Deep Regression via Multi-Channel Multi-Modal Learning for Pneumonia Screening

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

***Abstract*—Pneumonia screening is one of the most crucial steps in the pneumonia diagnosing system. This paper proposes a deep regression framework based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for automatic pneumonia screening, which simulates the clinical diagnosis process. Given a single case, the novel framework jointly learns the multi-channel images and multimodal information (i.e., clinical chief complaints and demographic information) then gives out the detection results. We demonstrate the advantages of the framework in several ways. First, we propose to treat chest CT scans as short video frames and analyze them using Recurrent Convolutional Neural Network (RCNN), which can make the most of 3D spatial information and reduce the need for calculation resources. Second, visual features from multi-channel images (Lung Window Images, High Attenuation Images, Low Attenuation Images), which are transformed from one-channel gray-scale CT scans, can provide supplementary features to each other and give qualitative information for pneumonia detection. Third, chief complaints can provide information like lesion locations and symptoms which can enhance the features from images and improve the specificity of the framework. Finally, demographic information (i.e., age and gender) contains prior information which can improve the overall performances. The proposed framework has been extensively validated in 1002 clinical cases collected from the First Affiliated Hospital of Army Medical University. Our network achieves 0.930 in accuracy and has a very balanced performance in sensitivity and specificity. As far as we know, we are the first to screen pneumonia using large scale clinical data with clinical and demographic information. Our method demonstrates that demographic and clinical information can provide more abundant information than image data only and get very convincing results. While the proposed framework is tailored for pneumonia screening, it can be extended to include more multimodal clinical data, and give out more reliable and explainable results.**

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

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

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

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

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

I. Introduction

PNEUMONIA is a prevalent thoracic disease in daily life. In clinical practice, radiologists need to consider multimodal information (i.e., images, chief complaints, patient age or gender) to screen pneumonia cases from or massive clinical data. Conventionally, this task relies on experts' manual operations, which is time consuming and inhibits fully automatic assessment. Thus, developing a fast, robust, and accurate CAD system to perform automated screening of pneumonia is meaningful and vital.

Many researches have devoted efforts in pneumonia screening, detection, monitoring and diagnosing like [1]–[5]. There are two major data types which are analyzed in these researches: chest X-Ray and CT.

The most common data type that used is chest X-Ray. Hoo-Chang Shin [1] proposed a method used CNN to extract features from chest X-Rays and used LSTM [4] to generated MeSH [6] terms for chest X-Rays. 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 to use deeper neural networks to analyze thoracic diseases. They tested different pre-trained CNN models on this dataset and showed that ResNet50 achieved the highest AUROC score of 0.6333 in classifying pneumonia. They also provided ChestX-ray14, which contains more kinds of thoracic diseases. Based on this database, later in 2017, Yao et al. [8] achieved AUROC of 0.713 in classifying pneumonia 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. [5] proposed TieNet, which could classify the chest X-Rays into different diseases and generate the report at the same time. In TieNet, CNN was used to capture features of chest X-Rays, RNN was used to learn these features and generate reports based on attention mechanism, which could help the model to focus on different parts of chest X-Rays alone with the generation of reports. In the pneumonia classification problem, they achieved 0.947 in AUROC based on reports, and reached 0.917 in AUROC on hand-labeled data.

A few studies focus on analyzing chest CT scans. Hoo-Chang Shin et al. [10] exploited three important, but previously understudied factors of employing deep convolutional neural networks to computer-aided detection problems. They used 2D CT slices for ILD (interstitial lung disease detection). Mingchen Gao et al. [11] presented a method to classify ILD imaging patterns on CT images. They also used 2D CT slices

as inputs of their models.

II. Motivation and Contributions

A. Motivation

Screening pneumonia cases from massive clinical data is time consuming. Thus, accurate screening of pneumonia can improve the working efficiency of radiologists and prevent delayed treatments. Methods mentioned above have some drawbacks in common.

First, chest X-Ray used to be the best available data for screening pneumonia, played a crucial role in clinical care and epidemiological studies [12], [13]. However, compared to chest X-Ray, CT scans have a more unobstructed view of patients' bodies and allow visualization of 3D lung structures [14], since bones, skin, vessels and lung tissues may cause overlapping shadows in chest X-Rays and cause misdiagnosis.

Extensive studies show that 3D CNN is the best choice for keeping 3D spatial information in CT [15]. However, 3D CNN cannot be applied to raw CT data directly since it will bring a heavy burden to computers. According to clinical requirements, radiologists need to measure the lesions accurately, so we cannot reduce the size of images by resizing at will. However, we can treat CT scans as short video frames, so that we can analyze CT 3D spatial information and reduce the need for calculation resources.

Second, these methods heavily rely on image information. Few of them combine image visual features with clinical information or demographic information. Models like TieNet do combine image visual features with descriptions about images written by radiologists. However, we believe using descriptions about images written by radiologists to improve models is not entirely convincing since descriptions like 'Findings' and 'Impressions' sometimes include diagnosis conclusions. Patients' chief complaints are valuable when doctors are making decisions [16], since chief complaints are patients' direct feeling about their physical condition, telling the patients' pain location, symptoms and how long have they been ill. Demographic information of patients is strongly connected to the condition of lungs and patients. Many studies have proved that demographic information can help improve the classification/regression performance in CAD systems [17]–[20]. However, as far as we know, few studies have used these information to improve CAD systems for pneumonia.

We show in Fig 1 the clinical process of pneumonia screening. Compared to clinical process, there are two major drawbacks of existing CAD systems for pneumonia: (1) They cannot handle raw CT scans, which allows visualization of lung structures; (2) Few studies consider multimodal clinical information like demographic information (e.g., age, gender) and clinical information (e.g., chief complaints), which is a conflict to clinical practice. If we use these methods directly, the results may not be ideal.

