MDDNet: Multimodal Data Diagnosis Network for Clinical Pneumonia Classification

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

Abstract*—*Objective: **Pneumonia detection is one of the most crucial steps in pneumonia diagnosing system. Clinical information of patients plays an important role in detection of pneumonia. In is paper, a Multimodal Data Diagnosing Network(MDDNet) is described for clinical pneumonia detection.** ***Method:*** **MDDNet is based on deep learning neural network and analyzes multimodal data. We use Recurrent CNN, which can keep 3-D spatial information and reduce the need of calculation resource, to capture visual features from CT image data. Each slice of CT is transformed into one 3-channel(Lung Window, High Attenuation, Low Attenuation) image which can provide more information of lung density. Meanwhile, patient clinical information like complaint, age and gender is adopted and provides more abundant information to improve the accuracy of pneumonia detection. A Long Short Term Memory(LSTM) network is used to analyze semantic features of patient complaints and provides information which image data cannot provide, like how many days the patient has been ill. Information about age and gender can provide priori information since age and gender is associated with certain kinds of pneumonia. CT visual features, complaint semantic features, patient age and gender will be fused together and calculate joint distribution to predict whether these cases are pneumonic.** ***Results:*** **We analyze 1002 clinical cases from The First Affiliated Hospital of Army Medical University. Our model 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 multimodal data.** ***Conclusion:*** **Our method proves that multimodal data provides more abundant information than image data only and improves the accuracy of pneumonia detection.** ***Significance:*** **Our model can be extended and include more kinds of clinical data to give out more reliable and explainable detection results.**

***Index Terms*—Long Short Term Memory(LSTM), Pneumonia Diagnosis, Recurrent Convolutional Neural Network, Computed tomography (CT), Computer-aided detection and diagnosis (CAD), Multimodal Data**

I. Introduction

PNEUMONIA is a very common thoracic disease in our life. In clinical practice, a radiologist needs to consider different source of information to decide on the next treatment plan, as a result, multimodal data plays the key role in decision making process. According to the survey, a radiologist of a major hospital need to diagnosis hundreds of pneumonia cases every day. Thus, developing a fast, robust and accurate CAD system to perform automated diagnosis of pneumonia is meaningful and important.

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised August 26, 2015.

There have been several methods and epidemiological studies [1][2][3] for pneumonia detection and diagnosing, most of them use image data like chest X-ray as their information source. Hoo-Chang Shin [4] combined CNN and LSTM [5], proposed a model which could describe the contexts of a detected diseases based on the deep CNN features. This model used CNN to extract features from chest X-Ray and used LSTM to generated MeSH [6] terms for chest X-Ray. In 2017, Xiaosong Wang et al. [7] provided hospital-scale chest X-ray database ChestX-ray8 which contained eight common thoracic diseases. This database allowed researchers use deeper neural network to analyze thoracic diseases. They tested different pre-trained CNN models on this dataset. Experiments showed that ResNet50 achieved highest AUROC score 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. [8] achieved 0.713 in AUROC score using DenseNet Image Encoder. Pranav Rajpurkar, Andrew Y. Ng et al. [9] developed CheXnet with 121 convolutional layers and achieved AUROC 0.7680 in pneumonia classification. In 2018, Xiaosong Wang et al. [10] 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 capture features of chest X-Rays, RNN learned these features and generated report based on attention mechanism, which could help model to focus on different parts of chest X-Ray alone with the generation of reports. In pneumonia classification problem, they achieved 0.947 in AUROC based on report, but they only achieved 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[1] and epidemiological studies[2]. However, compared to chest X-rays, CT scans have a clearer view of patients' bodies, since bones, skin, vessels, mediastinal and lung tissues may cause overlapping shadows in chest X-ray and cause misdiagnosis. CT allows visualization of lung structures[11], which can help to diagnosis pneumonia in early stage and avoid delayed treatments. Each slice of CT scans is a 2-D image of human body scan, besides 2-D visual features from CT, you can also reconstruct 3-D structure of human bodies using these slices. Extensive studies show that 3-D CNN is the best choice for keeping 3-D spatial information in CT[12]. However, 3-D CNN cannot be applied to raw CT data directly since it will bring a heavy burden to the server. Radiologists need to accurately measure the lesions, so we cannot reduce the size of images by resizing at will.

Second, these models are not designed following the radiologists' diagnosing process but are designed for the convenience of computer vision studies and deep learning model training. For models like CheXnet, image information is the key of models. Few of them combine image visual features with clinical information. Models like TieNet do combine image visual features with descriptions about images written by radiologists. But these descriptions only provide information related to images, which means no extra clinical information is provided to models. We also believe using descriptions about images written by radiologists to improve models is not quite convincing, since descriptions like 'Findings' and 'impressions' sometimes include diagnosis conclusions. Patients' complaints is a very useful information when doctors are diagnosing, since complaints is patients' direct feeling about their physical condition, telling us the patients' pain location, symptoms and how long have they been ill. Moreover, information of age and gender is also related to some certain diseases[13] [14]. However, as far as we know, few studies use this 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 clinical information like patients' complaints, which is conflict to clinical practice.

To address such drawbacks, we propose a novel Multimodal Data Diagnosis Network(MDDNet) for Clinical Pneumonia Classification. 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 about patient gender, age and complaints. In MDDNet, A Recurrent Convolutional Neural Network(RCNN) is used to capture visual features from CT slices. CT allows visualization of lung structures, but lots of redundant information will bring a heavy burden of calculation if we use 3-D CNN. RCNN uses a 2-D CNN to capture visual features from each 2-D slice, and LSTM captures relationships between slices, which means RCNN can analyze 3-D spatial information while reduce the need of calculation. Each CT image will be transformed into a 3-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. Besides image data, we also include clinical data in our MDDNet. Complaints can provide location of pain, symptoms and how long have patients been ill. These information is related to CT image and enhance the visual features extracted from CT. Moreover, information about age and gender can provide priori information since patients of different age and gender must have differences in the morphology of thoracic cavity and lungs. We use another LSTM(Long Short Term Memory) to analyze semantics from complaints, information of age and gender will be treated as two extra variables. Our model MDDNet(Multimodal Data Diagnosis Network), as shown in Fig 1, will learn a joint distribution of all features above and gives out the final classification result.

