Lung Nodule Classification Using Deep Features in CT Images

Devinder Kumar
Systems Design Engineering
University of Waterloo
Waterloo,Canada
Email: devinder.kumar@uwaterloo.ca

Alexander Wong
Systems Design Engineering
University of Waterloo
Waterloo,Canada
Email: alexander.wong@uwaterloo.ca

David A. Clausi
Systems Design Engineering
University of Waterloo
Waterloo,Canada
Email: dclausi@uwaterloo.ca

***Abstract*—Early detection of lung cancer can help in a sharp decrease in the lung cancer mortality rate, which accounts for more than 17% percent of the total cancer related deaths. A large number of cases are encountered by radiologists on a daily basis for initial diagnosis. Computer-aided diagnosis (CAD) systems can assist radiologists by offering a second opinion and making the whole process faster. We propose a CAD system which uses deep features extracted from an autoencoder to classify lung nodules as either malignant or benign. We use 4303 instances containing 4323 nodules from the National Cancer Institute (NCI) Lung Image Database Consortium (LIDC) dataset to obtain an overall accuracy of 75.01% with a sensitivity of 83.35% and false positive of 0.39/patient over a 10 fold cross validation.**

***Keywords*-Computer-aided diagnosis (CAD), LIDC, deep features, autoencoder, lung nodule**

I. Introduction

Lung cancer is the leading cause of cancer deaths in North America and worldwide among both men and women [1], [2]. The number of deaths caused due to lung cancer is more than prostate, colon and breast cancers combined. Also, most patients detected with lung cancer today are already at an advanced stage as lung cancer is hard to detect in early stages [2]. The reason for failure in detecting lung cancer in early stages is that there is only a dime-sized lesion growth known as nodule, inside the lung, and by the time it is detected it is already too late for the patient. Also, these small lesions cannot be detected by X-rays and are only detectable by a CT scan. Even after the detection, it takes a considerable amount of effort and experience on the part of radiologists to detect and label the nodules as benign or as a probable case of malignancy. Considering the large number of cases encountered by radiologists every day there is a constant pressure on them to analyse a huge amount of data and make a decision as quickly as possible based on the analysis.

A possible solution to decrease this burden on the radiologists is to use computer aided diagnosis (CAD) systems as a *second* opinion that can automatically detect and analyse lung nodules in CT images. Some of the studies in the past have shown an improvement in radiologists performance through the use of these CAD systems [3], [4].

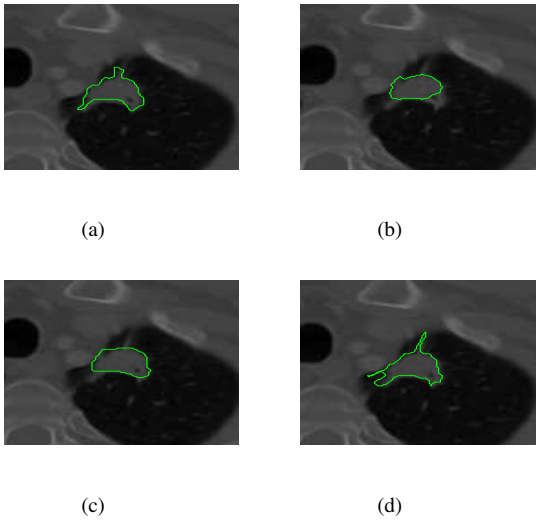

Figure 1. Illustration of annotations provided by four different radiologists for the same lesion in a single slice of CT study taken from LIDC-IDRI dataset.

To create such CAD systems, there is always a need for a reference standard dataset that can be used to obtain ground truths and can also act as a basis for comparison of different CAD algorithms. LIDC (Lung Cancer Data Consortium) [5], [6] is one such database which contains CT images of thoracic region for 1010 patients along with the annotation data of suspicious nodules (for both benign and malignant cases) for a size greater than 3 mm from up-to four radiologists collected over a long period of time. The dataset also contains diagnostic data for a limited number of cases (157 patient) obtained from biopsy, surgical resection, progression or reviewing the radiological images to show 2 years of nodule size.

