Fine Grain Lung Nodule Diagnosis Based on CT using 3-D Convolutional Neural Network

Qiuli Wang[1], Jiajia Zhang[1], Sheng Huang[1], Chen Liu[2], Xiaohong Zhang[1], and Dan Yang[1]

[1] School of Big Data & Software Engineering, Chongqing University, 400000, Chongqing , China
{wangqiuli, gagazhang, huangsheng, xhongzh, dyang}@cqu.edu.cn
[2] Radiology Department, The First Affiliated Hospital of Army Medical University,400000, Chongqing, China
cqliuchen@foxmail.com

Abstract. As the core step of lung nodule analysis, lung nodule diagnosis comprises two important tasks: False Positive Reduction (FPR) and Malignancy Suspiciousness Estimation (MSE). Many studies tackle these two tasks separately. However, these two tasks share a lot of similarities and have connections with each other, since MSE is the successive step of FPR, and both tasks can be deemed as the lung nodule labeling problems. In this paper, we split the label 'real nodule' defined in FPR into two new finer grain labels, namely 'low risk' and 'high risk', which are defined in MSE. In such way, we merge these two separated issues into a unified fine grain lung nodule classification problem. Finally, a novel Attribute Sensitive Multi-Branch 3-D CNN (ASMB3DCNN) is proposed for performing the fine grain lung nodule classification. We evaluate our model on LIDC-IDRI and LUNA16 datasets. Experiments demonstrate that ASMB3DCNN can efficiently address the two tasks above in a joint way and achieve the promising performances in comparison with the state-of-the-arts.

Keywords: Joint Learning · Lung Nodule Diagnosis · Convolutional Neural Network · Computed tomography (CT) · Computer-aided detection and diagnosis (CAD)

1 Introduction

Each year, there are 8.2 million deaths caused by cancer in the worldwide. Lung cancer accounts for the highest number of mortalities i.e. 1.59 million [**?**]. However, according to statistics, most patients diagnosed with lung cancer today already have advanced disease (40% are stage IV, 30% are stage III), and the current 5-year survival rate is only 16% [**?**], which indicates that early diagnosis and treatment can effectively improve survival chance of lung cancer patients. As the core step of lung nodule analysis, classifying a large number of detected nodules by the radiologists, which includes False Positive Reduction (FPR) and Malignancy Suspiciousness Estimation (MSE), can be very time-consuming.

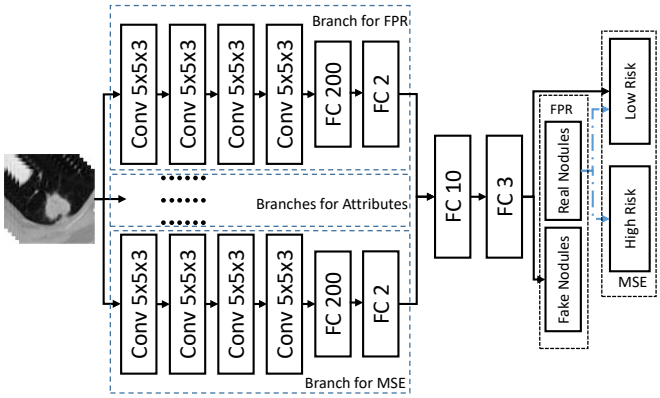

Fig. 1. Architecture of ASMB3DCNN

Small nodules are very difficult to be found and can be confused with normal tissues. Long time reading work can also cause the fatigue of the radiologists and reduce the work efficiency. Thus, developing a fast, robust and accurate CAD system to perform automated diagnosis of lung nodules is meaningful and important[**?**].

To the best of our knowledge, most studies considered FPR and MSE as two totally separated problems. In the task of FPR, Qi Dou et al. [**?**] designed a multilevel contextual 3-D Convolutional Neural Networks (3-D CNN) to encode richer spatial information and extract more discriminative representations via the hierarchical architecture trained with 3-D samples. In the task of MSE, Wei Shen et al. [**?**] exploited CNN to differentiate lung nodules and proposed a Multi-crop CNN network structure. Sarfaraz Hussein [**?**] used multi-task learning model based on 3-D CNN, which showed that information of high-level attributes can help to improve the performance of the model. Botong Wu[**?**] designed PN-SAMP, which can also provide related evidences and segmentation information to radiologists. Shiwen Shen [**?**] used CNN to capture high level features and splited model into branches to learn different semantic information, these information can help to classify malignant nodules. Jason Causey et al.[**?**] proposed NoduleX. In Nodulex, they leveraged a deep CNN (more than 10 CNN layers) to tackle FPR and MSE tasks separately, and they also proved that these two tasks are similar.