To address such drawbacks, we propose a novel deep regression framework for pneumonia screening to jointly learn the multi-channel image slices and multimodal clinical, demographic information. Moreover, our framework treats CT slices as short video frames so that we can keep 3D spatial information and reduce the need for calculation resources at the same time.

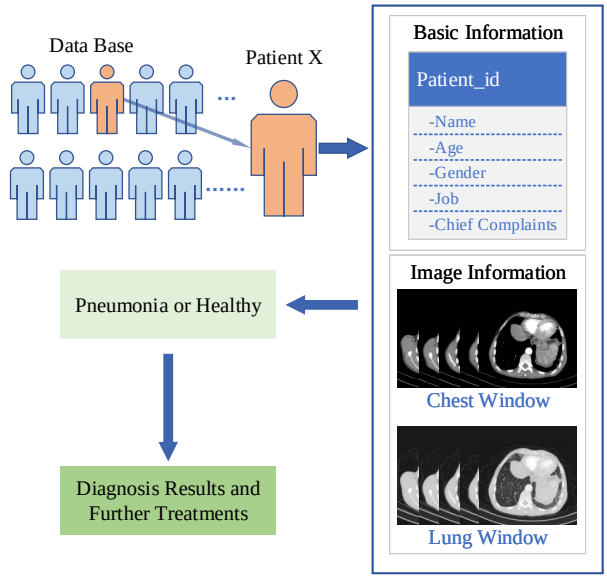


Fig. 1. Clinical process of pneumonia screening. The patients' information is kept in data base PACS (Picture Archiving and Communication Systems). Each patient case contains patient basic information like age, gender, name, chief compilations and so on. Meanwhile, image information like CT is kept in PACS with corresponding id. A radiologist needs to consider multi-modal information and refers to images under different windows to give out screening or diagnosing results.

Fig 2 illustrates schematic diagram of our framework. Herein, (a) each CT image will be transformed into a multi-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. All slices will be transformed in sequences and analyzed by a RCNN (Recurrent Convolutional Neural Network). (b) We also include clinical data in our framework. Chief complaints can provide the location of pain, symptoms, and how long have patients been ill. This information is related to the CT image and enhances the visual features extracted from CT. (c) Demographic information about age and gender can provide prior information since patients of different ages and genders have differences in the morphology of the thoracic cavity and lungs. (d) The RCNN captures visual features from each multi-channel slices in sequences. An other LSTM is used to analyze semantics from chief complaints. Demographic information will be used as confounding factors.

B. Contributions

The main contributions of this work are as follows.

i) We formulate the clinical pneumonia detection process as a regression problem. The proposed framework, which simulates the clinical diagnosing process, takes advantage of

the multi-channel images, multi-modal information and avoids the shortcomings of a single source of information.

ii) We propose to treat chest CT scans as short video frames and analyze them using Recurrent Convolutional Neural Network (RCNN). RCNN can make the most of CT image information and reduce the need for calculation resources. Meanwhile, we verify that ResNet50 performs the best when it is combined in RCNN.

iii) We demonstrate that different channels of images can provide different density information during the learning process and help to improve the detection accuracy. We further propose that clinical and demographic information can enhance visual features and provide prior information.

iv) We demonstrate experimentally good performance on a large clinical dataset collected from the First Affiliated Hospital of Army Medical University with 1002 cases.

III. Evaluation Datasets

TABLE I

Number of Male and Female Patients in HC and PC

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

Number of HC and PC in Different Ages

	Healthy	***pneumonia***	***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.

There have been many datasets for thoracic disease research, but datasets contains CT, demographic and clinical information are rare. To evaluate our framework, we use the raw data collected from the Radiology Department of The First Affiliated Hospital of Army Medical University. In this study, we have 552 pneumonia cases and 450 healthy cases (1002 cases total) from hospital PACS (Picture Archiving and Communication Systems) in the last three years (2016 - 2019). We show in Table I the details of this dataset.

Raw data from the hospital may have more than one series of images, and each series has specific data types, image windows, or view angles. Generally speaking, radiologists and doctors will use the series under lung window with the smallest 'Slice Thickness', but for deep learning models, each case can only have one series. So we design a protocol to pick up specific series for us:

(a) We choose the series with the specific 'Convolution Kernel'. Different 'Convolution Kernel' indicate different image windows. We need to notice that these names of 'Convolution Kernel' vary between hospitals and CT equipment. In our study, we choose 'B31f', 'I31f 3', 'B70f', 'B80f', 'B70s'. We notice that in the Radiology Department of The First Affiliated Hospital of Army Medical University, 'B70s' is the most common parameter used in clinical which contains 620 cases.

(b) We remove series like 'Patient Protocol', 'Topogram'. These series contain some basic parameters and information about CT equipment, which are not suitable for deep learning.

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

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

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

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

IV. Deep Regression Framework for Pneumonia Detection

A. Overview

The proposed framework, with a schematic illustration shown in Fig 2, can be described as a multi-channel multi-modal learning framework. This framework can be regarded as a regression model for classification. All cases will be classified into two Categories: pneumonia cases and healthy cases. As shown in Fig 2, there are three kinds of inputs: (1)multi-channel images, (2) clinical information, (3) demographic information. Multi-channel images have three image windows. High Attenuation Window can provide high density information of lungs, Low Attenuation Window can provide low density information, Lung Window can provide general lung information. The regression progress can be formulated as:

$$P(X) = Softmax(F(V(X) \otimes C(X) \otimes A(X) \otimes G(X)))$$

X is the input case, $\otimes$ is the concatenation operation, $V(X)$ is visual features captured from multi-channel images, $C(X)$ is semantic information captured from clinical chief compilations. $A(X)$ and $G(X)$ indicate patient age and gender. F is a function to fit the regression model of $V(X)$, $C(X)$, $A(X)$ and $G(X)$. $P(X)$ characterizes the likelihood of being pneumonic.