The remainder of the manuscript is organized as follows. Section II describes the architecture of MDDNet and details of our model. Section III describes the pre-processing steps of dataset and reports our experimental results. Section IV further discusses some key points of proposed model and some phenomenons shown during experiments. Our conclusions are drawn in section V.

II. Methodology

A. Construction of Recurrent Convolutional Neural Network

RCNN(Recurrent Convolutional Neural Network) has been proved to be very useful for video caption, description and classification [15][16], however, only a few work apply RCNN to medical image analyze. Zreik, Majd et al. [17] recently use RCNN for automatic detection and classification of coronary artery plaque, they use CNN extracts features out of $25 \times 25 \times 25$ voxels cubes, and use an RNN to process the entire sequence using gated recurrent units (GRUs)[18]. They proved that RCNN's potential for sequence information processing of medical 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 3-D 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 in sequence and get a sequence of visual features. Then this sequence of features will be fed into LSTM, so that we can reduce the need of calculation resource and keep 3-D spatial information at the same time. Follow the study[15], we use LSTM as our RNN cells cause LSTM has been demonstrated to be capable of large-scale learning of sequence data. We test three kinds of classic CNN models: VGG[19], ResNet[20] and GoogLeNet with Inception-V3 [21]. Experiments will be discussed in section III and section IV.

We use CNN without fully-connected layers as feature extractor. The input size of CNN is 512×512, so the outputs of CNN will be very large. We use global average pooling[22], as shown in Fig 2, to greatly reduce the number of neurons. 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 because the number of LSTM units is set to 256[15].

In order to get the best RCNN for CT scans, we run experiments to test best combination between CNN models and LSTM. Architecture of RCNN in this part is shown in Fig 3. After LSTM layer, we insert two fully-connected layers to give out classification results of RCNN, so that we can observe performances of different RCNNs and choose appropriate architecture. After building RCNN, we will keep architecture above LSTM(including LSTM) and insert into our MDDNet as encoder of visual features.

The experiments show that ResNet50 performs the best in these three models, so our RCNN use ResNet50 as its CNN

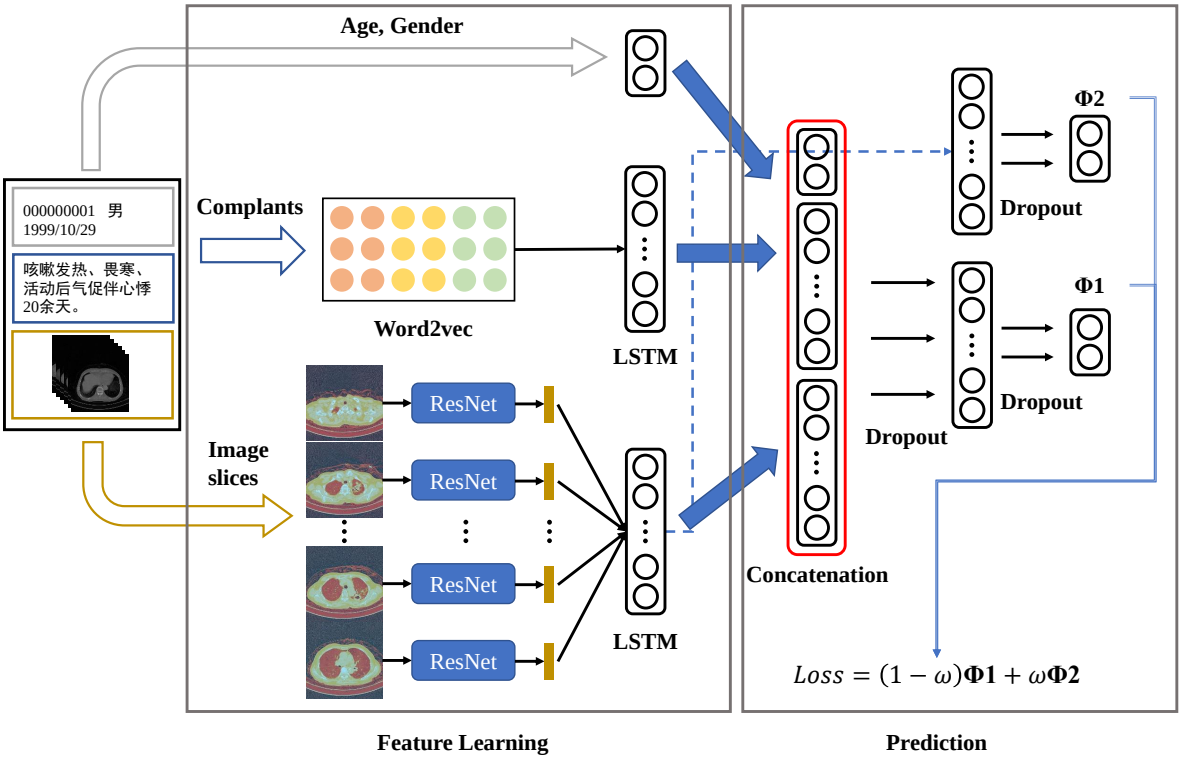

Fig. 1. Architecture of MDDNet. The black rectangle contains raw information from hospital. Information in grey rectangle is about age and gender, information in blue rectangle is complaints, information in yellow is CT image data. Complaints will be transformed into matrices by Word2vec and analyze by one LSTM network. Image will be fed into RCNN. Age and gender will be treated as two additional features. There three kinds of information will be concatenated in red rectangle and fed into fully-connected layers to get prediction

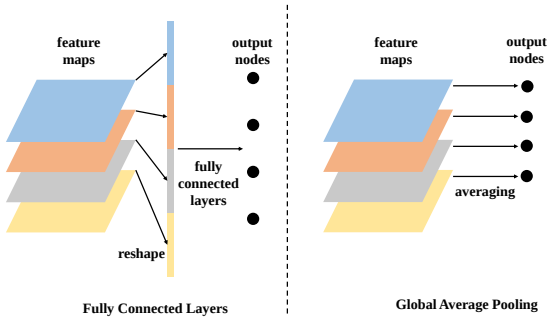

Fig. 2. Difference between Fully Connected Layers and Global Average Pooling

part, and use one layer of LSTM cells as its RNN part. This conclusion is similar to [7], their experiments showed that ResNet50 outperformed GoogLeNet and VGG16.