Another important criteria to create an effective CAD system for classification is to use a feature or combination of features that can effectively represent the structure and characteristics of region of interest in images. Recent studies have shown that the use of deep learning can significantly improve the performance of CAD systems. Deep learning architectures are effectively formed by combination of many

linear and non linear transformations to obtain more abstract and useful representations of data [7]. S. Liu et al. [8] report a significant improvement of over 20% in the overall accuracy over the earlier methods by using deep learning architecture for early diagnosis of alzheimer's disease; Cruz-Roa et al. [9] reported an improvement of around 7% over canonical representations while using a deep learning architecture for image representation and automatic basal cell carcinoma cancer detection. Xu Yan et al. [10] present the effectiveness of using deep neural networks (DNN) for feature extraction in medical image analysis and report that the performance of an automatic unsupervised approach for feature extraction from DNN is as effective as a supervised approach. The main reason for outstanding performances of deep learning architectures like DNN is that the higher level of features are derived from lower level features to form a hierarchical representation.

In this paper, we propose to use *deep* features extracted from an autoencoder along with a binary decision tree as a classifier to build a CAD system (Fig. 2) for lung cancer classification. In the first step of our CAD system, nodules are extracted from the 2D CT images using the annotations provided by different radiologists. These nodule areas are then individually fed into our autoencoder and learned features are then extracted from layer 4 of the 5 layer autoencoder. These features are then used as input to the trained classifier (binary decision tree in this case), which gives us the classification results for each nodule in the test set. To perform the classification we use CT scans for all the patients in the LIDC dataset for which diagnostic data was available as a ground truth for comparison purposes. The next section explains the related work done in recent past in the field of lung cancer classification and the use of deep learning in medical imaging. The methodology to extract the deep features is described in Section III . Further, the implementation methodology is explained in Section IV and the experimental results obtained by using the proposed CAD system is discussed in Section V. We conclude this paper by summarizing our work and by discussing future avenues in Section VI.

II. Related Work

In the past, several methods have been proposed to detect and classify lung cancer in CT images using different algorithms. For example, Camarlinghi et al. [11] have used three different computer aided detection techniques for identifying pulmonary nodules in CT scans. Abdulla and Shaharum [12] used feed forward neural networks to classify lung nodules in X-Ray images albeit with only a small number of features such as area, perimeter and shape. Kuruvilla et al. [13] have used six distinct parameters including skewness and fifth & sixth central moments extracted from segmented single slices containing 2 lungs along with the features mentioned in [12] and have trained a feed forward back propagation neural network with them to evaluate accuracy for different features separately. In Bellotti et al. [14], the authors have proposed a new computer-aided detection system for nodule detection using active contour based model in CT images. The paper reports a high detection rate of 88.5% with a average of 6.6 false positives (FPs) per CT scan on 15 CT scans. In the recent past a comparison between six different methods for detecting nodules in lungs was done by Ginneken et al. [15] that also proposed a method to combine the output of multiple systems for effectively detection of pulmonary nodules. In Riccardi et al. [16] the authors presented a new algorithm to automatically detect nodules with an overall accuracy of 71% using 3D radial transforms.

In the recent years, there has also been a renewed interest in the field of deep learning and the latest research in area of medical imaging using deep learning shows promising results. One such study is of Suk et al. [17] in which the authors propose a novel latent and shared feature representation of neuroimaging data of brain using Deep Boltzmann Machine (DBM) for AD/MDC diagnosis. The methods outperforms the competing methods and achieve a maximal diagnostic accuracy of 95.52% (AD vs. NC); Wu et al. [18] use deep feature learning for deformable registration of brain MR images demonstrating that a general approach can be built to imporve image registration by using deep features. A stacked autoencoder (a type of deep learning architecture) was used by Fakoor et al. [19] to diagnose and classify different types of cancer based on gene expression data, which eventually out performs contemporary methods for different datasets. To the best of our knowledge there has been no work that uses deep features for lung nodule classification.

The work that is closest to the proposed work is of Zinovev et al. [20]. In this paper, the authors propose to use belief decision trees for the classification of lung nodules in LIDC dataset. They use different features such as lobulation, texture, spiculation etc. to create a 63 dimensional feature vector for the classification of each of the 914 instances (1 instance/nodule). The paper reports an overall average accuracy of 68.66%.