In our framework, a RCNN is used to learn multi-channel CT images and get $V(X)$, a LSTM is used to learn semantics

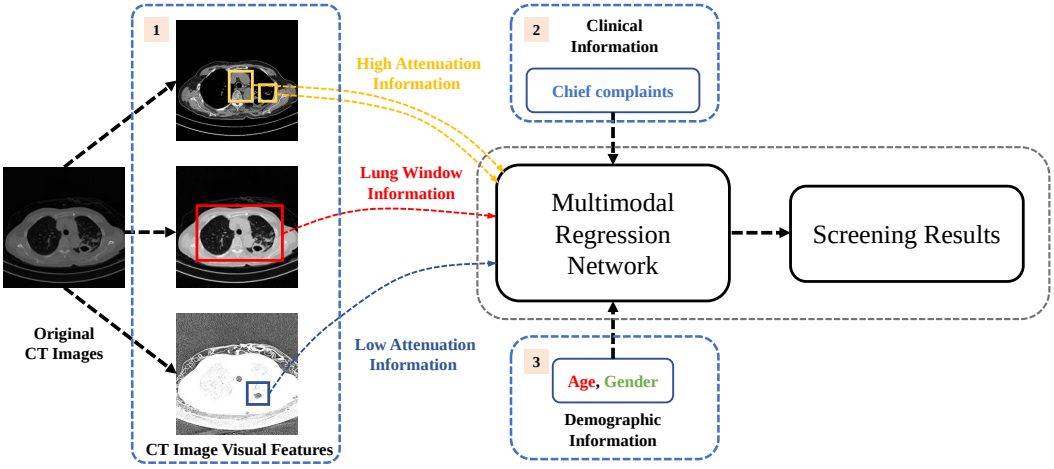


Fig. 2. Illustration of the proposed deep regression framework via multi-channel multi-modal learning for pneumonia detection. The inputs of this models are three-channel CT images, clinical chief complaints and demographic information. A RCNN (Recurrent Convolutional Neural Network) is used to capture the visual features. A LSTM is used to analyse chief complaints. Visual features, chief complaints and demographic information will be fed into a deep regression network.

of clinical chief complaints. Following studies like [21], [22], we treat the demographic information (i.e., age and gender) as confounding factors.

B. Recurrent Convolutional Neural Network

RCNN (Recurrent Convolutional Neural Network) has been proved to be very effective in video caption, description, and classification [23], [24], some studies have applied RCNN to medical image analysis. Majd Zreik et al. [25] recently used RCNN for automatic detection and classification of coronary artery plaque, they used CNN extracts features out of $25 \times 25 \times 25$ voxels cubes and used an RNN to process the entire sequence using gated recurrent units (GRUs) [26]. KL Tseng et al. [27] exploited convolutional LSTM to model a sequence of 2D slices, and jointly learn the multi-modalities and convolutional LSTM in an end-to-end manner to segment 3D biomedical images.

As mentioned in section II, CT allows visualization of lung structures, which brings a large amount of redundant information, like muscle, vessels, and bones. It will cost lots of calculation resource if we use 3D CNN directly. However, if we treat CT slices as short video frames, we can analyze them using RCNN instead. 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 RNN, so that we can reduce the need for calculation resource and keep 3D spatial information at the same time. This RCNN is actually a encoder for visual features from CT mulit-channel slices. V mentioned above can be calculated as:

$$V_t = LSTM(Fx_t, V_{t-1}, z_{t-1}) \tag{1}$$

Fx_t is the t-th visual features in CT slices, V_{t-1} is LSTM hidden state of $t-1$ step, z_{t-1} is LSTM output of $t-1$ step. t is the current step of LSTM. In this study, each scan has 32 steps.

1) Convolutional Neural Network: In this study, we use ResNet50 in RCNN. We compare three kinds of classic CNN models: VGG16 [28], ResNet [29] and GoogLeNet with Inception-V3 [30] and our experiments demonstrate that ResNet50 performs the best. This part of experiments will be discussed later in section V. We use CNN without fully-connected layers as a feature extractor. We use ResNet50 without fully-connected layers as a feature extractor. The input size of CNN is 512×512.

2) Global Average Pooling: Since the input size of CNN is 512×512, the outputs of CNN will be extensive. We use the global average pooling [31] to reduce the number of neurons significantly. It is a replacement of fully-connected layers to enable the summing of spatial information of feature maps. After global average pooling, we insert a fully-connected layer to reduce dimensions to 128 to fit the number of LSTM units.

3) Long Short-Term Memory: Recurrent neural networks (RNNs) [32] are a rich class of dynamic models that have been used to generate sequences in domains as diverse as text and motion capture data. There have been severl kinds of RNN units like GRU [33] and LSTM. In this study, we use LSTM as our RNN cells cause LSTM has been demonstrated to be capable of large-scale learning of sequence data.

C. Multi-Channel Images Representation

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

Following the study in [10], [11], HU value matrices will be transformed into images using different HU windows. Let

X denote the mulit-channel image input, then X can be calculated as:

$$X = \sum_{q \in \chi} Threshold_q[Min, Max]$$

$\chi = \{LW, HA, LA\}$, where LW is Lung Window, HA is High Attenuation Window, LA is Low Attenuation Window. $Threshold$ is a threshod method, Min and Max indicate upper and lower limits of HU values. Each image window has corresponding Min and Max. Our threshods are hand crafted and little different from studies [10], [11]:

$$LW = Threshold_{LW}[-1000, 400HU]$$
$$HA = Threshold_{HA}[-160, 240HU]$$
$$LA = Threshold_{LA}[-1400, -950HU]$$

For each slice, it will generate three one-channel grayscale images. Then we add three one-channel grayscale images into one three-channel false-color RGB image. The 'Slice Thickness' between each slice is adjusted into 10mm, and each case will keep 32 slices. As shown in Fig 3, three-channel images can show more information about lung density. Original CT images are grayscale images; high dense tissues are white; normal lung tissues and low dense tissues tend to be black. Relatively, three-channel false-color images have a larger scale of colors. First of all, high dense tissues will still tend to be white, like bones, high dense tissues in the lungs. Second, normal lung tissues will tend to be red, and low dense tissues tend to be black, which is very useful when patients have severe lung diseases.

