MPDNet: Multimodal Pneumonia Detection Network Simulating Clinical Diagnosis Process

Qiuli Wang, Zhihuan Li, Dan Yang, Chen Liu*, Xiaohong Zhang*

***Abstract*—Pneumonia detection is one of the most crucial steps in the pneumonia diagnosing system. Most of existing approaches heavily dependent on chest X-Ray. Since chest X-Ray may have shadows caused by bones, muscle, and so on, conventional methods may lead to delayed treatments. Also, demographic information (e.g., age, gender) and clinical information (e.g., chief complaints) of subjects may also be related to lung status, and thus can help improve the performance. However, conventional methods seldom incorporate such demographic and clinical information into the learning models. In this paper, a Multimodal Pneumonia Detection Network (MPDNet) is described for clinical pneumonia detection. MPDNet simulates clinical pneumonia detection process using multi-channel CT images, demographic information, and clinical information. Specifically, we firstly extract visual features from three-channel (Lung Window, High Attenuation, Low Attenuation) images, which are transformed from one-channel grey level CT images. Different channels can provide supplementary features to each other and give qualitative information for pneumonia detection. Then we extract information about lesion location, symptoms, or how long patients have been ill from chief complaints, which enhances visual features extracted from CT images. We finally propose MPDNet, which incorporates CT visual features, complaint semantic features, and demographic information of subjects into the learning process. The proposed MPDNet has been extensively validated in 1002 clinical cases from The First Affiliated Hospital of Army Medical University. Our network achieves 0.945 in accuracy and has a very balanced performance in sensitivity and specificity. As far as we know, we are the first to detect pneumonic cases using large scale clinical data with demographic information. Our method demonstrates that demographic and clinical information provide more abundant information than image data only and gets very convincing results. While MPDNet is tailored for pneumonia detection, it can be extended and include more multimodal clinical data, and give out more reliable and explainable detection results.**

***Index Terms*—Multimodal Data, Pneumonia Detection, Computed Tomography (CT), Computer-Aided Detection and Diagnosis (CAD)**

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61772093), the Chongqing Major Theme Projects (Grant Nos. cstc2018jszx-cyztzxX0017, cstc2017zdcy-zdzx0077), and Fundamental Research Funds for the Central Universities (Grant Nos. CDJZR14105501, CDXYRJ0011). Asterisks indicate corresponding authors.

Q. Wang, Z. Li and D. Yang are with the School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China. E-mail: wangqiuli@cqu.edu.cn.

L. Chen is with the Radiology Department, The First Affiliated Hospital of Army Medical University, 400032, Chongqing, China. E-mail:cqliuchen@foxmail.com.

X. Zhang is with the Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing 400044, China, also with the School of Software Engineering, Chongqing University, Chongqing 401331, China, and also with the State Key laboratory of Coal Mine Disaster Dynamics and Control, Chongqing University, Chongqing 400044, China. E-mail: xhongz@cqu.edu.cn.

I. Introduction

PNEUMONIA is a prevalent thoracic disease in daily life. In clinical practice, multimodal data plays a key role in the decision-making process. Radiologists need to consider multimodal information to decide on the next treatment plan. As a result, radiologists from major hospitals have a heavy burden of work. Thus, developing a fast, robust, and accurate CAD system to perform automated detection of pneumonia is meaningful and vital.

Many research works have devoted efforts in pneumonia detection, monitoring and diagnosing like [1]–[3].

Hoo-Chang Shin [1] proposed a method which used CNN to extract features from chest X-Ray and used LSTM [4] to generated MeSH [5] terms for chest X-Ray. In 2017, Xiaosong Wang et al. [6] provided hospital-scale chest X-ray database ChestX-ray8, which contained eight common thoracic diseases. This database allowed researchers to use deeper neural networks to analyze thoracic diseases. They tested different pre-trained CNN models on this dataset. Experiments showed that ResNet50 achieved the highest AUROC score of 0.6333 in classifying pneumonia. They also provided ChestX-ray14, which contains more kinds of thoracic diseases. Based on this database, later in 2017, Yao et al. [7] achieved AUROC of 0.713 in classifying pneumonia using DenseNet Image Encoder. Pranav Rajpurkar, Andrew Y. Ng et al. [8] developed CheXnet with 121 convolutional layers and achieved AUROC 0.7680 in pneumonia classification. In 2018, Xiaosong Wang et al. [9] proposed TieNet, which could classify the chest X-Rays into different diseases and generate the report at the same time. In TieNet, CNN was used to capture features of chest X-Rays, RNN learned these features and generated reports based on attention mechanism, which could help the model to focus on different parts of chest X-rays alone with the generation of reports. In the pneumonia classification problem, they achieved 0.947 in AUROC based on reports, but they only reached 0.917 in AUROC on hand-labeled data.

Studies above have something in common. First of all, they are designed for chest X-Rays. Chest X-rays used to be the best available method for detect pneumonia, played a crucial role in clinical care and epidemiological studies [10], [11]. However, compared to chest X-rays, CT scans have a more unobstructed view of patients' bodies and allow visualization of 3D lung structures [12], since bones, skin, vessels, mediastinal and lung tissues may cause overlapping shadows in chest X-ray and cause misdiagnosis. CT can help to diagnose pneumonia in early-stage and avoid delayed treatments. Extensive studies show that 3D CNN is the best choice

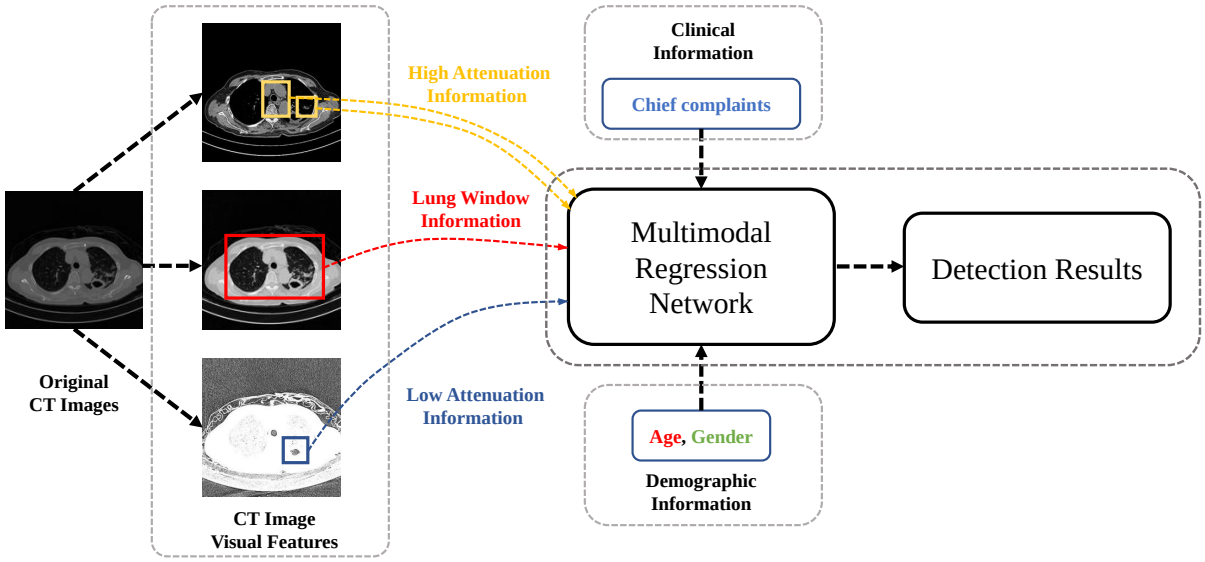


Fig. 1. Illustration of the proposed Multimodal Pneumonia Detection Network (MPDNet) for pneumonia detection. There are four main elements: (a) CT image processing; (b) visual features extraction from three channels; (c) clinical and demographic information processing; and (d) multimodal regression network for pneumonia detection