Extensive studies show that 3-D CNN is a powerful approach for addressing both MSE and FPR issues in a separated way [**?**][**?**][**?**][**?**], since it is deemed as the best choice for keeping 3-D spatial information in the CT scans [**?**]. As MSE is the successive step of FPR, the label 'real nodule' defined in FPR can be further divided into two new finer grain labels, namely 'low risk' and 'high risk'. Moreover, these two tasks share lots of high-level attributes.

According to the findings above, we intent to merge these two tasks into a unified fine grain lung nodule diagnosis problem and present a novel Attribute Sensitive Multi-Branch 3-D CNN (ASMB3-DCNN) to address this unified issue. The architecture of ASMB3DCNN is shown in Fig **??**.

Our main contributions are in four-folds:

(1) We analyze backgrounds of False Positive Reduction (FPR) and Malignancy Suspiciousness Estimation (MSE) tasks, and merge them into a unified task.

(2) We measure the sensitivities of attributes and select the most reliable attributes to yield a attribute sensitive version of 3-D CNN for fine grain lung nodule diagnosis.

(3) We design a method of normalization to capture different sizes of inputs according to the slices-thickness and pixel-spacing which has been empirically proved that such strategy can improve accuracy of classification and reduce the burden of calculation.

(4) Beyond FPR and MSE,we extend the proposed approach to predict nodule attributes,which could potentially assist radiologists in evaluating malignancy uncertainty.

The rest of paper is organized as follows: Section **??** introduces the methodology of our works; experiments are presented in section **??**; the conclusion is finally summarized in section **??**.

2 Method

Since Malignancy Suspiciousness Estimation (MSE) is the successive task of False Positive Reduction (FPR) in the lung nodule diagnosis, MSE essentially performs a further classification on the positive samples labeled by FPR. In such way, we can further categorize the positive samples into two finer categories, namely 'low risk' and 'high risk', and these two tasks can be unified as a fine grain lung nodule classification issue. In this paper, we present a novel 3-D CNN named Attribute Sensitive Multi-Branch 3-D CNN (ASMB3DCNN) to address such issue and then the tasks of FPR and MSE can be jointly tackled.

2.1 Architecture of ASMB3DCNN

There are eight branches in the proposed Attribute Sensitive Multi-Branch 3-D CNN(ASM3DCNN), and all branches share the same architecture as shown in Fig **??**. After many experiments, the best architecture of all is carried out: each branch consists of four 3-D convolutional layers (the kernel sizes are $5 \times 5 \times 3$, $5 \times 5 \times 3$, $5 \times 5 \times 3$, $1 \times 1 \times 1$, and the number of kernels are 64, 128, 256, 256 respectively), two fully-connected layers, and one softmax layer for giving an initial binary prediction to a specific attribute or a category. In these eight branches, there are two branches used for predicting false positive and malignancy suspiciousness of a lung nodule respectively, and the other six branches are used for predicting the selected attributes, which have been empirically verified to be helpful for lung nodule diagnosis [**?**]. We select the attributes based on the sensitivity analysis of attributes. This part will be discussed later in the subsection **??**. We fuse the outputs of these branches via two fully-connected layers and present the final classification via a softmax layer.

