# Clinical Pneumonia Classification via Multimodal Data Fusion

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

***Abstract*—Objective: Existing CAD(Computer Aided Detection) systems for pneumonia diagnosis are basically designed for chest X-ray image. But in clinical practice, CT(Computed Tomography) can provide more information and details which can help to increase the accuracy of diagnosing. Applying deep learning methods directly to raw CT scans requires much resources of computing. Moreover, patients' clinical information are often ignored in recent deep learning methods, which is conflicted to clinical practice. The objective of this paper is to effectively address the problems of absent of clinical information in deep learning methods and therefore to accurately diagnosis pneumonia via raw CT scans. Method: We imitate the radiologists' clinical diagnosis process and propose a novel model, MMDD(Multimodal Data Diagnosis), which combines CT visual features with patients' age, gender and complaints. In MMDD, we treat process of reading CT scans as playing short videos and use RCNN(Recurrent Convolutional Neural Network) to capture visual features from CT image data. Meanwhile, complaints will be transformed into word vectors using word2vec and analyzed by LSTM(Long Short Term Memory). Visual features and complaint textual features will be fused together with patients' information about age and gender. All information above will play its role in the decision making process. In order to provide more visual features to MMDD, we extract images from three different ranges of HU values based on original CT scans and transform 1-channel grey level CT images into three-channel false color RGB images. Moreover, we use a auxiliary loss to enhance the gradient propagated to CNN. 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 classify pneumonic cases based on large scale clinical raw data using multimodal clinical data.**

***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

CHEST X-rays used to be the best available method for diagnosing pneumonia, played a crucial role in clinical care[1] and epidemiological studies[2]. However, with the development of economic and technology, CT(Computed Tomography) has become a common technology for diagnosing pneumonia. Compared to chest X-ray, CT can provide more detailed information which can help to diagnosis pneumonia in early stage and avoid delayed treatment.

According to the survey, a radiologist of a major hospital need to read hundreds scans of CT every day. Thus, developing a fast, robust and accurate CAD system to perform automated diagnosis of pneumonia is meaningful and important.

There have been lots of models designed for diagnosing thoracic diseases. Hoo-Chang Shin [3] combined CNN and LSTM[4], 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[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 use deeper neural network to analyze thoracic diseases. They used different pre-trained CNN models on this dataset. Experiments showed that ResNet50 achieved highest AUROC score 0.6333. They also provided ChestX-ray14 with more kinds of thoracic diseases. Based on this database, later in 2017, Yao et al.[7] achieved 0.713 in AUROC score using DenseNet Image Encoder. Pranav Rajpurkar, Andrew Y. Ng et al. [8] developed CheXnet with 121 convolutional layers and achieved 0.7680 in AUROC. 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 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.

Studies above have something in common. First of all, they were designed for chest X-Rays, which means even if they can achieve good results in every indication, these models still cannot handle CT(Computed Tomography) scans, which can provide more detailed visual information. 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. Each slice in CT scans is a 2-D image of human body scan, using these 2-D slices can reconstruct 3-D structure of human bodies. Extensive studies show that 3-D CNN is the best choice for keeping 3-D spatial information in CT[10]. 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

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.

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. Moreover, information of age and gender is also related to some certain diseases. However, as far as we know, few studies use this information to improve the performance of CAD systems.

In general, there are two major problems of existing CAD systems: (1) They cannot handle raw CT scans, and this problem has become a practical limitation of deep learning models designed for medical problems; (2) Few studies consider clinical information like patients’ complaints, which is conflict to clinical practice.

In this study, 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. We treat slices of CT scans as video frames, and transform the problems of analyzing CT scans into the problems of analyzing short videos. We use RCNN(Recurrent Convolutional Neural Network) to extract features from slices of CT scans. Then we embed patients’ complaints into word vectors using word2vec[11][12]. We use another LSTM(Long Short Term Memory) to extract textual features, and concatenate it with features of images, alone with information about patients’ age and gender. Our mode Multimodal Data Diagnosis(MMDD) model, as shown in Fig 2, will learn a joint distribution of all features above and gives out the final classification result.

Our main contributions are listed as follow:

(1) We design a protocol for selecting particular series o scans from raw CT data collected from Radiology Departmen of The First Affiliated Hospital of Army Medical Universit according to ‘Slice Thickness’ and ‘Convolution Kernels’ an transform original CT images into HU value matrices fo future use.