for keeping 3D spatial information in CT [13]. However, 3D CNN cannot be applied to raw CT data directly since it will bring a heavy burden to computers. Radiologists need to measure the lesions accurately, so we cannot reduce the size of images by resizing at will.

Second, these methods heavily rely on image information. Few of them combine image visual features with clinical information or demographic information. Models like TieNet do combine image visual features with descriptions about images written by radiologists. We believe using descriptions about images written by radiologists to improve models is not entirely convincing since descriptions like 'Findings' and 'Impressions' sometimes include diagnosis conclusions. Patients' chief complaints are valuable information when doctors are making decisions [14], since chief complaints are patients' direct feeling about their physical condition, telling us the patients' pain location, symptoms and how long have they been ill. Demographic information of subjects is strongly connected to the condition of lungs and subjects. Many studies have proved that demographic information can help improve the classification/regression performance in CAD systems [15]–[18]. However, as far as we know, few studies have used these information to improve CAD systems for pneumonia.

In general, there are two major drawbacks of existing CAD systems for pneumonia: (1) They cannot handle raw CT scans, which allows visualization of lung structures; (2) Few studies consider multimodal clinical information like demographic information (e.g., age, gender) and clinical information (e.g., chief complaints), which is a conflict to clinical practice.

To address such drawbacks, we propose a novel Multimodal Pneumonia Detection Network (MPDNet) and simulate clinical pneumonia detection. The architecture of MPDNet is shown in Fig I. We use raw data collected from The First Affiliated Hospital of Army Medical University. Each case contains not only CT image information but also clinical information and demographic information.

Herein, (i) each CT image will be transformed into a three-channel image with three windows: Lung Window(LW), High Attenuation(HA) and Low Attenuation(LA). LW provides visual features of normal lung tissues, HA provides visual features of abnormal increase in lung density, LA provides visual features of abnormal decrease in lung density. Three channels complement each other, which not only maintains the ability to extract information from normal lung tissues but also increases the ability to extract information from abnormal lung tissues. (ii) We also include clinical data in our MPDNet. Chief complaints can provide the location of pain, symptoms, and how long have patients been ill. This information is related to the CT image and enhances the visual features extracted from CT. (iii) Demographic information about age and gender can provide priori information since patients of different age and gender have differences in the morphology of the thoracic cavity and lungs. (iv) To reduce the burden of calculation, we treat CT slices as short video frames and a Recurrent Convolutional Neural Network (RCNN) is used to capture visual features from CT slices. RCNN uses a 2D CNN to capture visual features from each 2D slice, and LSTM captures relationships between slices. We use another LSTM to analyze semantics from chief complaints. Information about age and gender will be treated as two extra variables. Our model MPDNet, as shown in Fig I, will learn a joint regression of all features above and give out the final detecting results.

The remainder of the manuscript is organized as follows. Section II describes the prepare of dataset and pre-process steps. Section III describes the architecture of MPDNet and details of our model. Section IV reports our experimental results. We further discusses some key points of proposed model and some phenomenons shown during experiments in this section. Our conclusions are summarized in section VI.

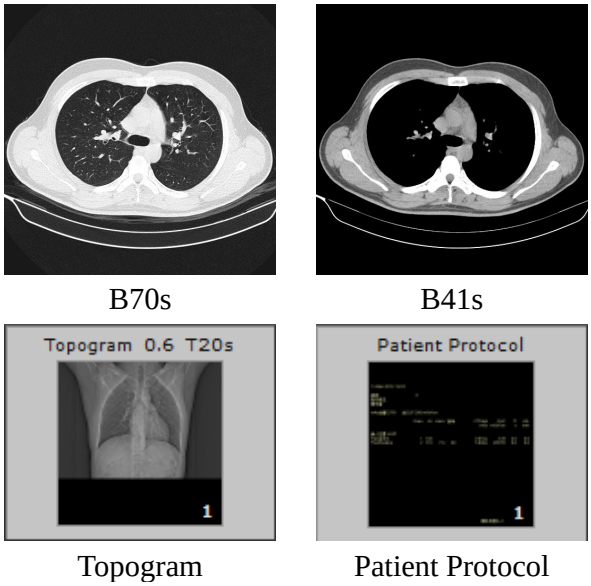


Fig. 2. Scans under Different 'Convolutional Kernel'. Slice under 'B70s' has clearer view of lungs, slice under 'B41s' has a more unobstructed view of the heart. 'Patient Protocol' and 'Topogram' contain some basic parameters of CT equipment or information about radiologists, which are not suitable for CNN.

II. Material and Image Processing

A. Data Generation

Because of the shortage of public available CT dataset for pneumonia, we use the raw data from the Radiology Department of The First Affiliated Hospital of Army Medical University. We get 1036 cases of CT (842 pneumonic cases, 464 healthy cases) from hospital PACS (Picture Archiving and Communication Systems) in the last three years (2016 - 2019). All cases are selected randomly. Raw data from the hospital may have more than one series of images, and each series has specific data types, image windows, or view angles. Generally speaking, radiologists and doctors will use the series under lung window with the smallest 'Slice Thickness', but for deep learning models, each case can only have one series. So we design a protocol to pick up specific series for us.

First of all, we eliminate these cases which start scanning from the middle of the chest. Then we pick up the best series from the whole cases according to the following requirements:

(a) We choose the series with the specific 'Convolution Kernel'. Different 'Convolution Kernel' may have different data types or different image windows, as shown in Fig 2. We need to notice that these names of 'Convolution Kernel' vary between hospitals and CT equipment, so if you want to adopt this protocol, you need to observe 'Convolution Kernel' in your environment. In our study, we choose 'B31f', 'I31f 3', 'B70f', 'B80f', 'B70s'. We notice that in the Radiology Department of The First Affiliated Hospital of Army Medical University, 'B70s' is the most common parameter used in clinical which contains 620 cases.

(b) We remove series like 'Patient Protocol', 'Topogram'. These series, as shown in Fig 2, contain some basic parameters and information about CT equipment.