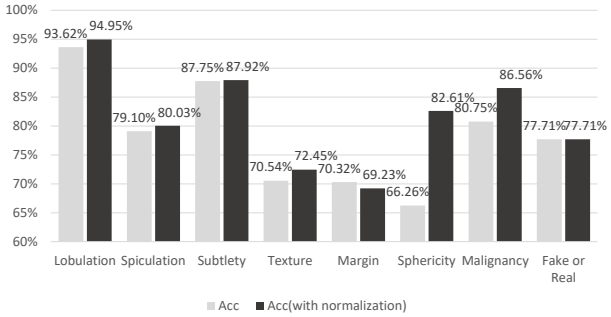

Fig. 2. Comparison of Accuracy After Normalization

Table 1. Proportion of Malignant Nodules in Each Level

Attributes Category	**Attributes Rate**						**Sensitivity (Variance)**	**Attribute Rank**
	1	2	3	4	5	6		
subtlety	2.56%	5.92%	5.53%	17.23%	44.61%	-	0.030	5
internalStructure	18.90%	47.62%	50.00%	0.00%	0.00%	-	0.060	1
calcification	0.00%	0.00%	0.00%	1.80%	10.81%	21.88%	0.057	2
sphericity	0.00%	17.28%	23.00%	20.79%	8.60%	-	0.009	6
margin	20.56%	23.83%	28.36%	21.80%	9.78%	-	0.004	7
lobulation	6.77%	26.67%	54.55%	56.12%	13.79%	-	0.052	3
spiculation	7.47%	31.39%	58.70%	57.39%	41.94%	-	0.045	4
texture	11.56%	20.35%	20.86%	23.57%	18.41%	-	0.002	8

In the training phase, we adopt the pre-training + fine-tuning strategy to train our network.Eight branches are trained separately and then the branches are combined with the fusion layers to perform a fine-tuning.Why we use this strategy to train the network will be discussed later in 3.6.

2.2 Normalization

The slice-thickness ranges from 0.6 mm to 5.0 mm, the pixel-spacing ranges from 0.4609375 mm to 0.9765625 mm. Wei Shen et al. [**?**] normalized images using spline interpolation to have a fixed resolution with 0.5 mm/voxel along all three axes, all images will be re-sampled before cropping nodules, which is time-consuming. Moreover, coordinates of nodules will be affected during re-sampling.

To reduce the burden of computation, we adjust the input size according to slice-thickness and pixel-spacing and resize the input after cropping. The length, width and height of inputs are set to 30mm (In clinical, nodules larger than 30mm in diameter are called lung masses and will not be discussed here). The larger the slice-thickness is, the fewer slices are needed. The larger the pixel-spacing is, the fewer pixels are needed. The coordinates of nodules are not affected.

Then we resize the input into $20 \times 20 \times 10$ using spline interpolation because ASMB3DCNN needs a size-fixed input. Fig **??** shows that after normalization, the accuracy of each branch is improved in LIDC-IDRI dataset.

Table 2. Number of Nodules in Each Level

Attributes	**Attributes Rate**					
Category	1	2	3	4	5	6
subtlety	117	321	506	1097	594	-
internalStructure	2609	21	2	2	1	-
calcification	3	2	187	111	74	2258
sphericity	5	191	613	1419	407	-
margin	107	277	342	1101	808	-
lobulation	1433	855	220	98	29	-
spiculation	1647	704	138	115	31	-
texture	173	113	139	488	1722	-

2.3 Sensitivity Analysis of Attributes

[**?**] indicates that the attributes can facilitate the solution of MSE. However, not all attributes are sensitive to (or strongly correlated to) the level of malignancy suspiciousness of lung nodules. Thus, here we intend to analyse the sensitivities of attributes to malignancy suspiciousness level via measure the variance of the proportion of high risk lung nodules under different levels of attributes as follows:

$$\mathcal{S} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1}$$

where n is the number of score levels, X_i is the proportion of malignant nodules, $\bar{X}$ is the average of proportion.

There are 8 attributes, namely 'subtlety', 'internalStructure', 'calcification', 'sphericity', 'margin', 'lobulation', 'spiculation' and 'texture'. We measure the sensitivity of these attributes in training set of LIDC-IDRI dataset, rank them by their sensitivities and tabulate the results in Table **??**. We find the top 2 sensitive attributes are 'internalStructure' and 'calcification' whose sample distributions are extremely unbalanced as shown in Table **??**. This makes us very hard to train the reliable predictors for these two attributes via using the training samples, and the unreliably predicted attributes may corrupt the performance of the lung nodule diagnosis system. Thus, we choose 'subtlety', 'lobulation', 'speculation' (their ranks are from 3 to 5) as complementary information for lung nodule classification. In order to prove that our selection of attributes can actually help the model, we also train the model with 'subtlety', 'lobulation', 'spiculation' ,'sphericity','margin' and 'texture', and results shown in Table **??** prove that model with 3 selected attributes is batter than model with 6 attributes. There are examples of nodules with different level attributes in Tabel **??**, nodules framed with red rectangles are high risk nodules.