(2) We transform CT images from one-channel grey leve images into three-channel fake color images using three dif ferent ranges of HU values, and analyze the effectiveness o three-channel fake color images via middle outputs of CNN

(3) We propose a novel model: MMDD(Multimodal Dat Diagnosis). This model can capture visual features by playin CT slices as short videos. Meanwhile, MMDD will fuse visual features with clinical information like patient complaint, age and gender to predict whether these cases are pneumonic.

The remainder of the manuscript is organized as follows. Section II describes the architecture of MMDD 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 [13][14], however, only a few work apply RCNN to medical image analyze. Zreik, Majd et al. [15] 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)[16]. They proved that RCNN’s potential for sequence information processing of medical images. Meanwhile, slices of CT scans can be seen as video frames, so in this study, we use RCNN to capture visual features from CT slices.

Follow the study[13], we use LSTM as our RNN cells cause LSTM has been demonstrated to be capable of large-scale learning of sequence data. However, few studies have paid attention to different CNN models’ performance on medical images, so we test three kinds of classic CNN models: VGG[17], ResNet[18] and GoogLeNet with Inception-V3 [19]. We wanted to test deeper models like ResNet101, but it will cost more resource of calculation so we give up. 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 $512 \times 512$, so the outputs of CNN will be very large. We use global average pooling[20], as shown in Fig 1, 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 and fit the requirements of LSTM. Following study in [13], the number of LSTM units is set to 256.

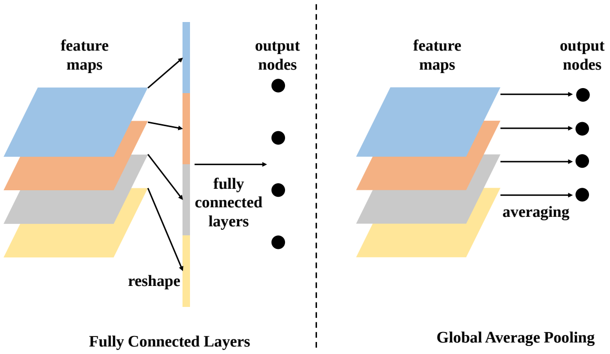


Fig. 1. Difference between Fully Connected Layers and Global Average Pooling

For example, if we use ResNet50, the final feature maps will be $16 \times 16 \times 2048$. If we use fully-connected layer, the length of first fully-connected layer will be $1 \times 524288$, which will make model very difficult to train. If we use global average pooling, outputs from CNN will be reshaped into a tensor of size $1 \times 2048$, then it will be easy to reduce the tensor to $1 \times 256$ using one fully-connected layer. If we have $n$ slice, we will have a matrix with size $n \times 256$, this matrix will be fed into LSTM by $n$ steps.

In order to get the best RCNN for CT scans, we run experiments to get the best combination between CNN models and LSTM. The experiments show that ResNet50 performs the best in these three models, so our RCNN use ResNet50 as its CNN part, and use one layer of LSTM cells as its RNN part. This conclusion is similar to [6], their experiments showed that ResNet50 outperformed GoogLeNet and VGG16.

In fact, LSTM layer plays the role of encoder, it encodes image feature sequences and gives out the output of the last step 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, $t$ is equal to 32.

### B. Multimodal Data Fusion and Multimodal Data Diagnosis

Besides CT image information, we also know patients gender, age, and complaints. For gender and age, we use them as additional features and set a tensor with size $1 \times 2$ to hold values. For patients' complaints, we will use Jieba Chinese word segmentation tool [1] to segment Chinese sentences into word sequences. This process steps will be discussed later in section III. 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. This LSTM is the second encoder to encode complaint. It is calculated in the same way as Eq. 1:

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

$Cx_{ct}$ 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.

We use cross-entropy as classification loss function[15]. After getting $hv_t$, $hc_{ct}$, we can calculate the prediction and loss $\Phi_1$ 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, $\Delta$ 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.

