

Improving Lung Cancer Detection

Abhishek Mishra -amishr20@uic.edu, Nikita Jaiswal- njaisw2@uic.edu,
Rashmi Choudhary- rchoud9@uic.edu

1 Problem description

Lung cancer is the leading cause of cancer-related death worldwide. Early detection is critical to give patients the best chance at recovery and survival.

It has been founded through research that lung cancer screening using annual low-dose computed tomography (CT) helps in early detection of lung cancer and hence plays a critical role. The objective is to design and develop computer-aided detection systems for optimized CT lung cancer screening that can be implemented on a large scale. We need to develop algorithms that accurately determine when lesions in the lungs are cancerous thereby reducing false positive rate that plagues the current detection technology.

We will use a data set of thousands of high-resolution lung scans provided by the National Cancer Institute to develop systems which help predicting whether a CT scan is of a patient who either has or will develop cancer within the next one year of the date the scan was taken. This will get patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

2 Data description

1. Low-dose CT images from high-risk patients in DICOM format.

A CT image contains a series with multiple axial slices of the chest cavity. Each 3D CT Scan consists of variable number of 2D slices, and this number depends on the resolution of the scanner. Each slice has an Instance Number associated with it which tells the index of the slice from the top. All the DICOM files for a CT Scan are inside one folder having the CT Scan's name/ID.

2. Labels of CT images containing the ground truth/class information.

This file contains the cancer ground truth for the training set images. This file includes patient id and the class information (0/1), where 1 represents the class of patients diagnosed with cancer. The ground truth labels were confirmed by pathology diagnosis.

Additional data set from Lung Nodule Analysis 2016 (LUNA2016) challenge:

A vital first step in the analysis of lung cancer screening CT scans is the detection of pulmonary nodules, which may or may not represent early stage lung cancer. We train a network to segment out potentially cancerous nodules and then use the characteristics of that segmentation to make predictions about the lung cancer diagnosis of the scanned patient within one year time frame. LUNA2016 challenge provides the relevant data for the segmentation and training purposes.

1. CT images in (MetaImage format) taken from LIDC-IDRI database

The CT images are stored in MetaImage (mhd/raw) format. Each .mhd file is stored with a separate .raw binary file for the pixeldata.




2. Annotations for CT images

This file contains one finding per line. It has 3D co-ordinates and the diameter of nodules located in the CT images. This is used as a reference/training data labels for the nodule detection/segmentation task. The annotated nodules can be a cancerous or non-cancerous.

3. Nodules location and ground truth/class for the CT images.

This file contains information of one nodule candidate per line. It has the scan name, 3D co-ordinates of each candidate located in the CT images with the ground truth label stating it to be a cancerous/non-cancerous candidate. There can be multiple candidates for the same nodule. This is used as a reference/training data labels for the nodule classification task.

CT Scan image details for each patient	Ground truth information about patients												
Patient '0a38e7597ca26f9374f8ea2770ba870d' has 110 scans Patient '0d19f1c627df49eb223771c28548350e' has 183 scans Patient '0ddeb08e9c9722785342bd71a2a695e' has 171 scans Patient '0a0c32c9e08cc2ea76a71649de56be6d' has 133 scans Patient '0b20184e0cd497028bdd155d9fb42dc9' has 196 scans Patient '0a099f2549429d29b32f349e95fb2244' has 128 scans Patient '0acbebb8d463b4b9ca88cf38431aac69' has 203 scans Patient '0bd0e3056cbf23a1cb7f0f0b18446068' has 280 scans Patient '0c98fcb55e3f36d0c2b6507f62f4c5f1' has 180 scans Patient '0c37613214faddf8701ca41e6d43f56e' has 164 scans Patient '0c59313f52304e25d5a7dcf9877633b1' has 244 scans Patient '0ca943d821204ceb089510f836a367fd' has 147 scans Patient '0c0de3749d4fe1757a5098b060982a1' has 123 scans Patient '0d06d764d3c07572074d468b4cf954f' has 435 scans Patient '0c60f4b87afcb3e2dfa65abbbf3ef2f9' has 136 scans Patient '0de72529c30fe642bc60dcb75c87f6bd' has 113 scans Patient '0d941a3ad6c889ac451caf89c46cb92a' has 177 scans Patient '00cba091fa4ad62cc3200a657aeb957e' has 134 scans Patient '0d2fcf787026fece4e57be167d079383' has 126 scans Patient '0c9d8314f9c69840e25febabb1229fa4' has 221 scans ---- Total patients 20 Total DCM files 3604	<table> <tr> <th>id</th><th>cancer</th></tr> <tr> <td>0015ceb851d7251b8f399e39779d1e7d</td><td>1</td></tr> <tr> <td>0030a160d58723ff36d73f41b170ec21</td><td>0</td></tr> <tr> <td>003f41c78e6acfa92430a057ac0b306e</td><td>0</td></tr> <tr> <td>006b96310a37b36cccb2ab48d10b49a3</td><td>1</td></tr> <tr> <td>008464bb8521d09a42985dd8add3d0d2</td><td>1</td></tr> </table>	id	cancer	0015ceb851d7251b8f399e39779d1e7d	1	0030a160d58723ff36d73f41b170ec21	0	003f41c78e6acfa92430a057ac0b306e	0	006b96310a37b36cccb2ab48d10b49a3	1	008464bb8521d09a42985dd8add3d0d2	1
id	cancer												
0015ceb851d7251b8f399e39779d1e7d	1												
0030a160d58723ff36d73f41b170ec21	0												
003f41c78e6acfa92430a057ac0b306e	0												
006b96310a37b36cccb2ab48d10b49a3	1												
008464bb8521d09a42985dd8add3d0d2	1												