B. Multimodal Data Fusion and Diagnosis

The whole RCNN, as mentioned in section II-A, can be seen as a encoder of CT images, it encodes image feature sequences and gives out the last output of LSTM as middle state hv_t:

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

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.

Besides CT image information, we also have clinical information about patients gender, age, and complaints. For gender and age, we use them as additional features and set a tensor with size 1×2 to hold both values. For patients' complaints, we will use Jieba Chinese word segmentation tool [1] to segment Chinese sentences into word sequences. We set length of Chinese word sequence to 16. Then we transform sequences of words into sequences of vectors using word2vec, which is commonly used in nature language process, since it can capture the relations between words. The width of vectors is set to 50, so does the number of LSTM units. Details of processing steps will be discussed later in section III. This LSTM is the second encoder to encode complaint. It is calculated in the same way as Eq. 1:

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

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

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

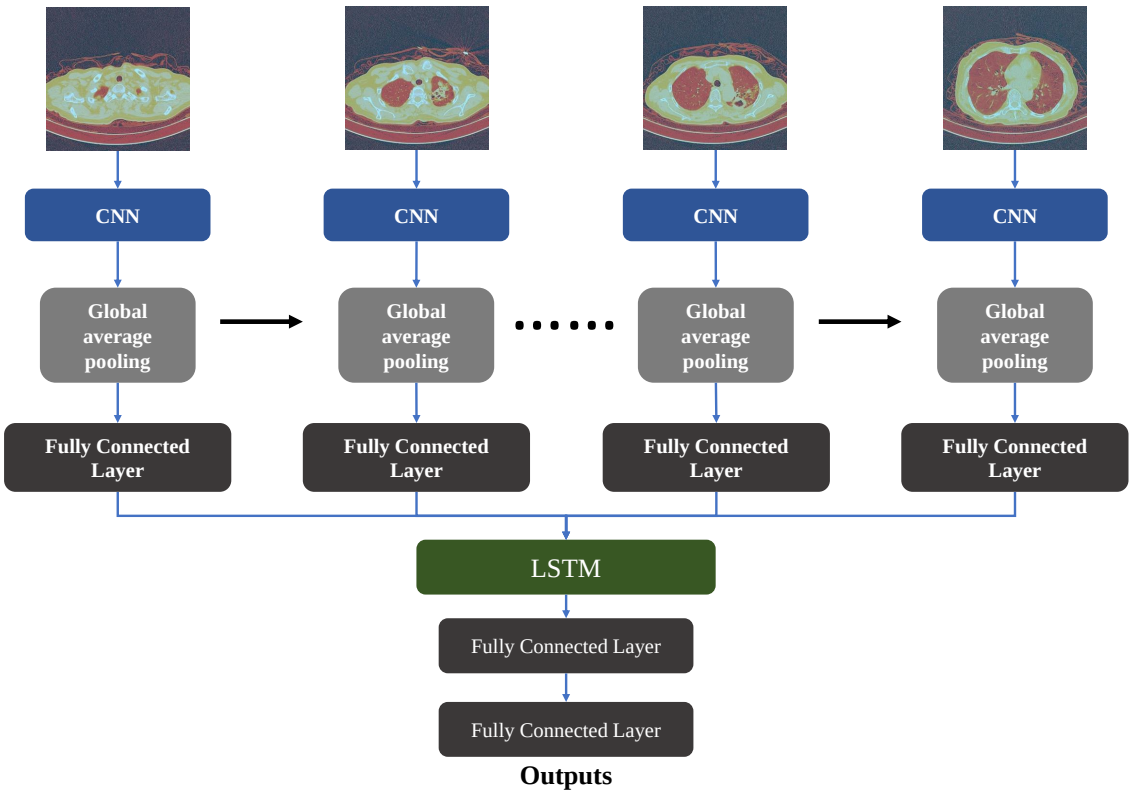

Fig. 3. Architecture of RCNN. Inputs of RCNN is a sequence which contains CT image slice. Each slice is sorted and analyzed by CNN in order. After CNN, there is a LSTM network. The outputs from CNN will be reshaped by fully-connected layers and fed into LSTM. In order to get best RCNN for CT, we add two fully-connected layers after LSTM so that we can get classification results.

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 the labels of patients, Δ is prediction after Softmax, $\bigotimes$ is the concatenation operation, A is patient age, G is patient gender. F is a function to calculate joint distributions of hv_t, hc_{ct}, A and G. In this study, we use two fully-connected layers to fit the function. Φ_1 is cross-entropy that can be used as classification loss [17].

Since LSTM need to encode 32 visual features, we assume that the gradients propagate to CNN will be very small, so that CNN will not be trained properly. Invoked by study in [21], we use a auxiliary loss to enhance signal of 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 get 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. We output weights of two losses during training MDDNet, as shown in Table I, 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 meas weight for CNN is 0.3762 at 602 steps, and it will drop to 0.2766 at the end. Experiments also show that RCNN with auxiliary loss can have a better performance, which will be discussed latter in section III.

TABLE I
WEIGHTS OF TWO LOSSES AT DIFFERENT TRAINING STEP

Number of Steps	$1-\omega$	ω
602	0.6238	0.3762
9030	0.6547	0.3453
18060	0.7027	0.2973
27090	0.7185	0.2815
36120	0.7234	0.2766

Finally, MDDNet is built, RCNN for image data and LSTM for clinical information will be trained jointly, the architecture of MDDNet is shown in Fig 1.

C. Training Process

There two steps during training process. The first step is to train difference kinds of RCNN to get the best combination between CNN models and LSTM, the outputs from

RCNN(1×256) will be feed into two fully connected layers to get classification results, as shown in Fig 3. We compare three kinds of classic CNN models: VGG16, GoogLeNet with Inception-V3, ResNet50. Experiments show that ResNet50 has the best performance. We tried to use deeper network like ResNet101, however using ResNet101 makes the training of model very slow and bring a heavy burden to servers. So we decide to use ResNet50 in RCNN. Moreover, we use CNN models pre-trained on ImageNet [23]. Models without pre-training is almost impossible to train because it won't converge or converge very slow during training. We test difference RCNNs, and experiments show that using pre-trained models can significantly improve the converging speed, as shown in Table II.