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 [19], we use a auxiliary loss to enhance signal of gradient for CNN. The auxiliary loss $\Phi_2$ and loss of whole model $Loss$ is defined as follow:

TABLE I
WEIGHTS OF TWO LOSSES AT DIFFERENT TRAINING STEP

| ***Number of Steps*** | $1-\omega$ | $\omega$ |
|---|---|---|
| 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 |

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

where $\omega$ is a parameter within the interval (0, 1). $\Phi_2$ is classification cross-entropy loss from CNN, $\Delta_i^c$ is Softmax prediction of CNN. $\omega$ 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, $\Phi_1$ tends to get small and keep updating parameters of LSTM. We output weights of two losses during training MMDD, as shown in Table I, weight for LSTM loss $(1-\omega)$ is 0.6238 at the beginning of training(602 steps), however, W1 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.

Finally, a Multimodal Data Diagnosis(MMDD) is built, RCNN for image data and LSTM for complaints will be trained jointly, the architecture of MMDD is shown in Fig 2.

### 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($1 \times 256$) 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. VGG16 is relatively shallow, ResNet50 is deepest in these three kinds of models. We tried to use deeper network like ResNet101, however using ResNet101 make the training of model very slow and bring a heavy burden to servers. So we keep ResNet50 in RCNN.

The second step is to train MMDD model. We use model 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 0.0005 and drops 50% every 3000 training steps. The dropout rate in fully-connected layers is set to 0.5.

Moreover, we use CNN models pre-trained on ImageNet[21]. Models without pre-training is almost

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

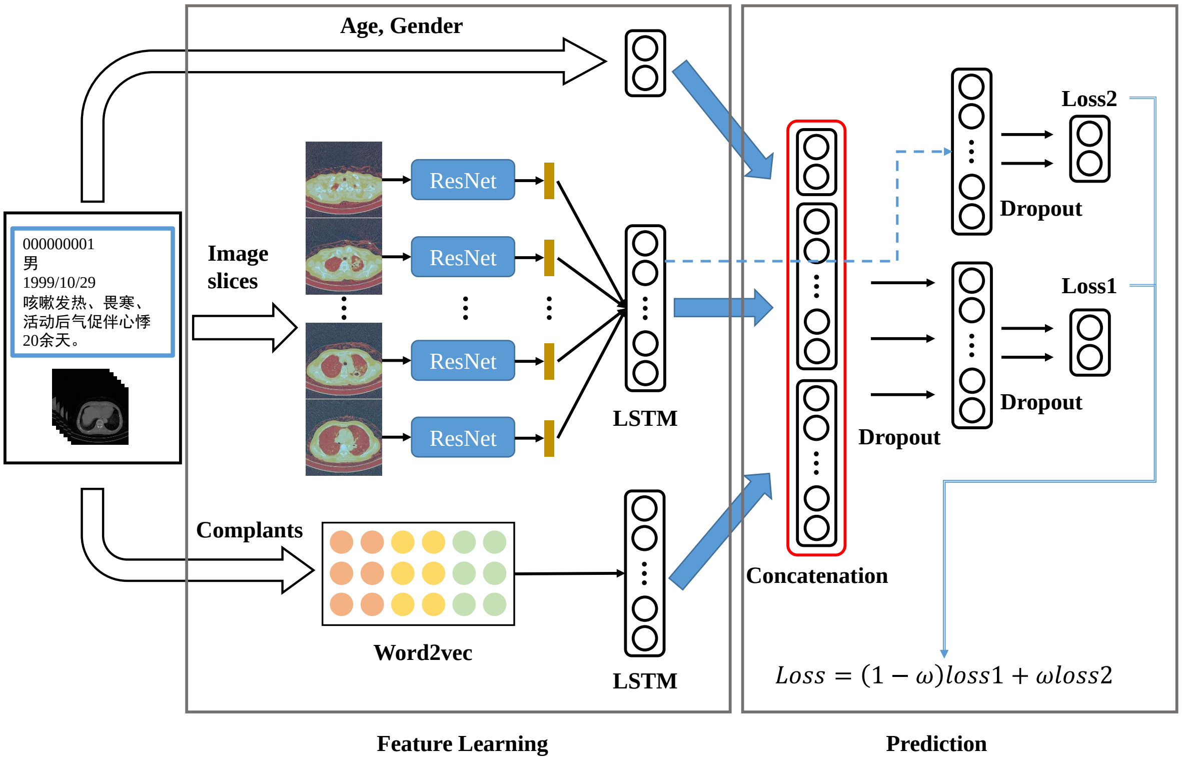


Fig. 2. Multimodal Data Diagnosis Model

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 Fig II.