III. Methodology

A. LIDC Dataset: Data Extraction

The Lung Image Database Consortium (LIDC) [5], [6] has made a database publically available that contains thoracic CT images of 1010 patients of lung cancers along with annotations (nodules outlines) from up-to four radiologists. Even though annotation are provided for over a thousand patients, the diagnostic data is only available for 157 patients containing information about ratings of nodules (0-Unknown,1-benign,2-Primary malignant,3-Malignant(metastatic)). The ratings were obtained by performing biopsy, surgical resection, progression or reviewing the radiological images to show 2 years of nodule state at two levels; first at the

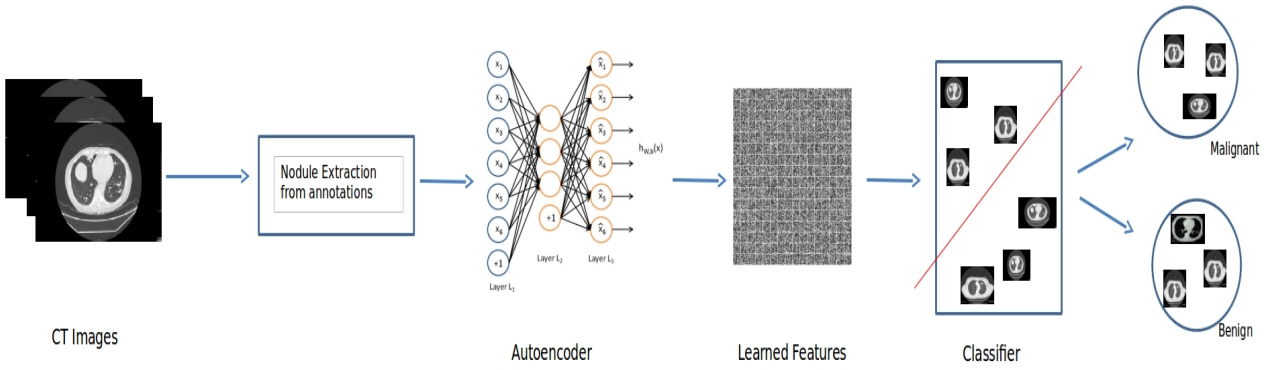

Figure 2. Illustration of different modules of the proposed CAD system.

patient level and second diagnosis at the nodule level. The LIDC database of thoracic CT studies for 1010 patients was acquired over a long period of time with various scanners. As there can be multiple nodules associated with a single study, a different number of slices may be associated with a particular nodule study. Also, each slice can contain annotations data for a nodule from up-to four radiologists. These annotations are available in LIDC dataset for only those nodules which are greater than 3 mm in size. As the diagnostic data is the only way to judge the certainty of malignancy, we chose to use the ratings from diagnostic data as the ground truth for training the classification system and evaluating the results instead of using the radiologists provided ratings in dataset.

B. Autoencoder: Feature Extraction

An autoencoder is primarily a two layer network that takes an input $f \epsilon [0,1]^d$ and then uses a linear or non-linear transformation to "encode" the data to a latent space. On the output layer, it uses a "decoding" transformation to reconstruct the data. Stating precisely we want to learn a representation:

$$l(f(i); W, b) = \phi(W f(i) + b) \tag{1}$$

such that $\phi(W^T l(f(i); W, b) + c)$ is approximately $f^{(i)}$.

$$\min_{W,b,c} \sum_{i=1}^{n} \parallel \phi(W^T \phi(W f^{(i)} + b) + c) - f^{(i)} \parallel_2^2 \tag{2}$$

For penalizing the error between input and output l-2 norm is used. Normally sigmoid or hyperbolic tangent functions are used as the activation function in autoencoder. In autoencoder the cost function can be accurately determined which results to a possibility of using more advance methods of optimization, such as L-BFGs [21] for training the networks. Since the optimization function in deep neural networks could be non convex, pre-training the filters greedily allows a way to trick the optimization objective by starting from a point that is likely to be closer to the optima [22]. The selection of network configuration parameters (number of hidden layers,iteration set, batch size etc.) seems to be somewhat an ad-hoc process with it been still an active topic of research.