The second step is to train MDDNet model. We use RCNN get in the first step as encoder for CT scan visual features, use LSTM as feature encoder for complaints, and combine them with information of age and gender. All these features will be feed into two fully-connected layers and one Softmax layer to get final classification results. 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.

MDDNet will be trained for 4 epoch, and each epoch contains 15 iteration for all training data.

TABLE II
COMPARISON BETWEEN TRAINING FROM SCRATCH AND TRAINING WITH PRE-TRAINED WEIGHTS

Structure	*Pre-trained*	*Data*	*Accuracy*
RCNN(ResNet)	No	Lung Window Image	0.545
RCNN(GoogLeNet)	No	Lung Window Image	0.545
RCNN(ResNet)	Yes	Lung Window Image	0.925
RCNN(GoogLeNet)	Yes	Lung Window Image	0.865

III. EXPERIMENTS

A. CT Image Data and Multimodal Data Generation

Because of the shortage of public CT dataset for pneumonia, we use raw data from the Radiology Department of The First Affiliated Hospital of Army Medical University. We get 1036 cases of CT(842 cases with pneumonia, 464 healthy cases) from hospital PACS(Picture Archiving and Communication Systems). Raw data from hospital may have more than one series of images(yellow rectangle), each series may have different data type, different image windows, or series with different angles. Generally speaking, radiologists and doctors will use series under lung window with smallest 'Slice Thickness', but for deep learning models, we need to pick up the most suitable series manually. This work is very heavy, so we design a protocol to pick up 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:

1. We use the series with the specific 'Convolution Kernel'. Different 'Convolution Kernel' may have different data types or differnet image windows, as shown in Fig 4. We need to notice that these names of 'Convolution Kernel' vary between

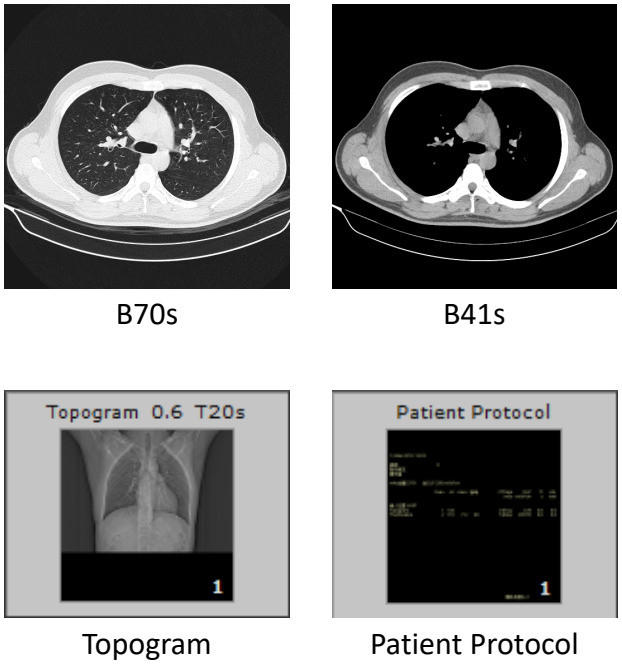

Fig. 4. Scans under Different 'Convolutional Kernel'. Slice under 'B70s' has clearer view of lungs, slice under 'B41s' has clearer view of heart. 'Patient Protocol' and 'Topogram' contain some basic parameters of CT equipments or information about radiologists, which are not suitable for CNN.

hospitals and CT equipments, 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'. Number of different 'Convolution Kernel' is shown in Fig 5. However, different 'Convolution Kernel' will not affect images we analyze in the end, because all slices will be calculated and transformed into HU value matrices, which will be discussed later in this section.

2. We remove series which are not showing cross section of human body like 'Patient Protocol', 'Topogram'. 'Patient Protocol' and 'Topogram', as shown in Fig 4, contain some basic parameters of CT equipments or information about radiologists, which should be eliminated.

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

4. If there were more than one series meet the last two requirements, we will keep the series with the largest number of slices, which can have a larger span of view.

As a result, we keep 552 cases with pneumonia and 450 cases of healthy people (1002 cases total). We split dataset in training/validation/testing as 60% / 20% / 20% and make them identically distributed in three parts of datasets, so we have 602 cases in training set, 200 cases in validation set, 200 cases in test set. Number of healthy and pneumonic cases in different slice-thickness is shown in Table III.

Each CT scan has a case file. In case files, we can get patient basic information: patient ID, gender, age and complaint. In this study, we will use gender, age as additional features, and use LSTM to extract textual features from patients' complaints, and combine them with features from CT image

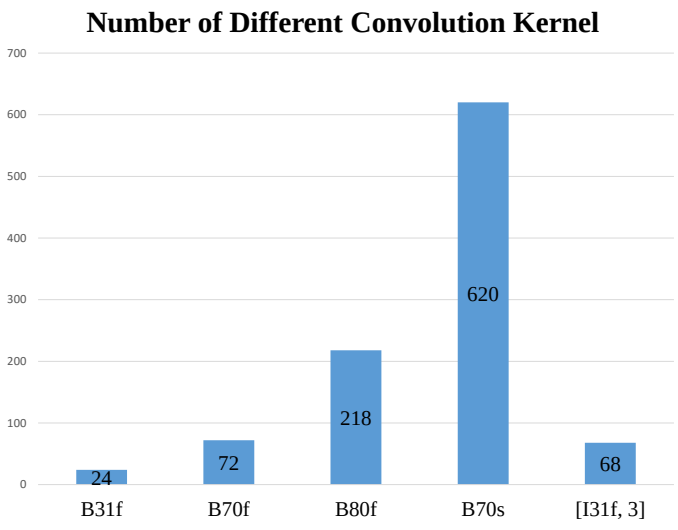

Fig. 5. Number of Different 'Convolution Kernel'. We notice that in the Radiology Department of The First Affiliated Hospital of Army Medical University, 'B70s' is the most common parameter used in clinical. However, this parameter varys between hospitals and clinics.

slices.