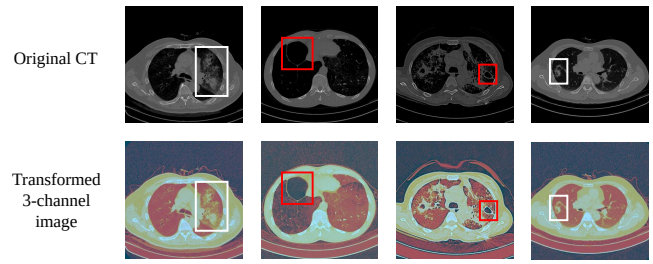


Fig. 3. Examples of three-channel images. In this figure, void space (in red rectangle) in original CT images is not very obvious since other normal tissues are in black too. But in the three-channel images, we notice the differences between normal tissues and low dense tissues. Moreover, the details of high dense tissues (in the white rectangle) are still kept in three-channel images.

D. Multi-Modal Information Representation

Studies like [21], [22] treated the demographic information as confounding factors. The main disadvantage of such a strategy is that the original representations of subjects will be modified because this strategy adds up several steps of engineered pre-processing in a directed and engineered way. Intuitively, it could further promote the learning performance by adding up demographic information in CAD systems [20]. Following these studies, we also treat demographic information as confounding factors. The demographic information of all studied patients is listed in Table I.

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

After word segmentation, we use word2vec [34], [35] to embed word sequences into vectors and use CBOW(Continuous Bag-of-Words) to capture relationship between words. Since our corpus is very small, the embedding size is set to 50, and the window size for CBOW is set to 3. We set length of Chinese word sequence to 16 since 16 is the maximum length among all chief complaint sequences. For those sequences whose length is less than 16, we add 'None' to fill up the voids and increase the length to 16. Then the sequences will be fed into LSTM:

$$C_{ct} = LSTM(Word_{ct}, C_{ct-1}, z_{ct-1}) \quad (2)$$

$Word_{ct}$ is word embedding matrix of the ct-th word in chief complaint, C_{ct-1} is LSTM hidden state of $ct-1$ step. ct is the current step of this LSTM.

E. Ensemble of Decisions

The over architecture of our framework is shown in Fig 4. All information will be fed into a regression model to calculate the likelihood P. We use cross-entropy function as loss function of our model:

$$\arg\min_{W} -\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{N}\sum_{X_n \in \chi} 1\{y_n^q = q\} log(P(y_n^q = q|X_n; W)$$

$\chi = \{X_n\}_{n=1}^{N}$ denotes the training set, X_n represents n-th case of training set. y^q are vectors for labels. In this study, the class labels are used in a back-propagation procedure to update the network weights in the convolutional layers, LSTM units, and learn the most relevant features in the fully-connected layers. W denotes the parameters of the model. Q is equal to 2, which indicates two classes in our data.

V. Experiments

A. Experimental Setup

The dataset is divided into training set, validating set and testing set. 602 cases are used to train the framework. Validating and testing sets each has 200 cases. Three sets have the same data distribution.

Source code for data pre-processing and the proposed framework will be released very soon. We will also release the model with trained parameters and some sample cases. But we cannot release dataset because of the privacy of patients.

B. Parameters

The initial learning rate is set to 0.0005 and drops 50% every 3000 training steps. The dropout rate in fully-connected layers is set to 0.5. Our framework will be trained for four epoch, and each epoch contains 15 iterations for all training