(c) We calculate 'Slice Thickness' of each series, and keep the series with the smallest 'Slice Thickness', since small thickness may keep more detailed information about body structure.

(d) If there were more than one series meet the last two requirements, we would keep the series with the largest number of slices, which could have a larger span of view.

As a result, 552 pneumonic cases and 450 cases of healthy people (1002 cases total) are left. Since our data are collected from the Radiology Department, the proportion of pneumonia are higher than normal proportion.

The dataset is divided into training / validation / testing as 60% / 20% / 20% and make them identically distributed in three parts of datasets, so we have 602 cases in the training set, 200 cases in the validation set, 200 cases in the test set. Each CT scan has a case file. In case files, we can get patient basic information: patient ID, gender, age, and chief complaint.

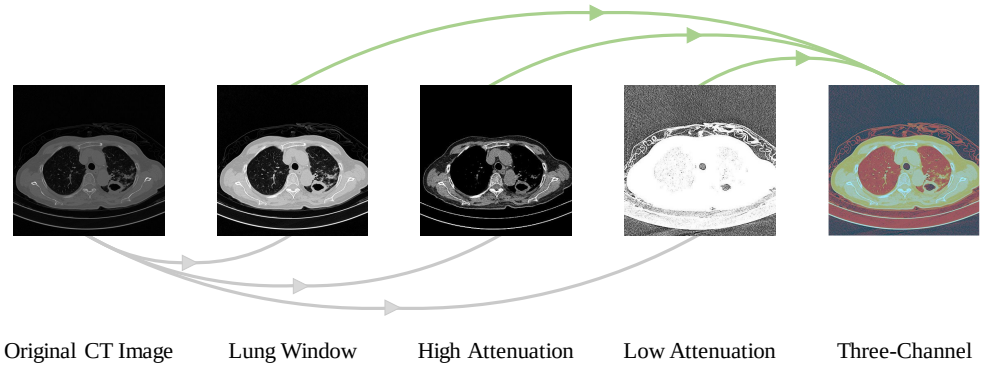
B. Data Pre-processing

1) Pre-processing of CT Image Data: There are different kinds of image windows for CT reader, such as windows for bone, brain, chest, or lung. Images under different image windows will highlight different tissues of bodies. As mentioned in sectionII-B1, each series of CT has one specific 'Convolution Kernel'. But it may make data inconsistent between different cases. So we transform raw data into HU(Hounsfield Unit) values. The Hounsfield Unit is a quantitative scale for describing radio-density. After transformed into HU value matrices, all slices form CT scans will have the same unit of measure.

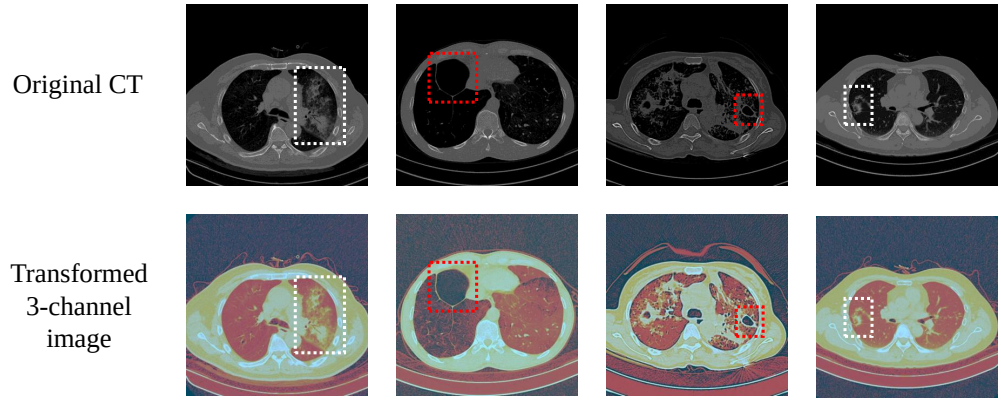
Following the study in [19], [20], HU value matrices will be transformed into images using three HU windows: Lung Window (LW) [-1000, 400HU], High Attenuation (HA) [-160, 240HU], Low Attenuation (LA) [-1400, -950HU]. For each slice, it will generate three one-channel grey level images. Then we compress three one-channel grey level images into one three-channel false-color RGB image. The 'Slice Thickness' between each slice is adjusted into 10mm, and each case will keep 32 slices.

As shown in Fig 3, three-channel images can show more density information about lung tissues. Original CT images are grey level images; high dense tissues are white; normal lung tissues and low dense tissues tend to be black. Relatively, three-channel false-color images have a larger scale of colors. First of all, high dense tissues will still tend to be white, like bones, high dense tissues in the lungs. Second, normal lung tissues will tend to be red, and low dense tissues tend to be black, which is very useful when patients have severe lung diseases. The influence of different HU value ranges will be discussed in section IV-B.

2) Pre-processing of Demographic and Clinical Data: Studies like [21], [22] treated the demographic information as confounding factors. That is, these methods often construct a regression model based on these factors by removing the confounding effects from measured features for subjects. The main disadvantage of such a strategy is that the original representations of subjects will be modified because this strategy



(a) Data Pre-precess for CT Scans



(b) Three-Channel Images

Fig. 3. Sub-figure (a) shows data pre-process for CT scans. Each scan will be transformed into three images with different windows (yellow arrow). Then three images will be compressed into one three-channel false-color image (green arrow). Sub-figure (b) exhibits examples of three-channel images. In this figure, void space (in red rectangle) in original CT images is not very obvious since other normal tissues are in black too. But in the three-channel images, we notice the difference between normal tissues and low dense tissues. Moreover, the details of high dense tissues (in the white rectangle) are still kept.

adds up several steps of engineered pre-processing in a directed and engineered way [18]. Following these studies, we also treat demographic information as confounding factors. The demographic and clinical information of all studied subjects is listed in Table I.

For patients' chief complaints, since all chief complaints are written in Chinese, we have to do Chinese word segmentation. Chinese word segmentation is a challenging problem, so we will take a short cut and use a mature tool: Jieba text segmentation [1] to segment Chinese sentences into Chinese word sequences.

After word-segmentation, we use word2vec [23], [24] to embed word sequences into vectors and use CBOW(Continuous Bag-of-Words) to capture relationship between words. Since our corpus is very small, the embedding size is set to 50, and the window size for CBOW is set to 3. We set length of Chinese word sequence to 16 since 16 is the maximum length among all chief complaint sequences. For those sequences whose length is less than 16, we add 'None' to fill up the voids and increase the length to 16. The details of word2vec will not be discussed here. After embedding, each word will be embedded into a vector of 50 dimensions.