Model 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 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). Open a CT scan with RadiAnt DICOM Viewer[2], as shown in Fig 4, we can see that raw data from hospital may have more than one series of images(yellow rectangle), each series may have different data type or different image windows. Generally speaking, radiologists and doctors will use series 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 let the computer pick up data for us.

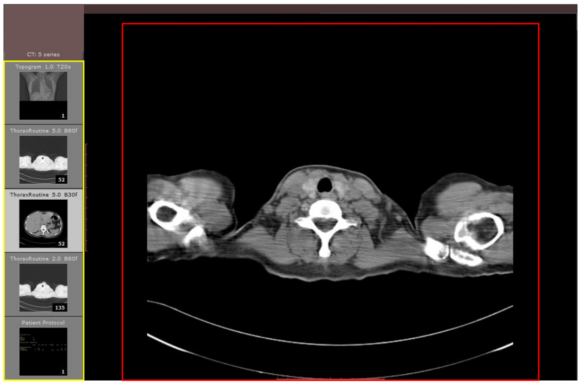


Fig. 4. Structure of Raw Data from Hospital

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'. Specific 'Convolution Kernel' can make the CT more suitable for observing the lungs or chest, for example, in Fig 5,

[2] www.radiantviewer.com

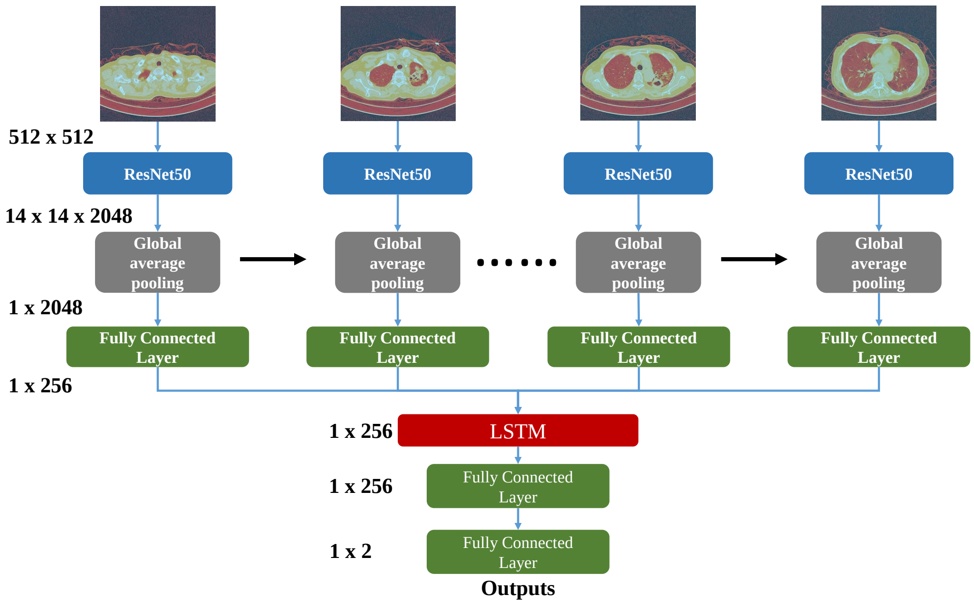


Fig. 3. Architecture of RCNN

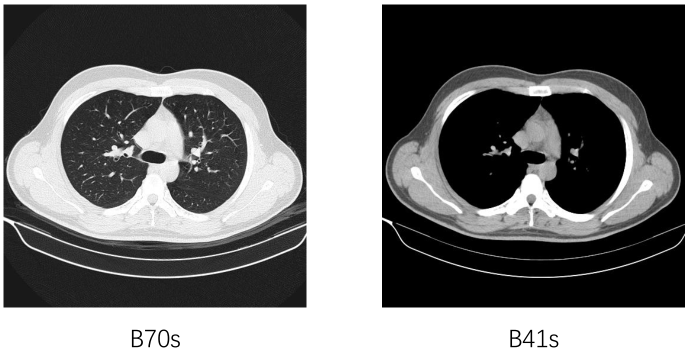


Fig. 5. Scans under Different Convolutional Kernel

slice under ‘B70s’ has clearer view of lungs, slice under ‘B41s’ has clearer view of heart. We need to notice that these names of ‘Convolution Kernel’ vary between 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’, since radiologists use these ‘Convolution Kernel’ more frequently to observe the lungs and chest than other kinds of ‘Convolutional Kernel’. Number of different ‘Convolution Kernel’ is shown in Fig 6. Other kinds of ‘Convolution Kernel’ will be eliminated since series with other kinds of ‘Convolution Kernel’ may have different perspectives of chest. 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 in this Section.