Extraction of deep features using an autoencoder is explained in detail in the following section.

IV. Implementation

For each nodule greater than or equal to 3 mm in diameter in the LIDC dataset, we extracted the annotations provided by the radiologists to be used later to extract features from the autoencoder (Section III B.). To extract the annotations, we used the same annotation extractor as used in Lampert et al. [23]. Using the annotaion extractor, we extracted annotations of 157 patients. Figure 1 shows the extracted annotations by four radiologists for a single nodule present in a study. It is interesting to note the different degree of variations in the annotation, from just a region inside the image (Fig. 1(b)) to an actual outline (Fig. 1(d)). Unlike many past methods which use only the largest area or the best outline, we used all the available annotations for feature extraction and classification. The reasons of using all available annotations is explained in Section V.

For extracting features we use a five layered de-noising auto-encoder trained by L-BFGS with a iteration set to 30 and batch size of 400, as the parameters seemed to work well for many datasets reported in past [24]. To extract features, we first created an adaptive window based on the nodule size to construct a rectangular window based on the max and min of (x,y) co-ordinates enclosing the nodule. We then resized each rectangular area to a fixed dimension to create a fixed length input for the auto-encoder. We gave this input to the auto-encoder and trained our network for 4303 instances (1 instance = 1 slice containing nodules) containing 4323 nodules for patients with rating 0, 1, 2, and 3 in the diagnostic dataset. The rating 0 i.e., unknown is treated as malignant for our experiments as it is better to flag such cases for doctors instead of ignoring them. Then we extracted features from different hidden and output layer of

autoencoder. The features from layer 4 of autoencoder of 5 layers were used to create a feature vector of 200 dimensions for each instance.

To evaluate the performance of the CAD system, we use a binary decision tree as a classifier as it can handle missing information in the input as well. 200 dimensional features for each of 4323 nodules were given as input to the decision tree and classification into benign and malignant classes were obtained.

V. Experimental Results

For evaluating the performance of our proposed CAD system which uses deep features, we used the 200 dimensional feature vector obtained from the layer 4 of our autoencoder architecture (Section III B.) for 4323 nodules from diagnostic dataset of the large LIDC dataset. Each annotation marking provided by each radiologist is considered to be one instance in a slice, which leads to up-to 4 annotations for one slice for certain cases. One of the advantages of this approach of using the annotation of all the radiologists as compared to many other similar approaches which use the best annotation for a slice or the annotation with largest area is that while testing in the real world, if one of the radiologists provide a partial annotation of a nodule (which can happen often), our system would still work. For example, if instead of annotation shown in Fig. 1(d), if Fig. 1(b) is given as input to our system, our system will still be able to provide reasonable output, where as systems which rely only on lobulation or spiculation as features would fail. As mentioned in section III, the diagnostic dataset had rating of 0, 1, 2, and 3 from two levels of inspection. We decided to take the ratings of the second level i.e., ratings at the nodule level because the second rating were the final ratings obtained from biopsy, surgical inspections or rate of growth which are considered conventionally to be the best methods for determining malignancy. We treated the rating 1-Benign as Benign and combined the rating 0-Unknown, 2-Primary malignant and 3- malignant (metastatic) as malignant. For the cases where level 2 rating wasn't available, we took the ratings of level 1. For all the cases associated with the benign and malignant, we extracted *200d* deep features. We used these deep features as input to the binary decision tree to perform the binary classification. We obtained an overall accuracy of 75.01% with a sensitivity of 83.25% and a false positive of 0.39 per patient (FP/patient) over 10- fold cross validation on our dataset. For obtaining the results we used 90% of dataset as training set and the rest 10% as testing set. This set-up was chosen similarly to the method used in Zinovev et al. [20] for the purpose of a better comparison as it is the state-of-the-art classification results for the LIDC dataset, to best of our knowledge. Zinovev et al. [20] use a 63 dimensional feature vector containing features such as lobulation, malignancy, etc to achieve an overall accuracy of 68.66%, measuring the area under the curve as a metric for