TABLE III
NUMBER OF HEALTHY AND PNEUMONIC CASES IN DIFFERENT SLICE-THICKNESS

Slice-Thickness	*Healthy*	*Pneumonic*
1 mm	0	24
1.5 mm	1	7
2 mm	444	386
3 mm	0	127
5 mm	5	8
Total	450	552

1) Pre-processing of CT Image Data: There are kinds of image windows for CT reader, such as windows for bone, brain, chest, lungs. Images under different image windows will highlight different tissues of bodies. As mentioned in sectionIII-A1, we can see that each series of CT actually has one specific 'Convolution Kernel' and show specific window for CT images directly from raw data. But it may make data inconsistent between different cases. So we transform raw data into HU(Hounsfield Unit) values. The Hounsfield Unit named after Sir Godfrey Hounsfield, is a quantitative scale for describing radio-density, its value is also termed CT number. After transformed into HU value matrices, all slices form CT scans will have the same unit of measure, then we will transform scans according to specific rules.

Following the study in [24] [25], we transform slices into images using three HU range: lung window [-1000, 400HU], high attenuation [-160, 240HU], low attenuation [-1400, -950HU]. For each slice, it will generate three 1-channel grey level images(lung window, high attenuation, low attenuation). Then we compress three 1-channel grey level images into one three-channel false color RGB image which fits the requirements of CNN models. The 'Slice Thickness' between each slice is adjusted into 10mm, and each case will keep 32 slices.

As shown in Fig 6, we can clearly see that three-channel images can show more density information about lung tissues. Original CT images are actually grey level images, which can only show lungs with white, black or grey. In original CT images, high dense tissues are white, normal lung tissues and low dense tissues are tend to be black. If you want to see details of low dense tissues, you have to adjust the image window to low attenuation range, but meanwhile you will lost the details of high dense tissues. In three-channel images, details of high dense and low dense tissues will both be kept. Three-channel fake 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 lungs. Second, normal lung tissues will tend to be red, 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 III-B.

2) Pre-processing of Patient Age, Gender and Complaints: The pre-process steps of age, gender and complaints is shown in Fig 7. For patient age and gender, we transform them into a two-dimensional array. For example, patient in 7 is an adult male, who was born in 1999-10-29. His gender and age will be transformed to $[1, 20]$. A female patient born in 1993 will have $[0, 26]$ to represent her information. 1 represents male patient, 0 represents female patient.

For patients' complaints, since we only have Chinese complaints, we have to do Chinese word segmentation. Chinese word segmentation is a very difficult problem so we will take a short cut and use a mature tools: Jieba text segmentation to segment Chinese sentences into Chinese word sequences. An example of Chinese word segmentation is shown in green rectangle in Fig 7. If you use data from English speaking countries, you may skip this step. After word segmentation, we use word2vec [26] [27] to embed word sequences into vectors and use CBOW(Continuous Bag-of-Words) to keep relationship between words. Since our corpus is very small, we set embedding size as 50, and window size for CBOW as 3. In order to simplify model, we set length of Chinese word sequence to 16 since 16 is the maximum length among all complaint sequences. For those sequences whose length is less than 16, we add 'None' to fill up the voids and increas 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.

B. Effectiveness of Three-Channel Image

In order to verify the effectiveness of three-channel pre-processing, we train four RCNNs with three-channel images, lung window images, high attenuation images and low attenuation images and output the feature maps of convolutional layer. Sample feature mapes are shown in Fig 8. 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. In order to keep experiments environment consistent, all experiments carried on in this part are based