2. 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.

3. 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. Patient complaints are descriptions from patients which describe their own feeling about their condition, which is very important in clinical practice. In this study, we will use gender, age as additional features, and use LSTM extract textual features from patients’ complaints, and combine them with features from CT 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

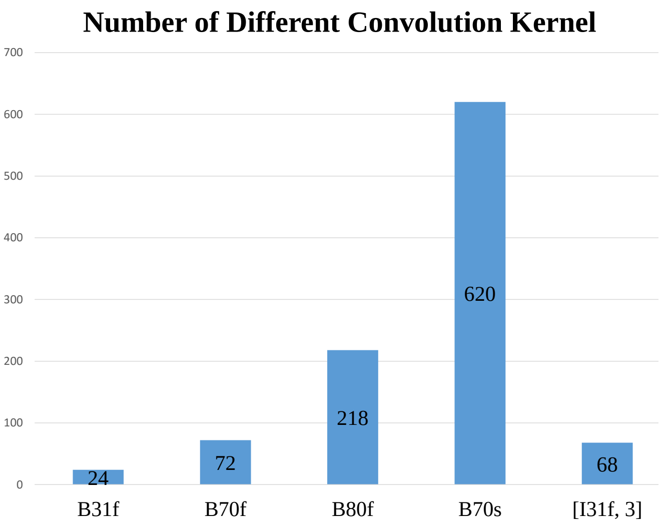


Fig. 6. Number of Different Convolution Kernel

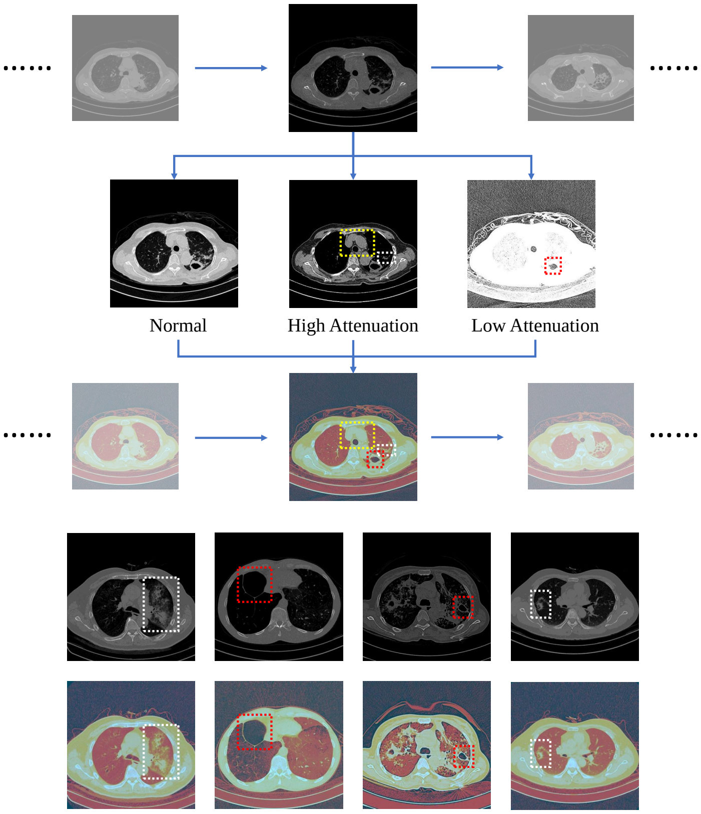


Fig. 7. Data Pre-precess for CT Scans

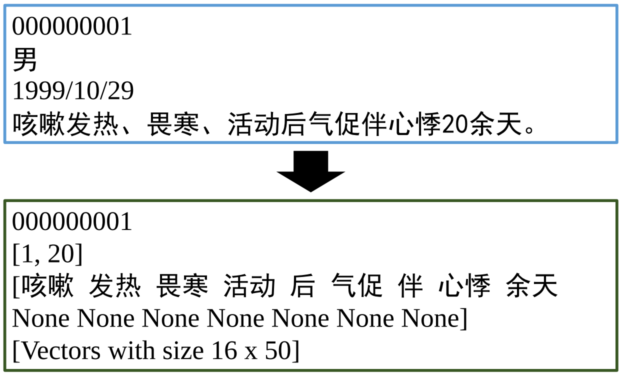
highlight different tissues of bodies. In sectionIII-A1, we ca see that each series of CT actually has one specific ‘Convolu tion Kernel’ and show specific window for CT images directl from raw data. But it may make data inconsistent betwee different cases. So we transform raw data into HU(Hounsfiel Unit) values. The Hounsfield Unit named after Sir Godfre 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 [22] [23], we transform slices into images using three HU range: lung window [-1000, 400HU], high attenuation [-160, 240HU], low attenuation [-1400, -950HU]. In Fig7 we can see that, compared to original CT image, image in ‘Lung Window’ is brighter, tissues in lungs are clearer and details are enhanced. Image in ‘High Attenuation’ have a clear view of hearts and vessels(in yellow rectangle). ‘High Attenuation’ range also enhance the difference between high dense pathological tissues(in white rectangle) and normal tissues. ‘Low Attenuation’ range highlight abnormal voids in lungs(in red rectangle) which is features of severe lung diseases. 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, as shown in Fig 7. The ‘Slice Thickness’ between each slice is adjusted into 10mm, and each case will keep 32 slices.

Compared to original CT images, 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 this modal, 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. For example, as shown in Fig 7, 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. The influence of different HU value ranges will be discussed int 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 8. For each patient, we have a list of information contains age, gender and complaints. For patient age and gender, we transform them into a two-dimensional array. For example, patient in 8 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



000000001
男
1999/10/29
咳嗽发热、畏寒、活动后气促伴心悸20余天。

000000001
[1, 20]
[咳嗽 发热 畏寒 活动 后 气促 伴 心悸 余天