Table I
Evaluation of Deep Features Classification Results from 10 Fold Cross Validation Experiments

	Accuracy(%)	**Sensitivity(%)**	**FP/patient**
1	75.05	83.92	0.40
2	75.93	84.42	0.39
3	73.14	81.41	0.38
4	76.02	84.56	0.40
5	72.71	80.65	0.33
6	75.98	86.38	0.41
7	75.76	83.56	0.40
8	74.67	82.35	0.42
9	75.32	80.95	0.38
10	75.54	85.37	0.39
Avg	**75.01**	**83.35**	**0.39**

Table II
Comparison of Results Obtained from Deep features Classification Technique and Belief Decision Tree Classification Technique [20]. The best results are highlighted in bold.

	Deep Features	Belief Decision Trees
Accuracy	**75.01%**	68.66%

914 instances from 154 patients of LIDC dataset. The results from the cross-validation study and a comparison with [20] is shown in Table I and II respectively. As noted in the Table II, our proposed CAD system reports an increase of about 6% in the overall accuracy compared to the results report by Zinovev et al. [20].

It is important to note here that there is a considerable difference between the part of LIDC dataset used in our proposed method and in [20]. In Zinovev et al. [20] the authors use the rating provided by the radiologists which corresponds to level 1 in our diagnostic data rating whereas we use the rating at level 2 i.e. rating at nodule level, which is generally considered as the final decisive rating in the domain of medical diagnosis. The reason for using the level 2 rating instead of just the radiologists provided rating was to create & test a system that can be obtained to replicate or at-least obtain comparable results along the lines of diagnosis data. It should also be noted that we obtained a false positive rate of 0.39 per patient which points towards a low specificity of the system. The main reason for such results is that many benign cases were visually very similar to the malignant cases as shown in Figure 3, which led to such results.

VI. Conclusion

In this paper, we presented a CAD classifier system for classifying lung nodules as either malignant or benign. The proposed system uses deep features extracted from an autoencoder for annotations provided by up-to four radiologists for 157 patients to precisely create a strong representation of

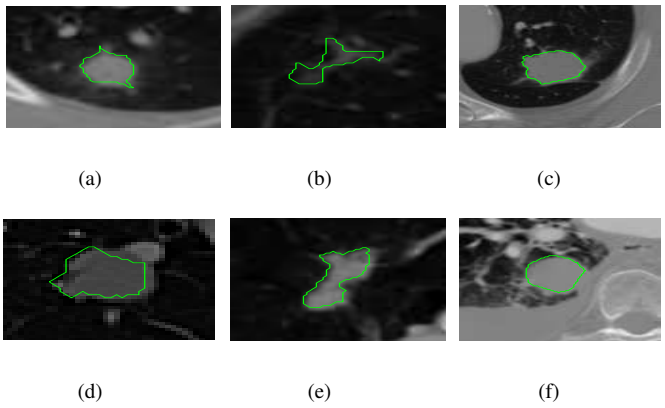

Figure 3. Examples of annotations for benign (a,b,c) and malignant cases (d,e,f) in the LIDC dataset. It can be observed that there are significant visual similarities between the annotated nodules in (a,d), (b,e) and (c,f), making it very difficult to differentiate between such nodules during the classification process. As such, such cases result in a high number of false positives obtained by the proposed CAD system. In general, from a clinicial decision support perspective, it is more important to catch all malignant cases and as such false positives are of lesser concern than false negatives.

nodules. Using the LIDC dataset, we showed that the proposed system convincingly outperforms the state-of-the-art method on overall accuracy metric even after experimenting with almost five times the data size (4323 vs. 914) used in the state-of-the-art method and considering the biopsy level clinical decision as ground truth. This is because the deep features not only take the different conventional semantic features like lobulation, spiculation etc. in to account but they also take into account the association between them.

In terms of future work, we plan to extend the proposed CAD system's capabilities by integrating the automatic detection of nodules module in it.

ACKNOWLEDGEMENT

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs Program, and the Ontario Ministry of Research and Innovation.