[1]https://github.com/fxsjy/jieba

III. Multimodal Pneumonia Detection Network

A. Construction of RCNN

RCNN (Recurrent Convolutional Neural Network) has been proved to be very effective in video caption, description, and classification [25], [26], some studies have applied RCNN to medical image analysis. Zreik, Majd et al. [27] recently used RCNN for automatic detection and classification of coronary artery plaque, they used CNN extracts features out of $25 \times 25 \times 25$ voxels cubes and used an RNN to process the entire sequence using gated recurrent units (GRUs) [28]. KL Tseng et al. [29] exploited convolutional LSTM to model a sequence of 2D slices, and jointly learn the multi-modalities and convolutional LSTM in an end-to-end manner to segment 3D biomedical images.

As mentioned in section I, CT allows visualization of lung structures, which brings a large amount of redundant information, like muscle, vessels, and bones. It will cost lots of calculation resource if we use 3D CNN directly. However, if we treat CT slices as short video frames, we can analyze them using RCNN. In RCNN, each slice will be fed into CNN

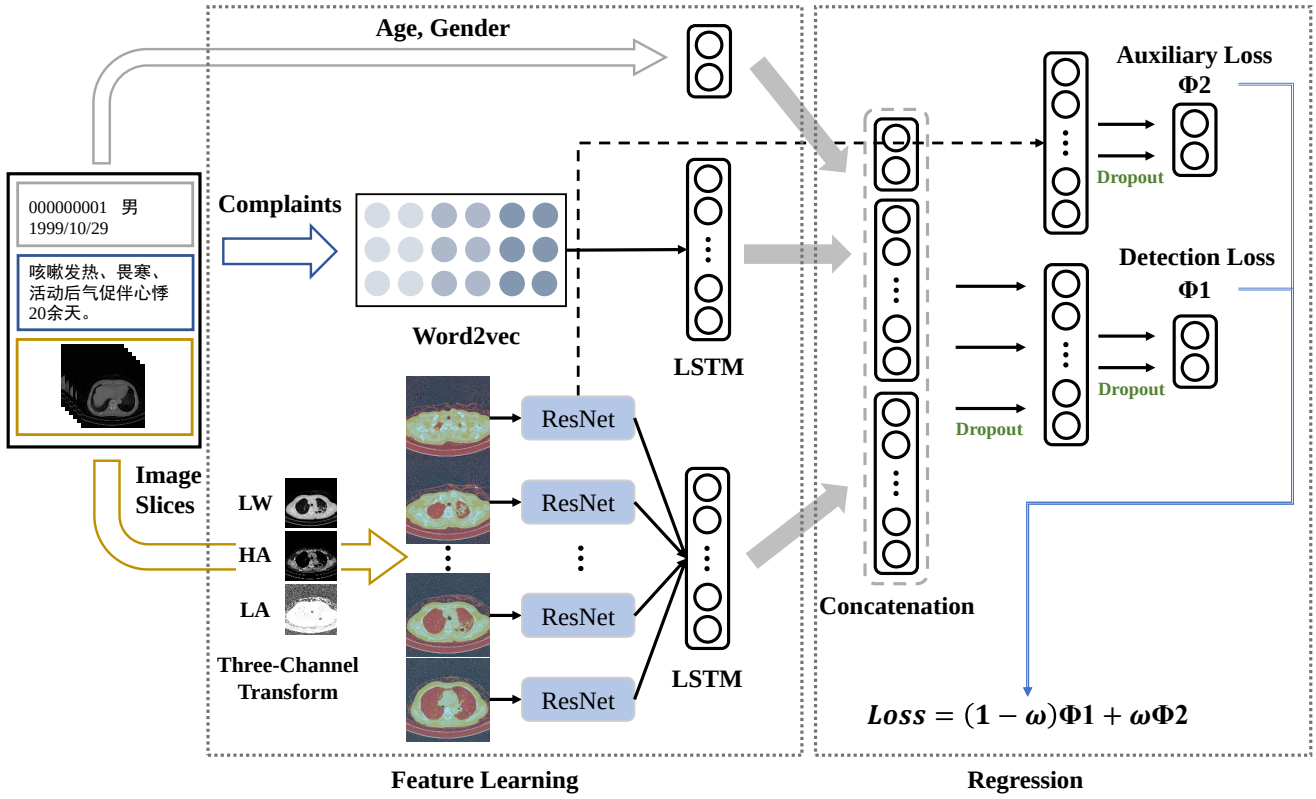


Fig. 4. Overview of the proposed MPDNet. The input data include three-channel CT slices, clinical and demographic information (i.e., chief complaints, age and gender) of each subject. The output is the regression result, which indicates whether this subject is pneumonic. Note that the term 'a' in '$a@b \times b$' denotes the number of kernels, while '$b \times b$' represents the size of a 2D convolutional kernel.

TABLE I

Number of Male and Female Patients in HC and PC

	Healthy	*Pneumonic*	*Total*	*Percentage**
Male	240	361	601	60.1%
Female	210	191	401	47.6%
Total	450	552	1002	55.1%

Number of HC and PC in Different Ages

	Healthy	*Pneumonic*	*Total*	*Percentage**
0-10	6	1	7	14.3%
10-20	31	2	33	6.1%
20-30	122	30	152	19.7%
30-40	124	45	169	26.6%
40-50	109	108	217	49.8%
50-60	53	131	184	71.2%
60-70	5	126	131	96.2%
70-80	0	82	82	100%
> 90	0	27	27	100%
Total	450	552	1002	55.1%

Percentage* is Percentage of Pneumonia Patients

in sequence and get a sequence of visual features. Then this sequence of features will be fed into RNN, so that we can reduce the need for calculation resource and keep 3D spatial information at the same time. Follow the study [25], we use LSTM with 256 units as our RNN cells cause LSTM has been demonstrated to be capable of large-scale learning of sequence data.

We use CNN without fully-connected layers as a feature extractor. The input size of CNN is 512×512, so the outputs of CNN will be extensive. We use the global average pooling [30] to reduce the number of neurons significantly. It is a replacement of fully-connected layers to enable the summing of spatial information of feature maps. After global average pooling, we insert a fully-connected layer to reduce dimensions to 256 to fit the number of LSTM units. To get the best RCNN for CT scans, after LSTM layer, we insert two additional fully-connected layers to give out classification results of RCNN, so that we can observe performances of different RCNNs and choose appropriate architecture. Experiments will be discussed later in section IV.

After building RCNN, we will keep architecture above LSTM (including LSTM) and insert into our MPDNet as encoder of visual features. It encodes image feature sequences and gives out the last output of LSTM as middle state hv_t:

$$hv_t = LSTM(Fx_t, hv_{t-1}, z_{t-1}) \tag{1}$$

Fx_t is the t-th visual features in CT slices, hv_{t-1} is LSTM hidden state of $t-1$ step, z_{t-1} is LSTM output of $t-1$ step. t is the length of slices, in this study, we set t as 32.

B. Construction of MPDNet

The whole RCNN, as mentioned in section III-A, can be seen as an encoder of CT images. Besides CT image

information, we also have clinical information about patients gender, age, and chief complaints.