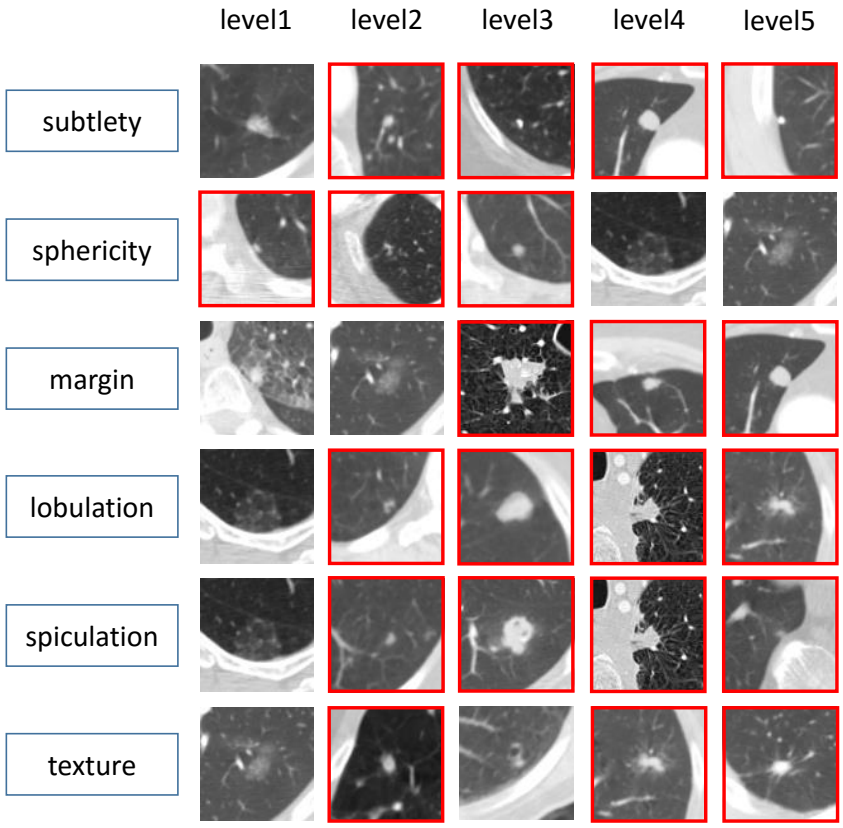

Fig. 3. Nodules with different level attributes

3 Experiments and Results

3.1 Datasets

The Lung Image Database Consortium Image Collection (LIDC-IDRI) [**?**], is a public-available dataset, consisting of 1010 patients with chest CT scans. This dataset provides 36378 nodules and 2635 of them are analyzed by four experienced radiologists. Because of the small number of data, we rotate nodules 90°, 180°, and 270° within the transverse plane.

Lung Nodule Analysis 2016 (LUNA2016) [**?**] is a challenge for lung nodules detection and FPR. Its dataset is built based on LIDC-IDRI, while it removes nodules which are less than 3mm in diameter and provides more than 700 thousand fake nodules for FPR. Since then, we train the branches of six attributes and MSE using LIDC-IDRI dataset while train the branch of FPR using LUNA2016 dataset (samples from LIDC-IDRI dataset + fake nodule samples).

Table 3. Comparison with Other Studies for MSE

Methods	***Accuracy***	***Sensitivity***	***Specificity***	***AUC Score***
ASMB3DCNN(ALL)	0.961	0.900	0.991	0.965
ASMB3DCNN(MSE+Attributes)	**0.978**	**0.957**	0.993	**0.976**
ASMB3DCNN(MSE)	0.866	0.696	**0.999**	0.845
NoduleX[**?**]	0.932	0.879	0.985	0.971
Fuse-TSD[**?**]	0.895	0.842	0.920	0.966
MC-CNN[**?**]	0.871	0.77	0.93	0.93
HSCNN[**?**]	0.842	0.705	0.889	0.856