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

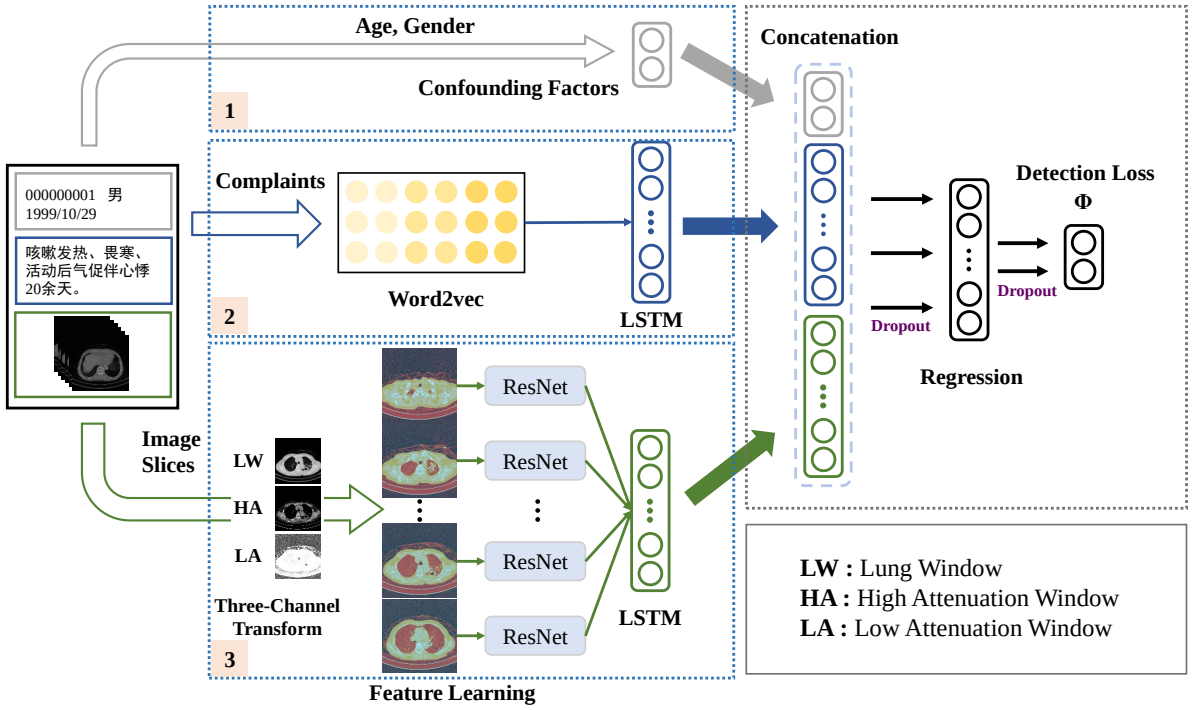


Fig. 4. Overview of the deep regression framework via multi-channel mulit-modal learning for pneumonia screening. The black box in the left indicates raw inputs from the hospital.The grey rectangle contains demographic information about age and gender; the blue rectangle contains chief complaints; the green rectangle provides CT image data. Chief complaints will be transformed into matrices by Word2vec and analyze by one LSTM network. Images will be fed into RCNN. Age and gender will be treated as two confounding factors. These three kinds of information will be concatenated and fed into the regression model to simulate the clinical process.

data. Moreover, we use CNN models pre-trained on ImageNet [36]. Experiments demonstrate that using pre-trained models can significantly improve the converging speed. In this work, all the experiments are run on a GPU of NVIDIA Tesla V100. Each experiment needs about 12G of GPU memory.

C. Baseline System

To mainifest the advantages offered by the proposed framework, we constructed a baseline framework for compartion: RCNN trained with lung window CT images. Generally speaking, we should have a 3D CNN as our baseline. However, the input size is so large that it is impossble to train a 3D CNN under our existing conditions.

We alse developed and repeated the experiments with different RCNN since different CNN models may have impacts on the performance of RCNN. We test three different classic CNN models: ResNet50, VGG16 and GoogLeNet with Inception-V3. We choose the RCNN with the best performance as our baseline system.

D. Analyze of Multi-Channel Images

In this section, we demonstrate the impacts of mulit-channel images. To remove the impact of clinical and demographic information, the experiments in this section will be carried on RNN models with different CNNs.

We test three kinds of classic CNN models: VGG16 [28], ResNet [29] and GoogLeNet with Inception-V3 [30]. Table II provides a comprehensive performance comparison on the validating and testing set with different conbination of RCNNs and input channels.

As can be seen in Table II, three-channel images can provide more complete visual information of lungs, and RCNN(ResNet) trained with three-channel images have the best performances compared to RCNN(ResNet) models trained with LW, HA, and LA images. According to experimental results, RCNN(ResNet) trained with lung window images performs the best in three kinds of image windows, which agrees with the clinical practice. As mentioned in section III, lung window (i.e., 'B70s') is the most common image window in clinical practice.

Taken alone, low attenuation images and high attenuation images cannot provide enough visual features for RCNN. But when three channels are combined, RCNN achieves the best performance, which means low attenuation images and high attenuation images provide supplementary information for lung window images. It also agrees with clinical practice, since radiologists need to change image windows and look for details if they need to deal with cases with severe diseases.

Moreover, as can be seen in Table II, ResNet50 has a better performance on visual features learning than VGG16 and GoogLeNet with Inception-V3. This conclusion is similar to the conclusion drawn in [7], and their experiments showed that ResNet50 outperformed GoogLeNet and VGG16. According to the Table II, RCNN(ResNet) and RCNN(GoogLeNet) both achieve 0.930 in validation accuracy, RCNN(GoogLeNet) achieves 0.932 in validatio AUROC, a little better than RCNN(ResNet). However, RCNN(ResNet) has the best performances in both test accuracy and test AUROC. It supports

TABLE II
COMPARISON OF ALL KINDS OF RCNN

Structure	*Input Channel*	*Validation Accuracy*	*Validation AUROC*	*Test Accuracy*	*Test AUROC*
RCNN(VGG)	Three Channels	0.895	0.894	0.845	0.841
RCNN(GoogLeNet)	Three Channels	**0.930**	**0.932**	0.900	0.902
RCNN(ResNet)	Three Channels	**0.930**	0.929	**0.915**	**0.914**
RCNN(ResNet)	Lung Window	0.925	0.920	0.895	0.891
RCNN(ResNet)	Low Attenuation	0.875	0.877	0.785	0.784
RCNN(ResNet)	High Attenuation	0.910	0.915	0.860	0.864

our selection of RCNN(ResNet50) trained with three-channel images as our baseline system.

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

In Fig 5, images in the first column are original false-color CT images, which are direct outputs from CT slices. The second, the third and the fourth columns are feature maps from LW CNN, HA CNN, and LA CNN. Images in the last column are feature maps from three-channel CNN.

According to Fig 5, HA window can keep high dense information, but HA has difficulty in capturing the difference between low dense tissues and normal tissues. Contrarily, LA can keep low dense information well, but high dense information tends to be blank in LA. LW window is close to the three-channel window. However, the three-channel window has better discrimination for normal tissues and low dense tissues.

E. Performance Comparison

In order to evaluate chief complaints and demographic information separatly, we design two frameworks: one contains all information, the ohter one contains only clinical chief complaints. 'RCNN' in this section refers to RCNN(ResNet).