Details of processing steps have been discussed in section II. The second LSTM is used to encode chief complaints. It is calculated in the same way as Eq. 1:

$$hc_{ct} = LSTM(Cx_{ct}, hc_{ct-1}, z_{ct-1}) \tag{2}$$

Cx_{ct} is word embedding matrix of the ct-th word in chief complaint, hc_{ct-1} is LSTM hidden state of $ct-1$ step. ct is the length of chief complaint, which is 16.

After getting hv_t, hc_{ct}, we can calculate the prediction and loss Φ_1 as follows:

$$\Phi_1 = \sum_i y_i log(\Delta_i),$$
$$\Delta = Softmax(F(hv_t \bigotimes hc_{ct} \bigotimes A \bigotimes G))$$

where y_i are vectors for labels, Δ is prediction after Softmax layer, $\bigotimes$ is the concatenation operation, A is patient age, G is patient gender. F is a function to fit the regression model of hv_t, hc_{ct}, A and G. In this study, we use two fully-connected layers to fit the regression function. Φ_1 is cross-entropy, which is commonly used as classification loss in many studies like [27].

Since each case has 32 slices but the loss is calculated for only one time in each case, we assume that the gradients propagate to CNN will be very weak, so that CNN will not be appropriately trained. Invoked by the study in [31], we use an auxiliary loss to enhance the signal of the gradient for CNN. The auxiliary loss Φ_2 and loss of whole model $Loss$ are defined as follow:

$$Loss = (1-\omega) \times \Phi_1 + \omega \times \Phi_2$$
$$\Phi_2 = \sum_i y_i log(\Delta_i^c)$$

where ω is a parameter within the interval (0, 1). Φ_2 is classification cross-entropy loss from CNN, Δ_i^c is Softmax prediction of CNN. ω can adjust the weight of two losses at different training phases. We expect that at the beginning of training, CNN gets stronger gradient and learn to capture features from CT images more quickly. After parameters of CNN get stable, Φ_1 tends to get small and keep updating parameters of LSTM. Experiments also show that RCNN with auxiliary loss can have a better performance, which will be discussed later in section IV. The parameters in MPDNet are optimized by minimizing the $Loss$.

Finally, MPDNet, which simulates clinical pneumonia detecting process, is built. RCNN for image data and LSTM for clinical information will be trained jointly. The architecture of MPDNet is shown in Fig III.

IV. Experiments

A. Experimental Setup

There are three parts of experiments in this section.

In the first part, we will try to analyse the effect of three-channel iamge and auxiliary loss. Meanwhile, we will test different kinds of RCNN networks to get the best combination between CNN and LSTM, the outputs from RCNN (1×256) will be feed into two fully connected layers to get classification results. Moreover, we use CNN models pre-trained on ImageNet [32]. Experiments demonstrate that using pre-trained models can significantly improve the converging speed.

In the second part, we will try to analyse the effect of demographic and clinical information. We will try to analyse the relationship between chief complaints and CT images by counting word frequency.

In the third part we will train MPDNet and show the experimental results on clinical data. We use RCNN as an encoder for CT scan visual features, use LSTM as an encoder for chief complaint features, and combine them with information of demographic information. All these features will be fed into two fully-connected layers and one Softmax layer to get final classification results. The initial learning rate is set to 0.0005 and drops 50% every 3000 training steps. The dropout rate in fully-connected layers is set to 0.5. MPDNet will be trained for four epoch, and each epoch contains 15 iterations for all training data.

In this work, all the experiments are run on the NVIDIA DGX Station with a GPU of NVIDIA Tesla V100. Source code for data pre-processing and MPDNet will be released very soon. We will also release the model with trained parameters and some sample cases for demo. But we cannot release dataset because of the privacy of patients.

B. Analysis of Three-Channel Image and Auxiliary Loss

In this section, we have conducted several experiments to analyze the effect of the three-channel image and auxiliary loss. We test three kinds of classic CNN models: VGG16 [33], ResNet [34] and GoogLeNet with Inception-V3 [31] with three-channel images, LW (Lung Window) images, HA (High Attenuation) images, and LA (Low Attenuation) images. Experiment results are shown in Table II.

We can see that RCNN(VGG) and RCNN(ResNet) trained by LW images perform better in sensitivity, but their specificity is lower than 0.9. ResNet50 trained by three-channel images performs the best in accuracy, specificity, AUROC.

These results demonstrate two phenomenons: (1) ResNet50 has a better performance on visual features learning than VGG16 and GoogLeNet with Inception-V3. This conclusion is similar to the conclusion drawn in [6], and their experiments showed that ResNet50 outperformed GoogLeNet and VGG16. (2) Three-channel images can provide more complete visual information of lungs. RCNN(ResNet), RCNN(VGG), and R-CNN(GoogLeNet) trained with three-channel images have the best performances compared to these trained with LW, HA, and LA images. As a result, our RCNN use ResNet50 as its CNN part, and use one layer of LSTM cells as its RNN part.

To experimentally analyse the effect of three-channel images, we output the feature maps of the convolutional layer, which are displayed in Fig 5. More specificity, we output the feature maps after one convolutional layer, one max-pooling layer, and three ResNet blocks, the size of feature maps are 128×128. To keep experiments environment consistent, all experiments carried on in this part are based on RCNN with ResNet50. Experiments show that CNN trained by three-

TABLE II
COMPARISON OF ALL KINDS OF RCNN AND MPDNET

Structure	*Data*	*Accuracy*	*Sensitivity*	*Specificity*	*AUROC*
RCNN(VGG)	Lung Window Image	0.805	**0.954**	0.626	0.790
RCNN(GoogLeNet)	Lung Window Image	0.865	0.826	0.912	0.869
RCNN(ResNet)	Lung Window Image	0.925	**0.954**	0.890	0.922
RCNN(GoogLeNet)	High Attenuation Image	0.880	0.853	0.912	0.883
RCNN(ResNet)	High Attenuation Image	0.875	0.908	0.835	0.872
RCNN(GoogLeNet)	Low Attenuation Image	0.860	0.890	0.824	0.857
RCNN(ResNet)	Low Attenuation Image	0.865	0.900	0.824	0.861
RCNN(VGG)	Three Channel Image	0.890	0.927	0.846	0.886
RCNN(GoogLeNet)	Three Channel Image	0.905	0.900	0.912	0.906
RCNN(ResNet)	Three Channel Image	**0.930**	0.927	**0.934**	**0.930**
RCNN(ResNet), One Loss	Three Channel Image	0.920	0.917	0.923	0.920