Table 4. Comparison with Other Studies for FPR

Methods	***Accuracy***	***Sensitivity***	***Specificity***	***AUC Score***
ASMB3DCNN(ALL)	**0.974**	0.935	0.992	0.967
ASMB3DCNN(FPR+Attributes)	0.969	0.898	0.996	0.969
ASMB3DCNN(FPR)	0.777	0.697	**0.997**	0.753
Multilevel 3D-CNN[**?**]	-	0.827	-	-
UACNN(ccanoespinosa)	-	0.824	-	-
NoduleX(CNN47+RF)[**?**]	0.946	**0.948**	0.943	**0.984**

3.2 Malignancy Suspiciousness Estimation

Following the study [**?**][**?**], we exclude the nodules whose average malignancy scores are equal to 3. We respectively label the nodules whose average malignancy scores higher than 3 and lower than 3 as 'high risk nodules' and 'low risk nodules'. Following the study [**?**], in the branches of attribute predictions, we binarize the malignancy levels of attributes. More specifically, we label the nodules with level equal to or higher than 3 as 'high', and the ones with level lower than 3 as 'low'. Therefore, each branch of attribute prediction can be considered as a naive binary classifier. Table **??** shows the performances of three versions of ASMB3DCNN in comparison with the state-of-the-arts. ASMB3DCNN(MSE), which uses the MSE branches only, is actually degraded as an ordinary single steam 3-D CNN. It obtains 86.6% in accuracy and 84.5% in AUC score respectively, while ASMB3DCNN (MSE+ Attributes) fuses MSE branches with six attribute branches, which obtains 11.2% and 11.1% gains in accuracy and AUC score respectively over ASMB3DCNN (MSE). These observations verify that the high-level attributes can indeed bring a boost in the performance of MSE. ASMB3DCNN(ALL) assembles all eight branches. It performs less worse than ASMB3DCNN (MSE+Attributes) but gets a similar performance to NoduleX. It is not hard to understand this phenomenon, since FPR performs a superclass labeling instead of a simple classification from the perspective of the MSE task. More specifically, the category 'real nodule' in FPR, which can be seen as a superclass, includes all two categories of MSE, namely 'low risk' and 'high risk'. In such way, the output of FPR branch actually has not offered any useful

information for discriminating the 'high risk' and the 'low risk' data, and even corrupts the performances of the whole system.

3.3 False Positive Reduction

We conduct experiments via following the same experimental settings for nodule labeling in section **??**. As Table **??** shows, ASMB3DCNN(FPR) obtains 77.7% in accuracy and 75.3% in AUC score respectively, while the ones of ASMB3DCNN (FPR+Attributes) are 96.9% and 96.9%, which shows the significant improvement over ASMB3DCNN (FPR). This phenomenon also indicates that the high-level attributes can benefit the solution of FPR problem. Another interesting observation is that ASMB3DCNN(ALL) outperforms ASMB3DCNN(FPR+Attributes) particularly in accuracy and sensitivity. The gains of ASMB3DCNN(ALL) over the FPR+Attributes version are 0.5% and 3.7% respectively. This verifies that the labels in a finer grain level is helpful to the solution of a classification task in a coarse level of labels. It also confirms the reasonability of our idea that these two separated issues can be merged into a unified fine grain lung nodule classification problem for better solving them together. Moreover, our proposed approaches have higher accuracy and specificity than other compared approaches. Although our approaches get slightly lower scores in comparison with NoduleX in sensitivity and AUC, NoduleX suffers from a heavier computation burden due to its deeper architecture. NoduleX has more than ten convolutional layers and max-pooling layers, while our models have only 4 convolutional layers.

Table 5. Results of ASMB3DCNN(ALL)

Method	***Acc of FN***	***Acc of LR***	***Acc of HR***
Fine Grain	0.957	0.996	0.923
MSE & FPR	0.992	0.991	0.900

3.4 Nodule Attributes Prediction Results

We perform nodule attributes prediction via following the same experimental setting in section **??**. Fig **??** presents the classification performance for each of the attribute-branches. We achieved mean accuracy(with normalization) of 94.95%, 80.03%, 87.92%, 72.45%, 69.23%, 82.61% for lobulation, spiculation, subtlety, texture, margin, sphericity, respectively. As mentioned before[**?**],nodule attributes can facilitate the solution of MSE. Therefore,we need these attributes prediction results to combine with the results of MSE classification and FPR classification to boost the performance of MSE and FPR tasks. At the same time, these attributes can potentially assist radiologists in evaluating the nodule malignancy result.