We show experimental results on testing set in Table III. As mentioned in section IV, the output of RCNN, features of clinical and demographic information will be concatenated together and fused by two fully-connected layers. It is simple yet effective. Compared to baseline, clinical chief complaints bring a significant improvements in specificity,but sensitivity will drop to 0.904. This phenomenon indicates that chief complaints can help to classifying healthy cases, but help little in finding pneumonia.

If we add demographic information (i.e., age and gender), the sensitivity will increase to 0.9302, which is the highest in Table III. However, the specificity will drop to 0.9298, but still higher than that in RCNN (0.907). The accuracy is still 0.930, but the framework has a more balance performance. However, in clinical practice, sensitivity is a more important indicator, so that we consider the framework trained with both clinical and demographic information has better clinical performances.

Moreover, we compare our method with studies in [5], [7]–[9]. Noted that these studies conducted experiments on chest X-Ray images, which is different from the data we use. Table III shows that TieNet [5] achieved an AUROC of 0.969, which is the highest among these studies. However, this AUROC was achieved by analyzing X-Ray images and reports written by radiologists. If they use images only, the AUROC will drop to 0.658, which indicates that reports provided very string information which is related to pneumonia. We believe it is because the reports written by radiologists contain some conclusions and diagnose results. Using reports to classify images is not quite convincing since radiologists have done the job. Actually in [5], reports only can achieve 0.947, which is much higher than X-Ray images. This pneumonia conforms our assumption.

We further report the confusion matrices of all experiments on validating set in Fig 6 so that we can comprehensive compare the impacts of multi-channel images and mulit-modal information. We have 200 cases in testing set. Among these cases, 114 cases are pneumonia, the rest are healthy cases. The framework trained with multi-channel multi-modal data detects 106 pneumonia cases correctly, higher than ohter models. RCNN(ResNet) trained with three-channel images detects 105 cases of pneumonia, 78 healthy cases. The framework trained with clinical information (i.e., chief complaints) detects 83 healthy cases correctly, which is the best among all models, but its performance in screening pneumonia cases is worse than the framework trained with multi-channel multi-modal data and RCNN(ResNet).

Four important points should be noticed from the Fig 6. Firstly, lung window images are able to cope with normal situations in clinical practice. 92.5% cases can be classified correctly using lung window images. Secondly, high attenuation window are second choice since normal pneumonia tends to cause increased lung density and high attenuation window can keep these features well. Thirdly, the low attenuation window is special, because abnormal decrease in lung density is rare. Only a few patients have such severe pneumonia that cause abnormal decrease in lung density. In spite of the small number, we still have to consider this situation. Framework trained with low attenuation images only performs the worst among three kinds of image windows. However, it can provide necessary supplementary information about low density information. Fourthly, clinical chief complaints and

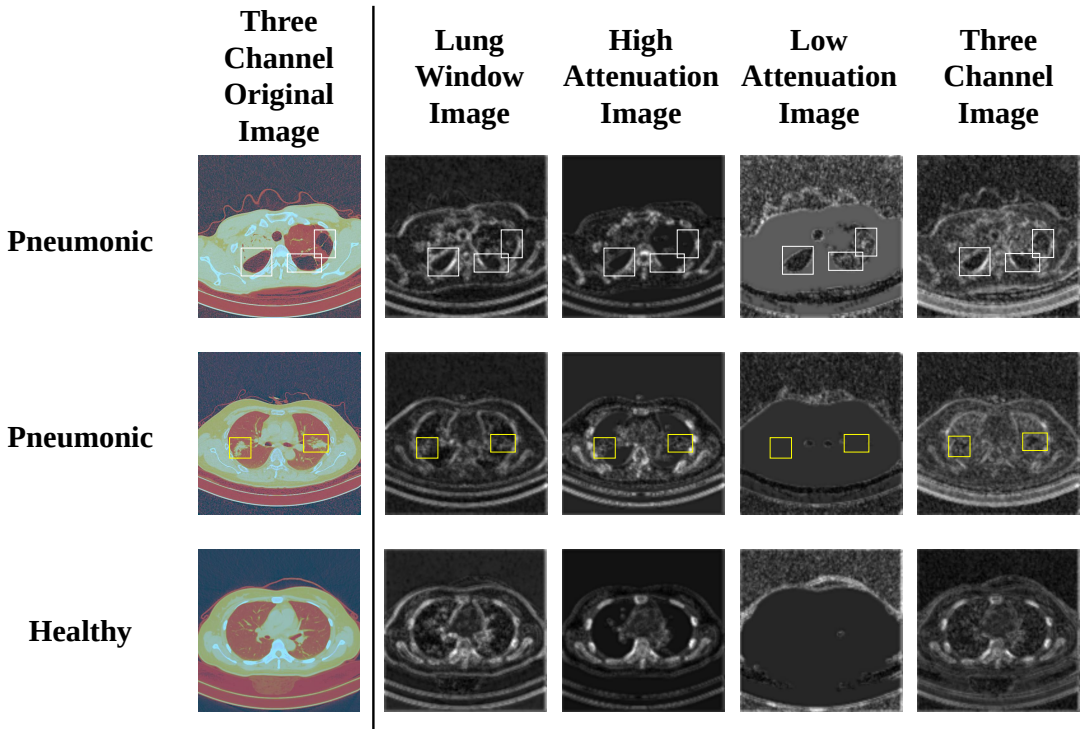


Fig. 5. Convolutional Feature Maps from CNN Models Trained by Different Images. The top row shows a pneumonia case which has abnormally low dense areas (white rectangles). The middle row shows a pneumonia case which has abnormal high dense areas (yellow rectangles). During convolutional process, three-channel images can provide both high dense and low dense information. However, low attenuation images can only provide low dense information, high attenuation images can only provide high dense information. Moreover, lung window images have difficulty in distinguishing low dense tissues form normal lung tissues. The bottom row is a healthy case. We can compare healthy lung tissues with abnormal low dense tissues and abnormal high dense tissues.

