# A Convolutional LSTM Network Approach for Identifying Diseases in Medical Volumetric Images with Limited Annotations


## Abstract

This paper presents a methodology for identifying disease characteristics from medical imaging data using 3D volumes, which have weak annotations. This approach converts 3D volumes into sequences of 2D images. We show the efficacy of our method when detecting emphysema using low-dose CT images taken from lung cancer screenings. Our method uses convolutional long short-term memory (LSTM) to sequentially "scan" through an imaging volume to detect diseases within specific areas. This structure enables effective learning by using just volumetric images and binary disease labels, facilitating training with a large dataset of 6,631 unannotated image volumes from 4,486 patients. When evaluated on a testing set of 2,163 volumes from 2,163 patients, our model detected emphysema with an area under the receiver operating characteristic curve (AUC) of 0.83. This method outperformed both 2D convolutional neural networks (CNN) using different multiple-instance learning techniques (AUC=0.69-0.76) and a 3D CNN (AUC=.77).


## 1 Introduction

This paper addresses the critical challenge of developing deep learning-based computer-aided diagnosis (CAD) systems in radiology, which is often limited by the need for large, annotated medical image datasets. It is particularly difficult to acquire manual annotations from radiologists, which is required to train deep models, especially for 3D imaging techniques like computed tomography (CT). As a result, it is frequently unfeasible to use a model trained using a large, labeled dataset. The detection of emphysema, a disease associated with shortness of breath and an elevated risk of cancer, is one such area. Emphysema is frequently observed as ruptured air sacs within a small portion of the lung volume. The wide range of manifestations in CT scans makes training a model to detect emphysema using solely volumetric imaging data and binary diagnostic labels difficult.

A common strategy to enable learning without precise labels is multiple instance learning (MIL). In MIL, sets of samples are organized into labeled bags, with a positive label indicating the existence of positive samples within the bag. Prior research has effectively used a MIL framework to identify emphysema and other lung disorders on CT scans. It has been demonstrated that MIL, when used with a handcrafted feature-based classifier to analyze a number of 2D patches from the lung, can identify emphysema and other lung diseases. More recently, researchers reported positive results in grading emphysema by summarizing the results of a convolutional neural network (CNN) across a set of 2D patches using a proportional method similar to MIL.

A drawback of MIL-based techniques is their failure to maintain inter-sample relationships. For instance, MIL does not retain the spatial relationship between samples collected from an image, despite being successful in summarizing data from a number of samples. Furthermore, the effectiveness of MIL depends on the pooling strategy used to summarize predictions across the bag, a variable that can greatly affect the instances in which a model succeeds or fails. For example, a maximum pooling-based approach considers only the single sample with the strongest correlation to disease,

.

disregarding any data from the bag's other samples. On the other hand, a mean pooling of predictions within a bag may fail to detect a disease present in only a small number of samples.

Recurrent neural networks, such as long short-term memory (LSTM), are highly adept at identifying correlations between connected samples, such as in pattern recognition across time series data. Convolutional long short term memory (Conv-LSTM) expands this capability to spatial data by applying convolutional operations to an LSTM. Conv-LSTM has been highly successful in identifying changes in image patterns over time, including applications like video classification and gesture recognition. Instead of utilizing Conv-LSTM to identify spatiotemporal patterns from time series image data, we suggest using it to "scan" through an imaging volume for the presence of disease without the need for expert annotations of the diseased regions. Our framework allows for the identification of emphysema-related image patterns on and between slices as it processes the image volume, unlike an MIL-based technique. The network stores emphysema-related image patterns through several bidirectional passes through a volume and produces a final set of characteristics that describe the full volume without the requirement for a possibly reductive bag pooling operation. Our method can make effective use of readily available, but weak, image labels (such as a binary diagnosis of emphysema as positive or negative) for abnormality identification inside image volumes.

## 2 Methodology

### 2.1 Dataset and Processing

A total of 8,794 non-contrast CT volumes from 6,648 unique participants in the National Lung Screening Trial (NLST) were used. We classified 3,807 CT volumes from 2,789 participants who were diagnosed with emphysema during the three years of the study as positive samples, and 4,987 CT volumes from 3,859 participants who were not diagnosed with emphysema in any of the three years as negative samples. 75% of these scans, with a balanced distribution of emphysema-positive and emphysema-negative patients, were utilized for model training. 4,197 volumes from 3,166 patients were used to directly learn model parameters, while 2,434 volumes from 1,319 patients were used to fine-tune hyper-parameters and assess performance in order to select the best-performing model. The remaining 2,163 volumes (578 emphysema positive, 1,585 emphysema negative), each from a unique patient, were held out for independent testing. Volumes were resized to 128x128x35, which corresponds to an average slice spacing of 9 mm.