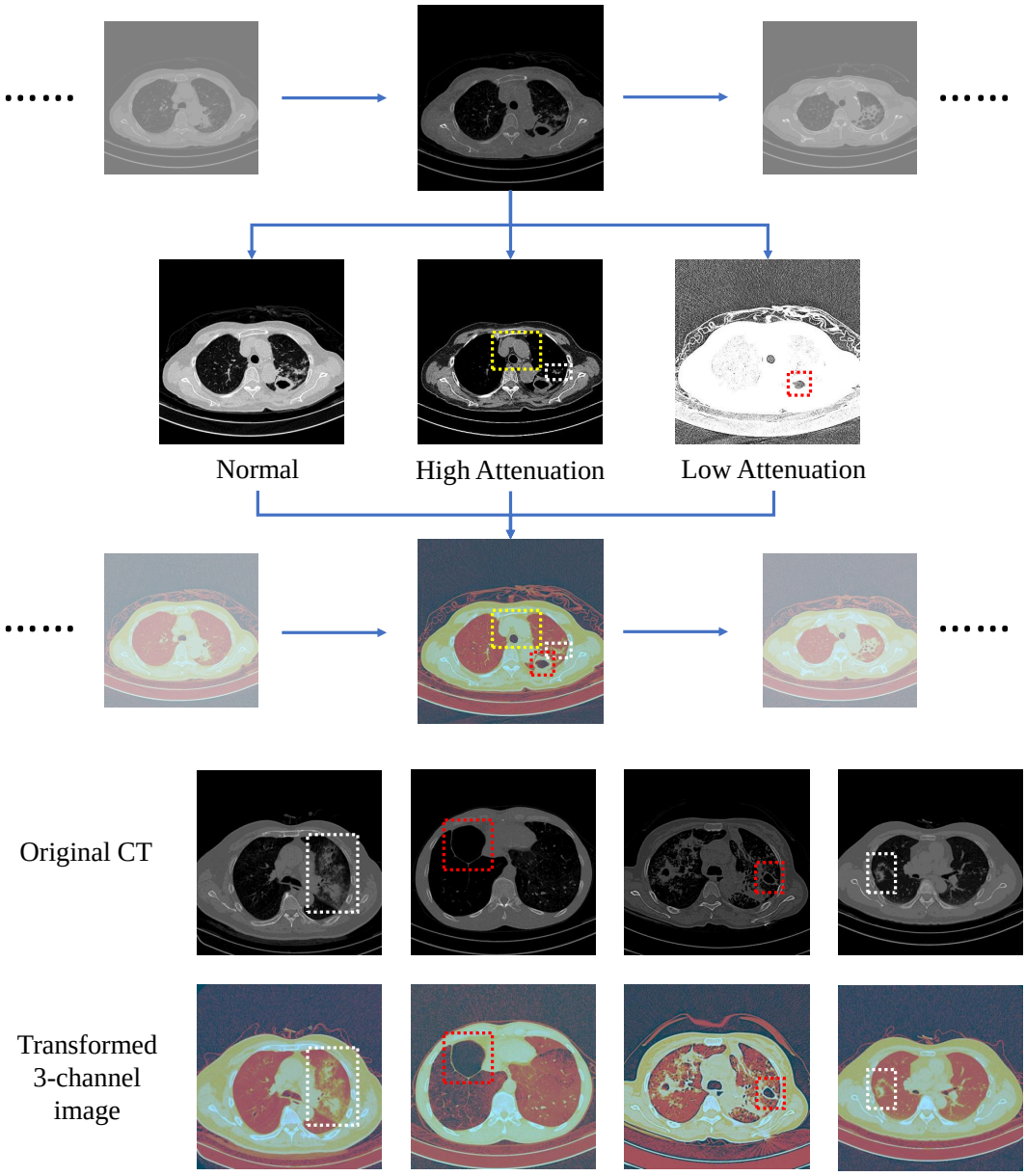

Fig. 6. Data Pre-precess for CT Scans. Void space(in red rectangle) in original CT images is not very obvious since other normal tissues is in black color too. But in three-channel image, we can clearly notice the difference between normal tissues and low dense tissues. Moreover, the details of high dense tissues(in white rectangle) are still kept.

on RCNN with ResNet50. We can see that CNN trained by three-channel images has advantages over CNNs trained by other kinds of images as shown in Fig 8.

In Fig 8, images in the first column are original fake color CT images, which are direct outputs from CT slices. The second, the third and the forth columes are feature maps from lung window CNN, high attenuation CNN and low attenuation CNN. Images in the last colume are feature mapes from three-channel CNN.

C. Effectiveness of Complaints, Age and Gender

As mentioned in I, information about age, gender and complaints can enhance the features extracted from CT image or become a supplement. We count word frequency in complaints. The frequency is shown in Table IV and Table V.

D. Results

In order to prove the effect of three-channel images, auxiliary loss and Multimodal Data, we run a lot of experiments to compare with each other, and the results of experiments is shown in Table VI.

First of all, we run experiments to prove the effect of images with three ranges of HU values and choose the best architecture of RCNN. We can see that RCNN(ResNet), RCNN(GoogLeNet) and RCNN(VGG) trained by three-channel image all have better performance than these models trained

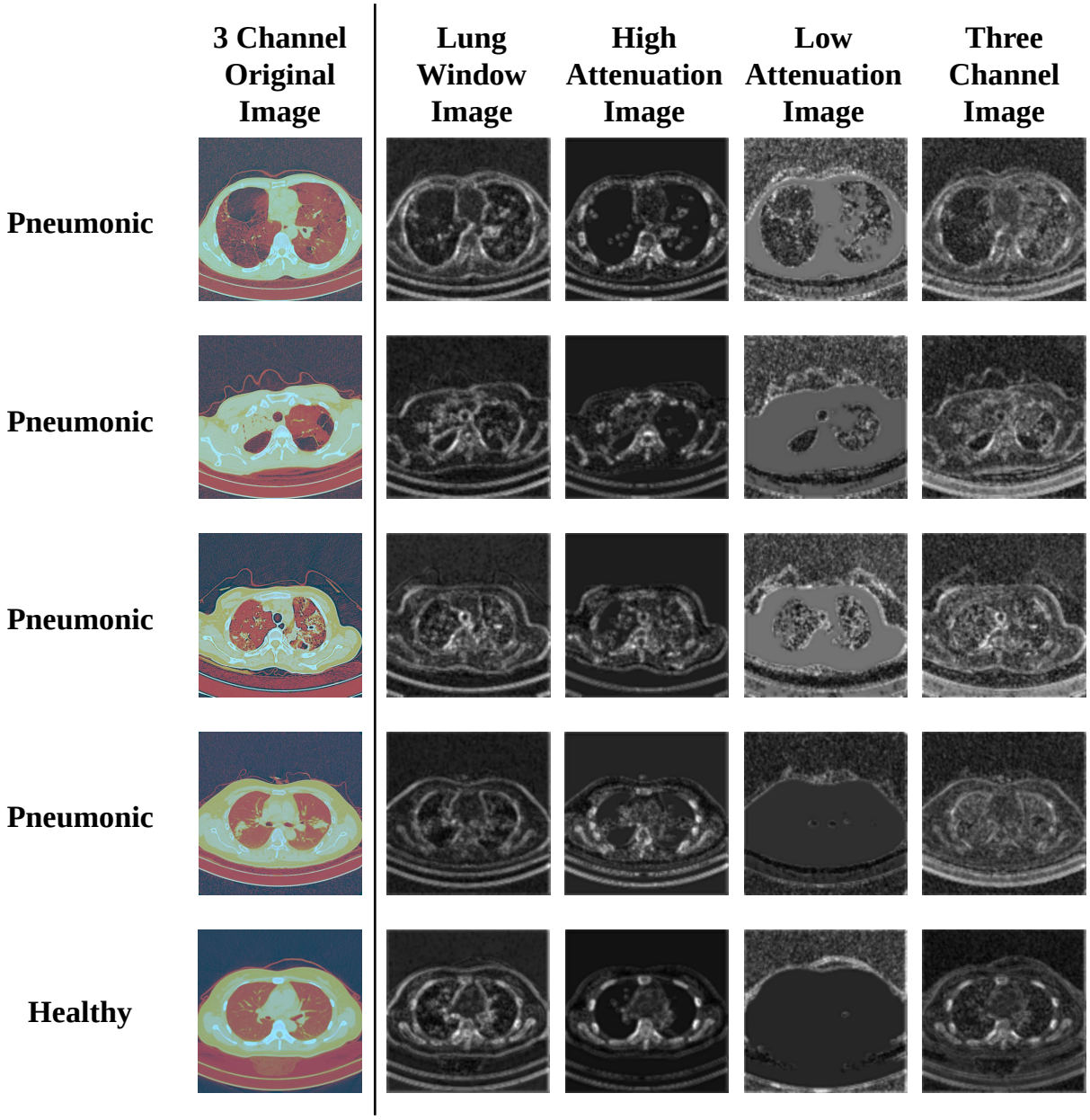

Fig. 8. Convolutional Feature Maps from CNN Models Trained by Different Images. In the first and the second rows, three-channel CNN can capture the low dense tissues of lungs, which are not very clear in lung window CNN and high attenuation CNN. Low attenuation CNN can notice the low dense tissues, but apparently the details of heart and vessels are ignored in the loa attenuation CNN. In the third and the forth row, three-channel CNN still has ability to capture high dense tissues, which is the same as normal CNN and high attenuation CNN, but low attenuation CNN has difficulty in doing so. In the third row, low attenuation CNN cannot distinguish the normal and unnormal tissues. Moreover, in the forth row, low attenuation CNN ignores high dense tissues. The last row shows a healthy case. Healthy case has a clear view in lung window CNN, high attenuation CNN and three-channel CNN, but shows nothing in low attenuation CNN.