3.5 Fine Grain Lung Nodule Diagnosis

We conduct experiments via extracting 7000 'fake nodules' from LUNA2016 and 'real nodules' from LIDC-IDRI as the dataset(1:1 fake nodules to real nodules). As Table **??** shows, the fine grain ASMB3DCNN(ALL) obtains 95.7% in the accuracy of classifying Fake Nodules (FN), while the accuracy of ASMB3DCNN (ALL) in section **??** is 99.2%. This is because the FPR branch is not trained with the 'real nodule' in LIDC-IDRI, thus this branch has difficulty in classifying 'low risk nodule' and 'fake nodule', which has side effects on system. On the other hand, ASMB3DCNN (ALL) in section **??** obtains 99.1% in the accuracy of classifying Low Risk Nodules(LR) and 99.0% in the accuracy of classifying High Risk nodules (HR), while the fine grain ASMB3DCNN(ALL) improve 0.5% and 2.3% in accuracy of classifying LR and HR respectively over ASMB3DCNN (ALL) in MSE,which means the FPR can help to make a distinction between 'low risk nodules' and 'high rick nodules'. In general,the results confirm our idea about FPR and MSE are actually similar tasks and these two separated issues can be merged into a unified fine grain lung nodule classification problem .

3.6 Further Experiments

We conduct further experiments via using one 3d-cnn branch to learn attributes and the two tasks jointly(ASOBJ3DCNN), compared with the proposed multi-branch cnns that needs to pre-train each single cnn separately. As Table 7 shows,the gains of ASMB3DCNN(ALL) over the ASOBJ3DCNN are 32% and 27.4% respectively in accuracy of FPR and MSE. This result shows that training the task branches separately performs better than training branches jointly. The better performance of the proposed method mainly lies in the following: According to Table 1, the proportion of each nodule attribute in the dataset LIDC-IDRI is imbalanced,which makes it more adaptive to optimize each attribute classification training task separately when using multi-branch cnns, instead of using only one cnn to learn all the tasks jointly.

Table 6. Comparison of ASMB3DCNN(ALL) and ASOBJ3DCNN on Accuracy

Method	***Acc of FPR***	***Acc of MSE***
ASMB3DCNN(ALL)	0.974	0.961
ASOBJ3DCNN	0.654	0.687

4 Conclusions

In this paper, we propose ASMB3DCNN which can merge two tasks: False Positive Reduction (FPR) and Malignancy Suspiciousness Estimation (MSE) into a

unified task: fine grain lung nodule diagnosis. Label 'real nodule' defined in FPR can be split into two new finer grain labels, namely 'low risk' and 'high risk', which are defined in MSE. We analyse the sensitivity of attributes and choose attributes to improve the performance of the model. Moreover, we design a method of normalization to improve the performance of each branch and reduce the burden of computation. Experiments show that our model can achieve convincing results in FPR, MSE and the task of fine grain lung nodule diagnosis,what's more,our method provides nodule attributes prediction to assist radiologists in evaluating malignancy uncertainty.

We wanted to use CNN to capture high level features jointly like study [**?**], but we found that this method's performance was not good enough, it is because different features has different distribution. Attributes, like sphericity, with unbalance distribution is difficult to predict. If we use CNN to share high level semantic information, we cannot avoid the problem of unbalance distribution of attributes. So we decide to train branches separately at first, and then train branches jointly, so that we can fine-tune each branch to improve the whole model. Moreover, decision process without branch of malignancy is actually not ideal, we believe it is because the branch of malignancy actually learn something more than 8 attributes. We find that this classification system for nodules (1-6 levels and 8 attributes) is problematic. For extremely unbalanced attributes, we should adapt different evaluation criterion and we should review all 8 attributes and study the relationship between attributes and nodules. This work requires years of professional training and we will work with radiologists to design a better evaluation criterion.

Because of the strong learning ability of CNN and small number of dataset, all models for nodules classification have a risk of over-fitting. Moreover, few models use information of patients like smoking history and family history, which is different from diagnosis process in clinical. Our future work will also focus on fusing information of patients with CNN models to improve ability of generalization, and overcome the problems of predicting unbalance attributes.