[31] D. P. Chakraborty, "A brief history of free-response receiver operating characteristic paradigm data analysis," *Academic radiology*, vol. 20, no. 7, pp. 915–919, 2013.

[32] H. Miller, "The froc curve: a representation of the observer's performance for the method of free response," *The Journal of the Acoustical Society of America*, vol. 46, no. 6B, pp. 1473–1476, 1969.

[33] P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free response approach to the measurement and characterization of radiographic observer performance," in *Application of Optical Instrumentation in Medicine VI*. International Society for Optics and Photonics, 1977, pp. 124–135.