None None None None None None None]
[Vectors with size 16 x 50]

Fig. 8. Data Pre-precess for Age, Gender and Complaints

a short cut and use a mature tools: Jieba text segmentation to segment Chinese sentences into Chinese word sequences. Before segmentation, we remove numbers, punctuation marks in order to get a better segmentation results. An example of Chinese word segmentation is shown in green rectangle in Fig 8. However, English patient complaints have no need to do segmentation. If you use data from English speaking countries, you may skip this step. After segmentation, we use word2vec to embed word sequences into vectors. We use CBOW(Continuous Bag-of-Words)[11] to capture 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 increase 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.

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

### B. Effectiveness of Three-Channel Image

In order to verify the effectiveness of three-channel preprocessing, we output the feature maps of convolutional layer with three-channel images, lung window images, high attenuation images and low attenuation images as show in Fig 9. In order to keep experiments environment consistent, all experiments carried on in this part is based on RCNN with ResNet50. In Fig 9, 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.

To verify the effectiveness of difference ranges of HU values, we output middle feature maps of ResNet. More specificity, we output the feature maps after one convolutional layer, one max pooling layer, and three blocks, the size of feature maps are $128 \times 128$, then we resize them into $512 \times 512$. We can see that CNN trained by three-channel images has advantages over CNNs trained by other kinds of 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 this. In the third row, low attenuation CNN cannot distinguish the normal and unnormal tissues. Moreover, in the forth row, low attenuation CNN simply 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.

### C. 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 IV.

First of all, we do 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 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 0.945 in accuracy, 0.956 in specificity, 0.945 in AUROC score, which are highest in experiments. RCNN(ResNet) trained by lung window image has 0.954 in sensitivity, which is the highest in experiments, but it only achieves 0.824 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 image, but we set $\omega$ 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 indication. It proves that, by using auxiliary loss, CNN will be trained in a better way.