by one-channel data. For RCNN(VGG), model trained by three-channel image outperforms RCNN(VGG) trained by lung window image in accuracy, specificity and AUROC score. RCNN(VGG) trained by lung window image has better performance in sensitivity, but we can see that it only get 0.626 in specificity, which means this model has not been trained well.

For RCNN(ResNet) and RCNN(GoogLeNet), we can see that these two models trained by three-channel image perform best compared to those trained by lung window image, high attenuation image and low attenuation image. Especially RCNN(ResNet) trained by three-channel image, it get 0.930 in accuracy, 0.934 in specificity, which are highest in different RCNN. RCNN(ResNet) trained by lung window image has 0.954 in sensitivity, which is the highest in experiments, but it only achieves 0.890 in specificity and corrupts the performance of whole model. As a result, we use RCNN(ResNet) as our visual feature encoder for CT.

Then we run a experiment to prove the effectiveness of auxiliary loss. We train RCNN(ResNet) with three channel

TABLE IV
TOP 10 FREQUENT KEY WORDS IN PNEUMONIA 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

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

TABLE V
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

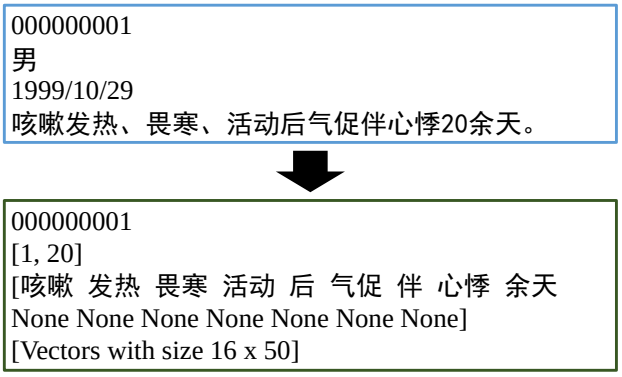

Fig. 7. Data Pre-precess for Age, Gender and Complaints

image, but we set ω to 1, which means we remove the gradient propagates directly to CNN, this model actually has only one loss. We can see that the performance of RCNN with single loss drops around 1% in all four indications. It proves that, by using auxiliary loss, CNN will be trained in a better way.

At last, we run experiments to prove that Multimodal Data can enhance the performance of CAD system. As shown in section 1, the output of RCNN(1×256), features of complaints(1×50), gender(1×1) and age(1×1) will be concatenated together(1×308) and fused by two fully-connected layers. It is simple, but effective. We can see that MDDNet has the highest score in accuracy, specificity and AUROC score. But it achieves 0.936 in sensitivity, 1.8% lower than the highest 0.954. It means MDDNet has the best performance of binary classification according to its AUROC score. We also remove the information about age and gender, we found that MDDNet without age and gender has a higher sensitivity and lower specificity than MDDNet with information about age and gender.

IV. DISCUSSION

If we treat RCNN(ResNet) trained with three-channel image as our baseline, we can see that complaint information can increase sensitivity with 1.8% to 94.5%, it means complaints do have information which can help diagnosis more accruatly. Meanwhile this information also decrease 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 information about 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 9. We can see that in Fig 9, MDDNet can achieve higher accuracy at the first phase of training, but RCNN(ResNet), RCNN(GoogLeNet) and RCNN(ResNet) with single loss perform better after. MDDNet's performance outperforms other methods again after 27000 training step. For RCNN(GoogLeNet), it converge slower and has a lower accuracy in the end. According to Fig 9, we can see that information about age and gender

TABLE VI
COMPARISON OF ALL KINDS OF RCNN AND MDDNET

Structure	*Data*	*Accuracy*	*Sensitivity*	*Specificity*	*AUROC*	*AUROC Rank*
RCNN(VGG)	Lung Window Image	0.805	**0.954**	0.626	0.790	13
RCNN(GoogLeNet)	Lung Window Image	0.865	0.826	0.912	0.869	10
RCNN(ResNet)	Lung Window Image	0.925	**0.954**	0.890	0.922	4
RCNN(GoogLeNet)	High Attenuation Image	0.880	0.853	0.912	0.883	8
RCNN(ResNet)	High Attenuation Image	0.875	0.908	0.835	0.872	9
RCNN(GoogLeNet)	Low Attenuation Image	0.860	0.890	0.824	0.857	12
RCNN(ResNet)	Low Attenuation Image	0.865	0.900	0.824	0.861	11
RCNN(VGG)	Three Channel Image	0.890	0.927	0.846	0.886	7
RCNN(GoogLeNet)	Three Channel Image	0.905	0.900	0.912	0.906	6
RCNN(ResNet)	Three Channel Image	0.930	0.927	0.934	0.930	2
RCNN(ResNet), One Loss	Three Channel Image	0.920	0.917	0.923	0.920	5
MDDNet	Three Channel Image & Complaints	0.925	0.945	0.901	0.923	3
MDDNet	Multimodal Data	**0.945**	0.936	**0.956**	**0.945**	1

can improve accuracy to 0.7 at the very beginning, it means the dataset we are using must be influenced by some certain distribution. So we count the number of male patients and female patients in healthy cases and pneumonic cases(Table VII) and number of patients in different ages(Table VIII).