TABLE III
PERFORMANCES OF RCNN AND THE PROPOSED FRAMEWORK IN VALIDATING SET

Structure	*Data*	*Accuracy*	*Sensitivity*	*Specificity*	*AUROC*
Wang et al. [7]	Chest X-Ray Image	-	-	-	0.633
Yao et al. [8]	Chest X-Ray Image	-	-	-	0.713
Rajpurkar et al. [9]	Chest X-Ray Image	-	-	-	0.768
Wang et al. [5]	Chest X-Ray Image & Report	-	-	-	0.969
Wang et al. [5]	Chest X-Ray	-	-	-	0.658
Wang et al. [5]	Report	-	-	-	0.947
Baseline	Lung Window Image	0.895	0.921	0.860	0.891
This Work	Three-Channel Image	0.915	0.921	0.907	0.914
This Work	Three-Channel Image & Complaint	0.930	0.904	0.965	0.934
This Work	TC Image & Complaint & Age & Gender	0.930	0.9302	0.9298	0.925

TC indicates Three-Channel Images. Clinical and Demographic Information indicates clinical chief complaints, age and gender.

demographic information further improve the performance of the proposed framework.

The validation loss and accuracy during training is shown in Fig 7. According to this figure, the performances in validation accuracy of three models are almost the same. However, the losses of the framework and the framework & Chief Complaints drop more quickly than RCNN(ResNet), which means clinical and demographic information accelerates the training process.

This figure indicates that these models do not have much difference in the training process, additional information like chief complaints, age and gender pushes the performance forward a little in the decision making process. It is a little different from our assumption. We assumed that clinical and demographic information can improve the speed of training and performances. But the fact is this additional information only works in the final stage: the decision making stage. This pneumonia indicates that image information is still the most important information resource. Demographic and clinical information can only play a supporting role.

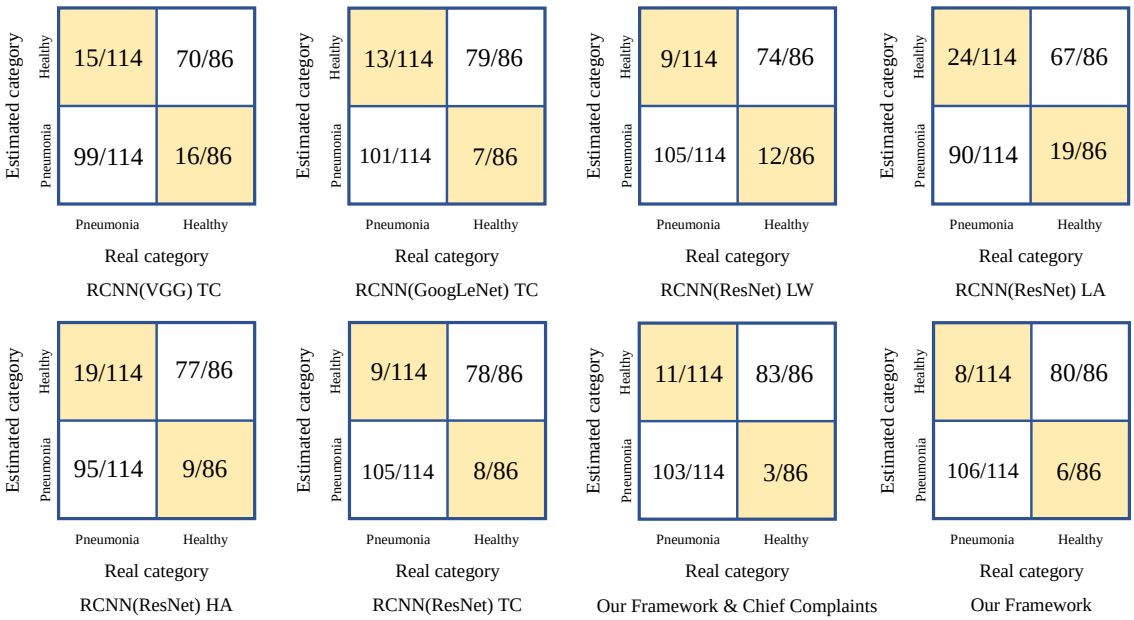


Fig. 6. Confusion matrices achieved by eight different methods in in detection results. We have 144 pneumonia cases and 86 healthy cases in validation set. LW: Lung Window Image, HA: High Attenuation Image, LA: Low Attenuation Image, TC: Three-Channel Image

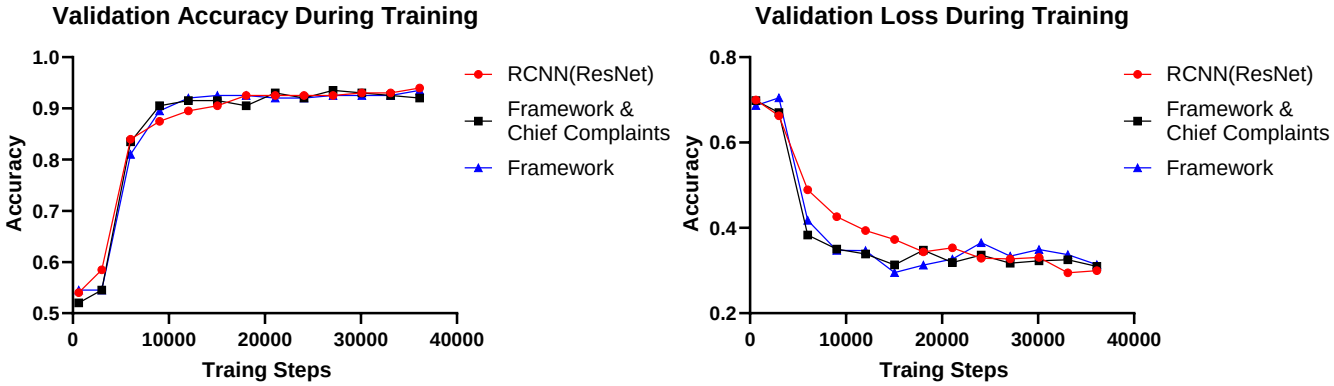


Fig. 7. Validation accuracy and loss during training.