### 2.2 Convolutional Long Short Term Memory (LSTM)

The architecture includes four units, each consisting of convolution operations applied to each slice individually and a conv-LSTM to process the volume slice by slice. Two 3x3 convolutional layers with batch normalization are followed by max-pooling. The output of the convolutional layers for each slice is then processed sequentially by the conv-LSTM layer in either forward or reverse order. This outputs a set of features collected through convolutional operations using both the current slice and previous slices within the volume. All layers within a unit have the same number of filters and process the volume in either ascending or descending order. The four convolutional units have the following dimensionality and directionality: Ascending 1: 32 filters, Descending 1: 32 filters, Ascending 2: 64 filters, Descending 2: 64 filters. The final Conv-LSTM layer produces a single set of features that summarizes the network's results after processing the full imaging volume multiple times. Finally, a fully-connected layer with sigmoid activation calculates the probability of emphysema. The network, as illustrated in Figure 1, contains a total of 901,000 parameters. All models were trained for 50 epochs or until validation set performance stopped improving.

### 2.3 Comparison Experiments

Multiple Instance Learning: We developed an MIL-based network in which each slice of the CT volume was treated as a sample from a bag. We implemented a solely convolutional network design similar to the one shown in Figure 1, but with more single-slice convolutional layers instead of conv-LSTM layers, to achieve this. Various methods for summarizing predictions across the entire volume into a single bag probability were investigated. The following methods can be used to compute the overall probability, P, for a bag containing N samples with an individual probability of emphysema, pi, i  1, ..., N:

1. Max Pooling: $P = \max(p_i)$
2. Mean Pooling: $P = \frac{1}{N} \sum_{i=1}^{N} p_i$
3. Product Pooling: $P = 1 - \prod_{i=1}^{N} (1 - p_i)$

3D CNN: Conv-LSTM was also compared to a 3D CNN with a similar structure to the 2D CNN used with MIL, with the exception of a single dense layer and no pooling action on the final convolutional layer. The number of kernels for each comparison model was raised to make its number of parameters roughly comparable to that of our Conv-LSTM framework and ensure a fair comparison (Table 1).

## 3 Results

Convolutional-LSTM demonstrated high accuracy in the detection of emphysema when trained using only weakly annotated imaging volumes, achieving an AUC of 0.82. It outperformed a CNN with MIL, regardless of the pooling strategy (Max pooling: AUC=0.69, Mean Pooling: AUC=0.70, Product pooling: AUC=0.76). At the optimal operating point corresponding to the Youden Index, our model achieved a sensitivity of 0.77 and a specificity of 0.74. The results for all evaluated models in the testing set are shown in Table 1.

| Model | Kernels | # Parameters | AUC | Sensitivity | Specificity | F1 |
|---|---|---|---|---|---|---|
| MIL - Max Pooling | 64 | 1,011,393 | 0.69 | 0.59 | 0.68 | 0.63 |
| MIL - Mean Pooling | 64 | 1,011,393 | 0.70 | 0.76 | 0.57 | 0.66 |
| MIL - Product Pooling | 64 | 1,011,393 | 0.76 | 0.61 | 0.79 | 0.69 |
| 3D CNN | 36 | 958,213 | 0.77 | 0.61 | 0.80 | 0.69 |
| Conv-LSTM | 32 | 901,793 | 0.83 | 0.77 | 0.74 | 0.75 |

Table 1: Emphysema detection results in the testing set (2,219 CT volumes) and model size.

Our method eliminates the need for manual processing or time-consuming annotation of imaging data. Our framework makes it possible to train for disease detection using simple binary diagnostic labels, even when the disease is confined to a small area of the image. As a result, our network can be trained easily using information that can be gathered automatically by mining radiology reports. This significantly increases the amount of volumetric imaging data that can be used for this kind of application and enables easy retraining and fine-tuning of an algorithm when used in a different hospital. This strategy can be used in other disease/abnormality detection problems outside of emphysema when the amount of volumetric imaging data accessible is greater than the capacity of radiologists to offer manually drawn ground truth, but when labels may be readily retrieved from radiology reports.