In Table VII, we can see that a male patient has a larger chance of being pneumonic. In 601 male cases, about 60% of them are pneumonic, however, in 401 female cases, only 47.6% are pneumonic. This may be related to smoking since male in Chinese suffer a serious smoking problem. In Table VIII, we can see that age is also related to the chance of being pneumonic. We can still observe that people older than 40 have much larger chance of being pneumonic. 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. It is not hard to understand this phenomenon, since young people are very sensitive about their healthy condition, they will go to have physical examination as long as they feel uncomfortable, even if they have a lower chance of having pneumonia. However, old people have to face up with another condition. Most old people only go to hospital or clinic when their conditions are very bad. These two tables explain why accuracy can achieve 0.7 at very beginning of training and why information about age and gender can imporve specificity to 95.6%.

TABLE VII
NUMBER OF MALE AND FEMALE PATIENTS IN HEALTHY AND PNEUMONIC CASES

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

Percentage* is Percentage of Pneumonia Patients

In Fig 9, we can see that RCNN(GoogLeNet) has the highest loss in the end, so it performs the worst in accuracy. RCNN(ResNet) and MDDNet without age and gender has similar performance. RCNN(ResNet) with single loss drops quickly at first, but its loss is very close to RCNN(ResNet) in the end. MDDNet has the lowest loss at the beginning of training, even if it has the highest loss for a moment during training, it has the lowest loss after 27090 training steps. This phenomenon is not hard to understand, since our dataset is influenced by distributions shown in Table VII and Table VIII, MDDNet need more time to fit joint distribution between age, gender, visual features and textual features.

TABLE VIII
NUMBER OF HEALTHY AND PNEUMONIC CASES 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

V. CONCLUSION

In this study, we propose a novel model, MDDNet(Multimodal Data Diagnosis Network), which combines CT visual features with patients' age, gender and complaints. In MDDNet, CT scans will be treated like video frames, and analyzed by RCNN(Recurrent Convolutional Neural Network), complaints will be transformed into word vectors by word2vec and analyzed by LSTM. Features from CT images and complaints will be fused together with patients' age and gender. All these features will be used to classify cases into healthy cases or pneumonic cases.

We analyze 1002 cases(450 healthy cases and 552 pneumonic cases). In fact, 1002 cases is far small than 'big data', so our model's performance is restricted by data distribution and quality. However, in clinical practice, it is very difficult to construct a big scale medical dataset for deep learning, cause raw data is affected by radiologists' personal habits, data acquisition equipments, and hospital work rules. Our future work will focus on methods of data pre-processing which can over come difficulties mentioned above. Moreover, our future work will also focus on fusing more source of

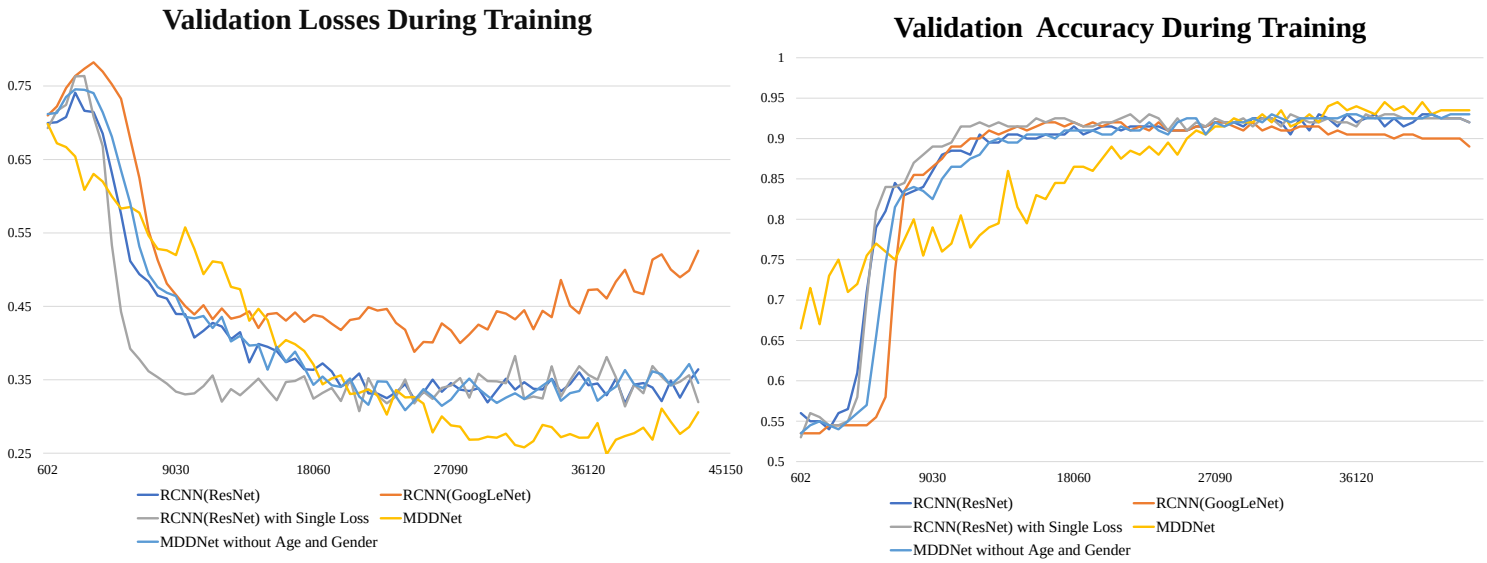

Fig. 9. Validation Loss During Training

information, like medical history, family history, blood test and other information which will be considered during clinical practice. All works above will be carried out under the premise of respecting the privacy of the patients.

Code for data pre-processing and MDDNet will be released very soon. We will release model with trained parameters and some sample cases for demo. But we cannot release dataset because of the privacy of patients.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] M. Xiaojian and W. Heyong, "Analysis of pathogens of pneumonia in children based on aprior algorithm," in *2011 IEEE International Symposium on IT in Medicine and Education*, vol. 1. IEEE, 2011, pp. 460–463.

[14] J. S. Huang, Y. F. Chen, and J. C. Hsu, "Design of a clinical decision support model for predicting pneumonia readmission," in *2014 International Symposium on Computer, Consumer and Control*. IEEE, 2014, pp. 1179–1182.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *international conference on learning representations*, 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *computer vision and pattern recognition*, pp. 770–778, 2016.

[21] 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.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," *international conference on learning representations*, 2014.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.