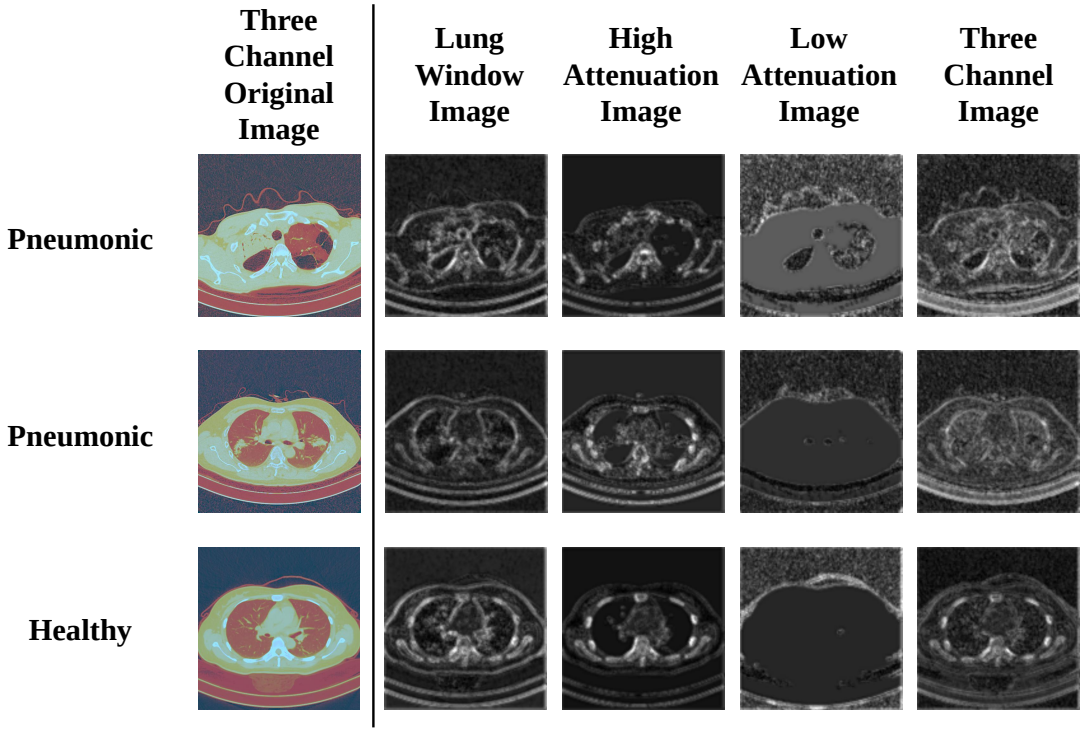
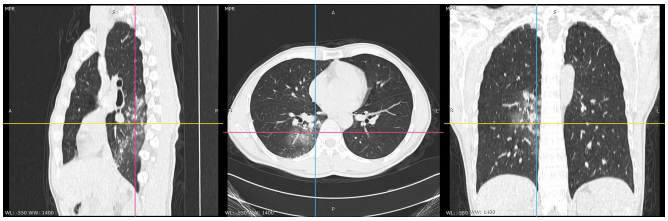


Fig. 5. Convolutional Feature Maps from CNN Models Trained by Different Images. In the first row, three-channel CNN can capture the low dense tissues of lungs, which are not very clear in LW (Lung Window) CNN and HA (High Attenuation) CNN. LA (Low Attenuation) CNN can notice the low dense tissues, but the details of heart and vessels are ignored in the low attenuation CNN. In the second row, three-channel CNN still can capture high dense tissues, which is the same as LW CNN and HA CNN, but LA CNN has difficulty in doing so. The last row shows a healthy case. The healthy case has a clear view in LW CNN, HA CNN, and three-channel CNN, but shows nothing in LA CNN.

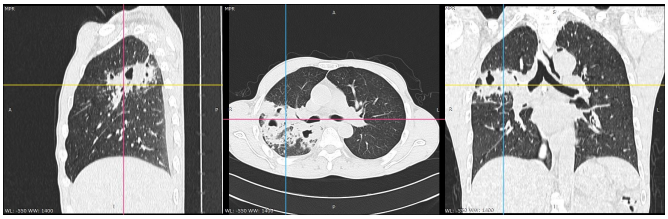
channel images has advantages over CNNs trained by other kinds of images.

In Fig 5, images in the first column are original false-color CT images, which are direct outputs from CT slices. The second, the third and the fourth columns are feature maps from LW CNN, HA CNN, and LA CNN. Images in the last column are feature maps from three-channel CNN. According to Fig 5, HA window can keep high dense information, but HA has difficulty in capturing the difference between low dense tissues and normal tissues. Contrarily, LA can keep low dense information will, but high dense information tends to be blank in LA. LW window is close to the three-channel window. However, the three-channel window has better discrimination for normal tissues and low dense tissues.

Moreover, we run an experiment to prove the effect of auxiliary loss. We train RCNN(ResNet) with three-channel images, but we set ω to 1, which means we remove the gradient propagates directly to CNN, this model has only one loss. As a result, the performance of RCNN with single loss drops around 1% in all four indications. We also output weights of two losses during training and observe that weight for LSTM loss $(1-\omega)$ is 0.6238 at the beginning of training (602 steps), however, $(1-\omega)$ will increase to 0.7234 when training process comes to 36120 steps, it means weight for



院外检查右下肺阴影，伴咯血，7天。
Has been examined by another hospital, shadow in right lower lung, hemoptysis, 7 days.



反复咳嗽、咳痰伴右胸痛半年，加重1月。
Repeated cough, sputum, pains in right chest for half a year, gets worse by one month.

Fig. 6. Chief complaints can provide information related to CT images. In this figure, we show two pneumonic cases, and each case has chief complaints provided by patients. Words marked red give the location, and words marked blue provide symptoms. English chief complaints are translated from Chinese above. Location and symptoms information provided by chief complaints are related to abnormal tissues in CT images.

CNN is 0.3762 at 602 steps, and it will drop to 0.2766 at the end. These two phenomenons prove that, by using auxiliary loss, CNN will be trained in a better way.

C. Analysis of Clinical and Demographic Information

In this section, we demonstrate the effect of clinical and demographic information. As mentioned in section I, clinical and demographic information can enhance the features extracted from the CT image or provide priori information.

To verify that cheif complaints can provide symptoms which are related to pneumonia, we count word frequency about symptoms. Table III shows that the top 10 keywords in HC (healthy cases) and PC (pneumonic cases) have certain regularity. 'Cough' is the most frequent keyword in both HC and PC. It appears 256 times (46.4%) in PC, 183 times (40.7%) in HC. However, symptoms like 'Expectoration', 'Fever', 'Coughing blood' appear more frequently in PC. For example, 'Coughing blood' appears 47 times in PC, but only appears one time in HC. According to this table, patient who have symptoms like expectoration, repeat condition, shortness of breath have larger chance of having pneumonia. Patients who have chest pain, or feel unconfortable have less chance of having pneumonia. Having cough, on the other hand, is a symptom with minimal discrimination.

We count the number of words which can provide information of location. In 1002 cases, every 2 cases contain 1 words which can help to locate the lesions (appear 504 times).