At last, we run experiments to prove than Multimodal Data can enhance the performance of CAD system. As shown in section 2, the output of RCNN($1 \times 256$), features of complaints($1 \times 50$), gender($1 \times 1$) and age($1 \times 1$) will be concatenated together($1 \times 308$) and fused by two fully-connected

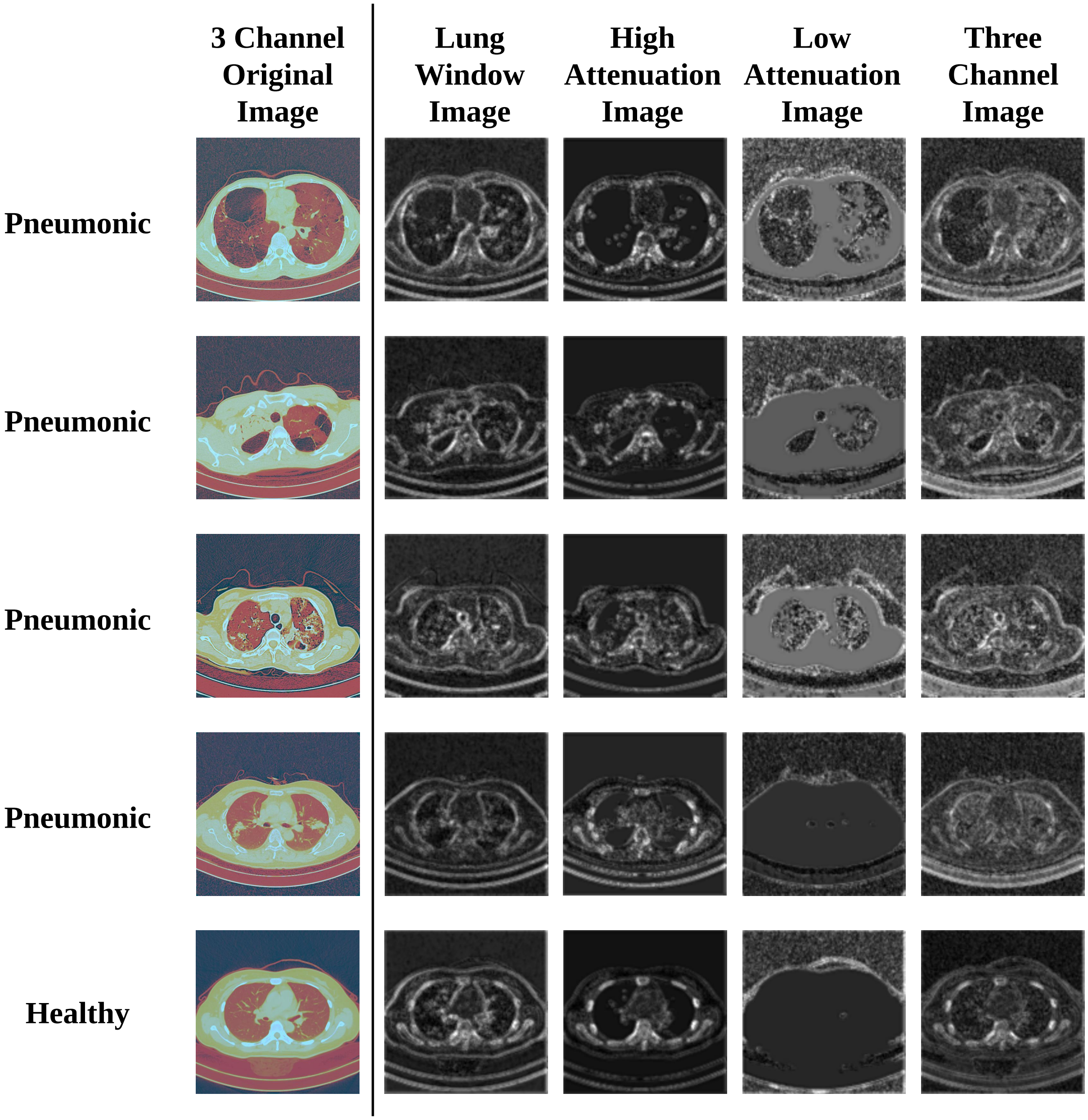


Fig. 9. Convolutional Feature Maps from CNN Models Trained by Different Images

layers. It is simple, but effective. We can see that MMDD 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 MMDD has the best performance of binary classification according to its AUROC score. We also remove the information about age and gender, we found that MMDD without age and gender has a higher score in sensitivity and lower score in specificity.