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. Data used in this research were obtained from The Cancer Imaging Archive (TCIA) sponsored by the Cancer Imaging Program, DCTD/NCI/NIH. Also, the authors would like to thank Prof. H.R. Tizhoosh for the numerous discussions on the dataset.

REFERENCES

[1] D. M. Parkin, "Global cancer statistics in the year 2000," *The lancet oncology*, vol. 2, no. 9, pp. 533–543, 2001.

[2] P. B. Bach, J. N. Mirkin, T. K. Oliver, C. G. Azzoli, D. A. Berry, O. W. Brawley, T. Byers, G. A. Colditz, M. K. Gould, J. R. Jett *et al.*, "Benefits and harms of ct screening for lung cancer: a systematic review," *Jama*, vol. 307, no. 22, pp. 2418–2429, 2012.

[3] F. Li, H. Arimura, K. Suzuki, J. Shiraishi, Q. Li, H. Abe, R. Engelmann, S. Sone, H. MacMahon, and K. Doi, "Computer-aided detection of peripheral lung cancers missed at ct: Roc analyses without and with localization 1," *Radiology*, vol. 237, no. 2, pp. 684–690, 2005.

[4] G. D. Rubin, J. K. Lyo, D. S. Paik, A. J. Sherbondy, L. C. Chow, A. N. Leung, R. Mindelzun, P. K. Schraedley-Desmond, S. E. Zinck, D. P. Naidich *et al.*, "Pulmonary nodules on multi–detector row ct scans: Performance comparison of radiologists and computer-aided detection 1," *Radiology*, vol. 234, no. 1, pp. 274–283, 2005.

[5] 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.

[6] S. G. Armato III, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon *et al.*, "Lung image database consortium: Developing a resource for the medical imaging research community 1," *Radiology*, vol. 232, no. 3, pp. 739–748, 2004.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[8] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of alzheimer's disease with deep learning," in *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*. IEEE, 2014, pp. 1015–1018.

[9] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer, 2013, pp. 403–410.

[10] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, E. I. Chang *et al.*, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1626–1630.

[11] N. Camarlinghi, I. Gori, A. Retico, R. Bellotti, P. Bosco, P. Cerello, G. Gargano, E. L. Torres, R. Megna, M. Peccarisi *et al.*, "Combination of computer-aided detection algorithms for automatic lung nodule identification," *International journal of computer assisted radiology and surgery*, vol. 7, no. 3, pp. 455–464, 2012.

[12] A. A. Abdullah and S. M. Shaharum, "Lung cancer cell classification method using artificial neural network," *Information Engineering Letters*, vol. 2, no. 1, pp. 49–59, 2012.

[13] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for ct images," *Computer methods and programs in biomedicine*, vol. 113, no. 1, pp. 202–209, 2014.

[14] R. Bellotti, F. De Carlo, G. Gargano, S. Tangaro, D. Cascio, E. Catanzariti, P. Cerello, S. C. Cheran, P. Delogu, I. De Mitri *et al.*, "A cad system for nodule detection in low-dose lung cts based on region growing and a new active contour model," *Medical Physics*, vol. 34, no. 12, pp. 4901–4910, 2007.

[15] B. van Ginneken, S. G. Armato, B. de Hoop, S. van Amelsvoort-van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. Schilham, A. Retico, M. E. Fantacci *et al.*, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study," *Medical image analysis*, vol. 14, no. 6, pp. 707–722, 2010.

[16] 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.

[17] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative *et al.*, "Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[18] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of mr brain images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer, 2013, pp. 649–656.

[19] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "using deep learning to enhance cancer diagnosis and classification," in *Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare (WHEALTH). Atlanta, GA*, 2013.

[20] D. Zinovev, J. Feigenbaum, J. Furst, and D. Raicu, "Probabilistic lung nodule classification with belief decision trees," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 4493–4498.

[21] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[22] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle *et al.*, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[23] T. A. Lampert, A. Stumpf, and P. Gançarski, "An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation," *arXiv preprint arXiv:1307.0426*, 2013.

[24] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng, "On optimization methods for deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 265–272.