VI. Discussion

A. Investigating the Causes of Improvements

In this section, we demonstrate how the effect of clinical and demographic information imporve the performance of the framework.

To verify that cheif complaints can provide symptoms which are related to pneumonia, we count word frequency about symptoms. Table IV shows that the top 10 keywords in HC (healthy cases) and PC (pneumonia cases) have certain regularity. 'Cough' is the most frequent keyword in both HC and PC. It appears 256 times (46.4%) in PC, 183 times (40.7%) in HC. However, symptoms like 'Expectoration', 'Fever', 'Coughing blood' appear more frequently in PC. For example, 'Coughing blood' appears 47 times in PC, but only appears one time in HC.

According to Table IV, patient who have symptoms like expectoration, repeat condition, shortness of breath have larger chance of having pneumonia. Patients who have chest pain, or feel unconfortable have less chance of having pneumonia. Having cough, on the other hand, is a symptom with minimal discrimination.

Moreover, we count the number of words which can provide information of location. In 1002 cases, every 2 cases contain 1 words which can help to locate the lesions (appear 504 times). Fig 8 shows two examples. According to the location and symptom information provided by chief complaints, we can accurately locate lesions in CT.

In Fig 8, words marked red is information related to location, words marked blue are related to symptoms. In the first case, its chief complaint locates the symptoms in the right lower lung, and then we find shadows in the accurate place. In the second case, it chief complaint says that this patient has pains in right chest, then we also find shadows in the right lung in CT images. This phenomenon demonstrates that information from chief complaints is related to CT images, and can assist deep learning model.

Then we further demonstrate the effect of demographic in-

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

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

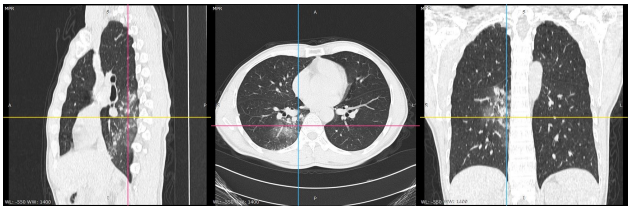
Percentage is frequency divided by number of cases. PC is pneumonia Cases. HC is Healthy Cases.
Key words in bold are both top 10 key words in healthy and pneumonia cases.

formation. As mentioned before, demographic information can provide prior information. According to the Table I mentioned above, we observe some interesting patterns:

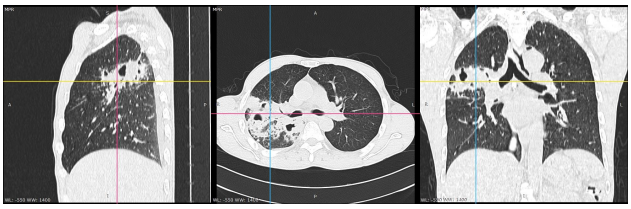
(i) A male patient has a more significant chance of getting pneumonia. In 601 male cases, about 60% of them are pneumonia; however, in 401 female cases, only 47.6% are pneumonia. This phenomenon may be related to smoking since males in Chinese suffer a severe smoking problem;

(2) The table shows that age is also associated with the chance of getting pneumonia. We can observe that people older than 40 have a much larger chance of getting pneumonia. There are about half of healthy cases between 40-50, but this indication drops so quickly that it goes down to 28.8% between 50-60.

According to these two patterns, we can clearly observe that demographic information can provide a strong prior information. These two kinds of demographic information will be treated as confounding factors and help to imporve the performance of the whole framework.



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



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

Fig. 8. Chief complaints can provide information related to CT images. In this figure, we show two pneumonia cases, and each case has chief complaints provided by patients. Words marked red give the location, and words marked blue provide symptoms. English chief complaints are translated from Chinese above. Location and symptoms information provided by chief complaints are related to abnormal tissues in CT images.

B. Limitations and Future Work

Even if our framework can screen pneumonia using multi-modal data, there are still some shortcomings in our work.

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

Secondly, we only consider chest CT scans, chief complaints, gender, and age. In clinical practice, besides the tests mentioned above, patients usually need to take blood pressure measurements, blood tests, heartbeat measurements,

and other tests. These examinations can help doctors gain a more objective and comprehensive understanding of the patient condition so that doctors can make a more accurate diagnose.

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

VII. Conclusions

In this study, we propose a novel mulit-channel multimodal deep regression framework, which combines CT visual features with patients' age, gender, and chief complaints to simulates clinical practice. The proposed framework extracts visual features from three-channel images, semantic features from chief complaints, and fuses these features with prior information provided by age and gender.

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

References

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

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

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

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

[5] 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.

[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] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

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

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

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

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

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3d biomedical segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6393–6400.

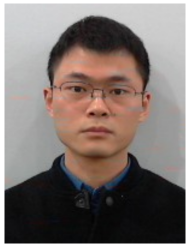
[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *international conference on learning representations*, 2015.

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

[30] 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.

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

[32] Y. Bengio, P. Simard, P. Frasconi *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
[33] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[35] 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.
[36] 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.



Qiuli Wang received the B.E. degree in the School of Information Engineering, Yangzhou University in 2016. He is currently working toward the Ph.D. degree in the School of Big Data & Software Engineering, Chongqing University. His research interests include medical image computing, deep learning, so on.

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

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

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

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