Fig 6 shows two examples. According to the location and symptom information provided by chief complaints, we can accurately locate lesions in CT. In these two cases, words marked red is information related to location, words marked blue are related to symptoms. In the first case, its chief complaint locates the symptoms in the right lower lung, and then we find shadows in the accurate place. In the second case, it chief complaint says that this patient has pains in right chest, then we also find shadows in the right lung in CT images. This phenomenon demonstrates that information from chief complaints is related to CT images, and can assist deep learning model.

D. Results on Clinical Data

In this section, the comprehensive evaluation of Multimodal Pneumonia Detection Network (MPDNet) is shown, results of experiments are shown in Table IV-D.

As mentioned in section III-B, the output of RCNN, features of clinical and demographic information will be concatenated together and fused by two fully-connected layers. It is simple yet effective. Experiments indicate that MPDNet trained by multimodal data has the highest score in accuracy, specificity, and AUROC score. But it achieves 0.936 in sensitivity, 0.009 lower than the most top 0.945. It means MPDNet has the best performance of binary classification according to its AUROC score. Besides, we remove the information about age and gender and found that MPDNet without age and gender has a higher sensitivity and lower specificity than MPDNet with information about age and gender.

If we treat RCNN(ResNet) trained with the three-channel image as our baseline, clinical chief complaints can increase sensitivity with 1.8% to 94.5%. It means chief complaints do have information which can help to detect pneumonia. Meanwhile, this information also decreases specificity to 90.1%, which is not hard to understand cause patients sometimes cannot accurately describe his feelings or even exaggerate his condition.

If we add demographic information (i.e., age and gender), the sensitivity drops a little bit, but the specificity increases to 95.6%, which means age and gender add information strongly connected to specificity.

The validation loss and accuracy during training is shown in Fig 7. According to this figure, demographic information can improve accuracy to 0.7 at the very beginning. It means some certain distribution must influence the dataset we are using.

According to the table I mentioned above, we observe some interesting phenomenons. (i) A male patient has a more significant chance of getting pneumonia. In 601 male cases, about 60% of them are pneumonic; however, in 401 female cases, only 47.6% are pneumonic. This phenomenon may be related to smoking since males in Chinese suffer a severe smoking problem; (2) The table shows that age is also associated with the chance of getting pneumonia. We can observe that people older than 40 have a much larger chance of getting pneumonia. There are about half of healthy cases between 40-50, but this indication drops so quickly that it goes down to 28.8% between 50-60.

These two tables explain why accuracy can achieve 0.7 at the very beginning of training and why information about age and gender can improve specificity to 95.6%. These findings explain why priori information provided by gender and age can have such a remarkable effect on our network.

TABLE III

Top 10 Frequent Key Words in Pneumonic Cases

Key Words	*Frequency in PC*	*Percentage*	*Frequency in HC*	*Percentage*
咳嗽, Cough	256	0.464	183	0.407
咳痰, Expectoration	103	0.187	42	0.093
反复, Repeat Condition	65	0.118	48	0.107
气促, Shortness of Breath	60	0.109	17	0.038
发热, Fever	51	0.092	14	0.031
咯血, Coughing Blood	47	0.085	1	0.002
加重, Aggravation	46	0.081	13	0.029
痰, Sputum	32	0.058	19	0.042
乏力, Weak	29	0.053	7	0.016
感染, Infection	28	0.051	1	0.002

Top 10 Frequent Key Words in Healthy Cases

Key Words	*Frequency in HC*	*Percentage*	*Frequency in PC*	*Percentage*
咳嗽, Cough,	183	0.407	256	0.464
胸痛, Chest Pain	67	0.149	17	0.031
不适, Unconfortable	54	0.120	25	0.045
疼痛, Pain	53	0.118	25	0.045
反复, Repeat Condition	48	0.107	65	0.118
咳痰, Expectoration	42	0.093	103	0.187
背痛, Backache	28	0.062	8	0.014
痰, Sputum	19	0.042	32	0.058
胸闷, Chest Tightness	19	0.042	16	0.029
气促, Shortness of Breath	17	0.038	60	0.109

Percentage is frequency divided by number of cases. PC is Pneumonic Cases. HC is Healthy Cases

TABLE IV
COMPARISON OF ALL KINDS OF RCNN AND MPDNET

Structure	*Data*	*Accuracy*	*Sensitivity*	*Specificity*	*AUROC*
RCNN(ResNet)	Three Channel Image	0.930	0.927	0.934	0.930
MPDNet	Three Channel Image & Complaints	0.925	**0.945**	0.901	0.923
MPDNet	Multimodal Data	**0.945**	0.936	**0.956**	**0.945**

V. DISCUSSION

Even if MPDNet can detect pneumonia using multimodal data, there are still some shortcomings in our work.

Firstly, we analyze 1002 cases in this study. But 1002 cases are far small than 'big data', so our model's performance is restricted by data distribution and quality.

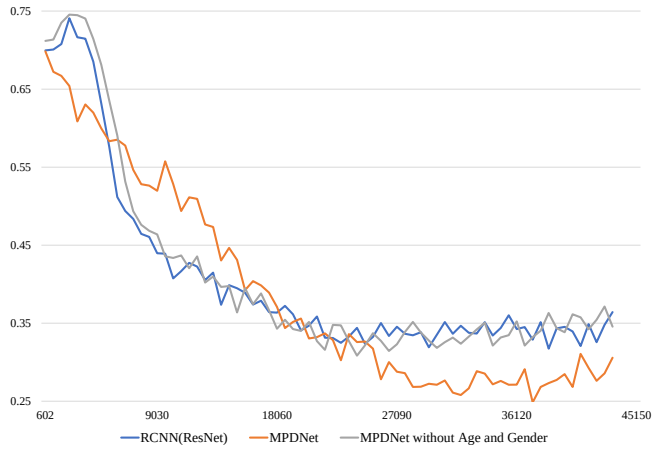
Secondly, we only consider chest CT scans, chief complaints, gender, and age. In clinical practice, besides the tests mentioned above, patients usually need to take blood pressure measurements, blood tests, heartbeat measurements, and other tests. These examinations can help doctors gain a more objective and comprehensive understanding of the patient condition so that doctors can make a more accurate diagnose.

However, it is very difficult to overcome these two shortcomings mentioned above since data collected from PACS are disorder. Constructing a big scale medical dataset with consistent data is a very challenging task, cause raw data is affected by radiologists' habits, data acquisition equipment, and hospital work rules. Our future work will focus on finding a method which can perform accurate diagnose on disorder data and include multimodal information from more medical tests.