## IV. Discussion

The validation loss and accuracy during training is shown in Fig 10 and Fig 11. We can see that in Fig 10, MMDD can achieve higher accuracy at the first phase of training, but RCNN(ResNet), RCNN(GoogLeNet) and RCNN(ResNet) with single loss perform better after. MMDD's performance outperforms other methods again after 27000 training step. We can also see that RCNN(ResNet), RCNN(ResNet) with single loss have similar performance during training, but RCNN(ResNet) with auxiliary loss perform a little bit better than RCNN(ResNet) with single loss at the end of training. For RCNN(GoogLeNet), it converge slower than RCNN(ResNet) and RCNN(ResNet) with single loss, and has a lower accuracy in the end. Without age and gender, MMDD has similar performance compared to RCNN(ResNet) with single loss. According to the experiments above, we can see that information about age and gender 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

TABLE IV
Comparison of All Kinds of RCNN and MMDD

| ***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 |
| MMDD | Three Channel Image & Complaints | 0.925 | 0.945 | 0.901 | 0.923 | 3 |
| MMDD | Multimodal Data | **0.945** | 0.936 | **0.956** | **0.945** | 1 |

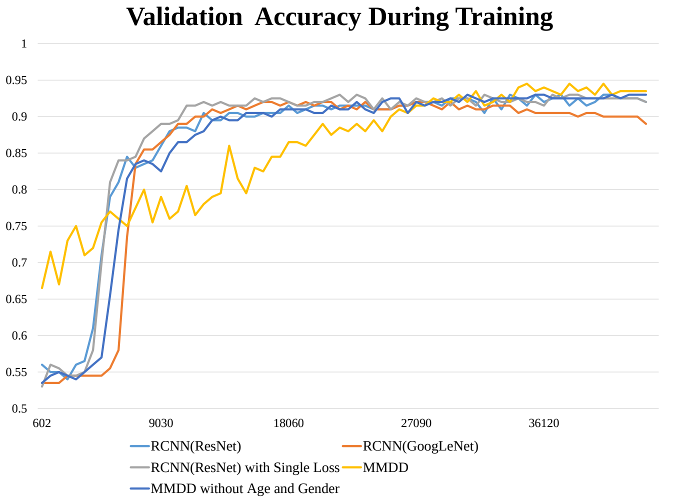


Fig. 10. Validation Accuracy During Training

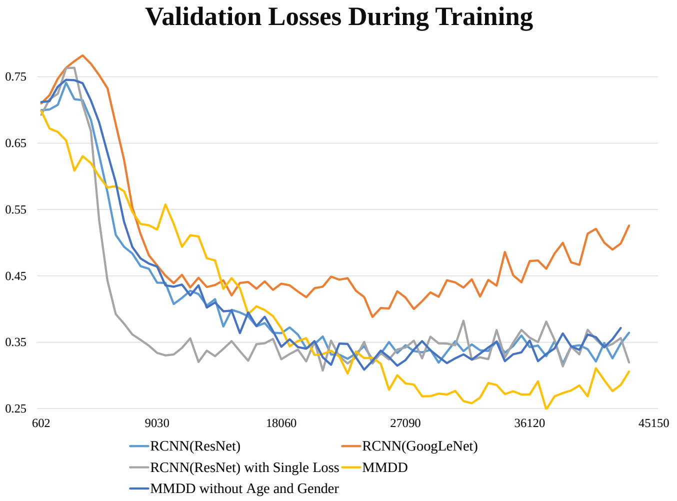


Fig. 11. Validation Loss During Training

the number of male patients and female patients in healthy cases and pneumonic cases(Table V) and number of patients in different ages(Table VI).

In Table V, 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 VI, we can see that age is also related to the chance of being pneumonic. People turn to hospital or clinic may have noticed that something goes wrong about their healthy condition, but we can still see that people older than 40 have much larger chance of being pneumonic. There are about half of healthy cases between 40-50, but this number 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. It reminds us that there is a lot of work need to be done about old people health care.

TABLE V
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 11, we can see that RCNN(GoogLeNet) has the highest loss at the end of training, so it perform the worst in accuracy. RCNN(ResNet) and MMDD 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. MMDD has the lowest loss at 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 distribution, MMDD has the lowest loss at beginning, which is corresponding to the

TABLE VI
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

phenomenon in Fig 10. But MMDD has more parameters to train, so its loss drops slower than others.

## V. CONCLUSION

In this study, we propose a novel model, MMDD(Multimodal Data Diagnosis), which combines CT visual features with patients' age, gender and complaints. In MMDD, CT scans will be treated like videos, 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 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.

## 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] 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.

[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] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

PLACE
PHOTO
HERE

**Michael Shell** Biography text here.

**John Doe** Biography text here.

**Jane Doe** Biography text here.