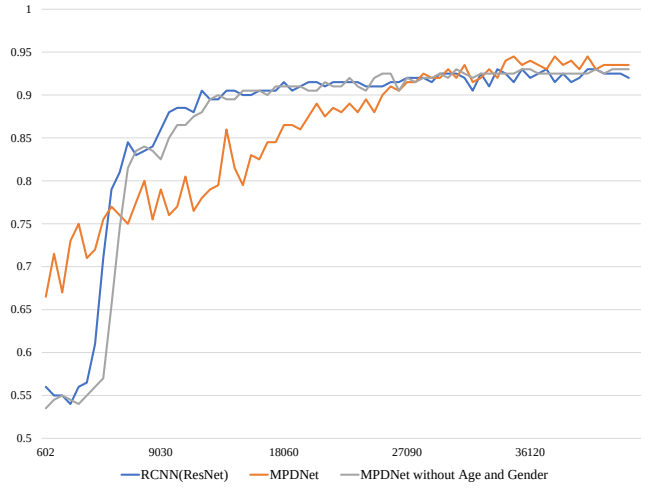
VI. CONCLUSION

In this study, we propose a novel model MPDNet (Multimodal Pneumonia Detection Network), which combines CT visual features with patients' age, gender, and chief complaints to simulates clinical practice. MPDNet extracts visual features from three-channel images, semantic features from chief complaints, and fuses these features with priori information provided by age and gender.

We analyze 1002 cases (450 healthy cases and 552 pneumonic cases) from the Radiology Department of The First Affiliated Hospital of Army Medical University. Experiments demonstrate that MPDNet achieves promising performance.



(a) Validation Loss During Training



(b) Validation Accuracy During Training

Fig. 7. Sub-figure (a) shows the validation loss during training; Sub-figure (b) shows the validation accuracy during training. We can see that MPDNet's performance outperforms others.

REFERENCES

[1] H. C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *Computer Vision & Pattern Recognition*, 2016.

[2] N. Deepika, P. Vinupritha, and D. Kathirvelu, "Classification of lobar pneumonia by two different classifiers in lung ct images," in *2018 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2018, pp. 0552–0556.

[3] D. K. Iakovidis, S. Tsevas, M. A. Savelonas, and G. Papamichalis, "Image analysis framework for infection monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 1135–1144, 2012.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] "Mesh: Medical subject headings," https://www.nlm.nih.gov/mesh/meshhome.html.

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *computer vision and pattern recognition*, pp. 3462–3471, 2017.

[7] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.

[8] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[9] X. Wang, Y. Peng, L. Le, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *IEEE CVPR 2018*, 2018.

[10] F. T, "Imaging of pneumonia: trends and algorithms," *European Respiratory Journal*, vol. 18, no. 1, pp. 196–208, 2001.

[11] T. Cherian, E. K. Mulholland, J. B. Carlin, H. Ostensen, R. Amin, M. De Campo, D. Greenberg, R. Lagos, M. G. Lucero, S. A. Madhi *et al.*, "Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies," *Bulletin of The World Health Organization*, vol. 83, no. 5, pp. 353–359, 2005.

[12] P. D. Korfiatis, A. N. Karahaliou, A. D. Kazantzi, C. Kalogeropoulou, and L. I. Costaridou, "Texture-based identification and characterization of interstitial pneumonia patterns in lung multidetector ct," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 675–680, 2009.

[13] T. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Translation Journal on Magnetics in Japan*, vol. 2, no. 8, pp. 740–741, 1987.

[14] J. Wu, X. Liu, X. Zhang, Z. He, and P. Lv, "Master clinical medical knowledge at certificated-doctor-level with deep learning model," *Nature communications*, vol. 9, no. 1, p. 4352, 2018.

[15] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson, "The clinical use of structural mri in alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 2, p. 67, 2010.

[16] P. Coupé, S. F. Eskildsen, J. V. Manjón, V. S. Fonov, D. L. Collins, A. disease Neuroimaging Initiative *et al.*, "Simultaneous segmentation and grading of anatomical structures for patient's classification: application to alzheimer's disease," *NeuroImage*, vol. 59, no. 4, pp. 3736–3747, 2012.

[17] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, A. D. N. Initiative *et al.*, "Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects," *Neuroimage*, vol. 104, pp. 398–412, 2015.

[18] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1195–1206, 2018.

[19] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Three aspects on using convolutional neural networks for computer-aided detection in medical imaging," in *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 2017, pp. 113–136.

[20] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers *et al.*, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 1, pp. 1–6, 2018.

[21] J. Dukart, M. L. Schroeter, K. Mueller, A. D. N. Initiative *et al.*, "Age correction in dementia–matching to a healthy brain," *PloS one*, vol. 6, no. 7, p. e22193, 2011.

[22] M. de Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," 2016.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

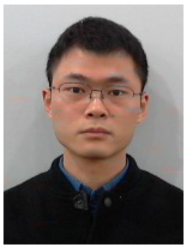
[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[25] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[26] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," *arXiv preprint arXiv:1902.10322*, 2019.

[27] M. Zreik, R. W. V. Hamersvelt, J. M. Wolterink, T. Leiner, and I. Isgum, "A recurrent cnn for automatic detection and classification of coronary artery plaque and stenosis in coronary ct angiography," *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, pp. 1–1, 2018.

[28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
[29] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3d biomedical segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6393–6400.
[30] M. Lin, Q. Chen, and S. Yan, "Network in network," *international conference on learning representations*, 2014.
[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *computer vision and pattern recognition*, pp. 2818–2826, 2016.
[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *international conference on learning representations*, 2015.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *computer vision and pattern recognition*, pp. 770–778, 2016.



Qiuli Wang received the B.E. degree in the School of Information Engineering, Yangzhou University in 2016. He is currently working toward the Ph.D. degree in the School of Big Data & Software Engineering, Chongqing University.

His research interests include medical image computing, deep learning, so on.

Zhihuan Li received the B.E degree in Geological Engineering from China University of Mining and Technology, Xvzhou, China in 2016. He is currently working toward the M.S. degree in Software Engineering from the Department of Big Data & Software Engineering, Chongqing University, Chongqing. His research interests include medical image analysis , segmentation and so on.

Chen Liu received the M.D. degree in Medical Imaging from Army Medical University, China, in 2015. He is currently an attending physicians in the Radiology Department of Southwest Hospital which is the first affiliated hospital of Army Medical University. He has hosted more than 6 research including National Natural Science Foundation and got funded more than 1.6 million. He published 6 articles as first author. His current research interests include brain functional MRI, clinical data mining, medical imaging deep learning.

Dan Yang received the B.Eng. degree in automation, the M.S. degree in applied mathematics, and the Ph.D. degree in machinery manufacturing and automation from Chongqing University, Chongqing. From 1997 to 1999, he held a post-doctoral position with the University of Electro-Communications, Tokyo, Japan. He is currently the President of Southwest Jiaotong University. He is also a Professor with the School of Big Data & Software Engineering, Chongqing University. He has authored over 100 scientific papers and some of them are published in some authoritative journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, and BMVC. His research interests include computer vision, image processing, pattern recognition, software engineering, and scientific computing.

Xiaohong Zhang received the M.S. degree in applied mathematics from Chongqing University, China, where he also received the Ph.D. degree in computer software and theory, in 2006. He is currently a Professor and the Vice Dean with the School of Big Data & Software Engineering, Chongqing University. His current research interests include data mining of software engineering, topic modeling, image semantic analysis, and video analysis.