Sample file listings for CT Scans		
	1.3.6.1.4.1.14519.5.2.1.6279.6001.105756658031515062000744821260	RAW File
	1.3.6.1.4.1.14519.5.2.1.6279.6001.105756658031515062000744821260.mhd	MHD File
	1.3.6.1.4.1.14519.5.2.1.6279.6001.108197895896446896160048741492	RAW File
	1.3.6.1.4.1.14519.5.2.1.6279.6001.108197895896446896160048741492.mhd	MHD File
	1.3.6.1.4.1.14519.5.2.1.6279.6001.10900252552452225658609808059	RAW File
	1.3.6.1.4.1.14519.5.2.1.6279.6001.10900252552452225658609808059.mhd	MHD File
	1.3.6.1.4.1.14519.5.2.1.6279.6001.111172165674661221381920536987	RAW File
	1.3.6.1.4.1.14519.5.2.1.6279.6001.111172165674661221381920536987.mhd	MHD File

3 Proposed model, algorithm and techniques

The images in the data set are of sizes (z,512,512) where z is the number of slices in the CT scan and varies depending on the resolution of the scanner. These images cannot be fed directly into CNN because of the limit of computational power. Hence, we will find regions that have more chances of being cancerous.

We will start with reading the data and visualizing the same. After that, we will segment the lungs and find the nodules using image processing methods. We will use the LUNA dataset(annotations.csv) to generate an appropriate training set for U-net. We will use the nodule locations given in annotations.csv and extract slices that contain the largest nodule from each patient scan. Mask will be created for those slices based on nodule dimension. The nodule locations are given in terms of millimeters relative to a coordinate system defined by the CT scanner. The image data is given as a varying length stack of

512 X 512 arrays. In order to translate the voxel position to the world coordinate system, one needs to know the real world position of the [0,0,0] voxel and the voxel spacing in mm.

The next step is to isolate the lungs in the images. We'll need to import some skimage processing modules for this step. The general strategy is to threshold the image to isolate the regions within the image, and then to identify which of those regions are the lungs. These images and the correspondingly trimmed and rescaled masks are randomized and sent to a single file that contains a numpy array of dimension [num_images,1,512,512]. The 1 is important as the U-net is enabled for multiple channels.

We can call `LUNA_train_unet` from the command line which will automatically attempt to load a `unet.hdf5` file from the current directory and train the model according the parameters. Further it will also make predictions on the test data.

Then we will preprocess the LUNA 16 dataset for training architectures for candidate classification. The nodules can be used for classification by cutting 3D voxels around them and passing it through 3D CNN. This 3D CNN will be trained on LUNA 16 dataset.

We will import useful analytics libraries to perform different data transformation and processing in python.

- Numpy : Fundamental package for scientific computing in Python
- Scikit.image : Collection for algorithm for image processing such as performing segmentation or morphology.
- Matplotlib : Used for visualization
- Pydicom : Package for working with DICOM files and manipulating it into easy structures.
- Keras : High-level neural networks library, written in Python and capable of running on top of either TensorFlow or Theano

4 Possible Experiments and metrics of success

The segmentation of lung structures and identifying nodules is a very challenging problem because homogeneity is not present in the lung region, similar densities in the pulmonary structures, different scanners and scanning protocols. We will use different image processing techniques to decrease the loss function for nodules segmentation.

Image Processing Techniques

- Convert the image into binary image
- Remove the objects associated with the border of the images
- Label the image to do segmentation of lungs
- Keep the two largest labels

- Perform Erosion to remove the lung nodules from blood vessels
- Perform closure to close small holes inside the foreground object
- Superimpose the binary mask on the image

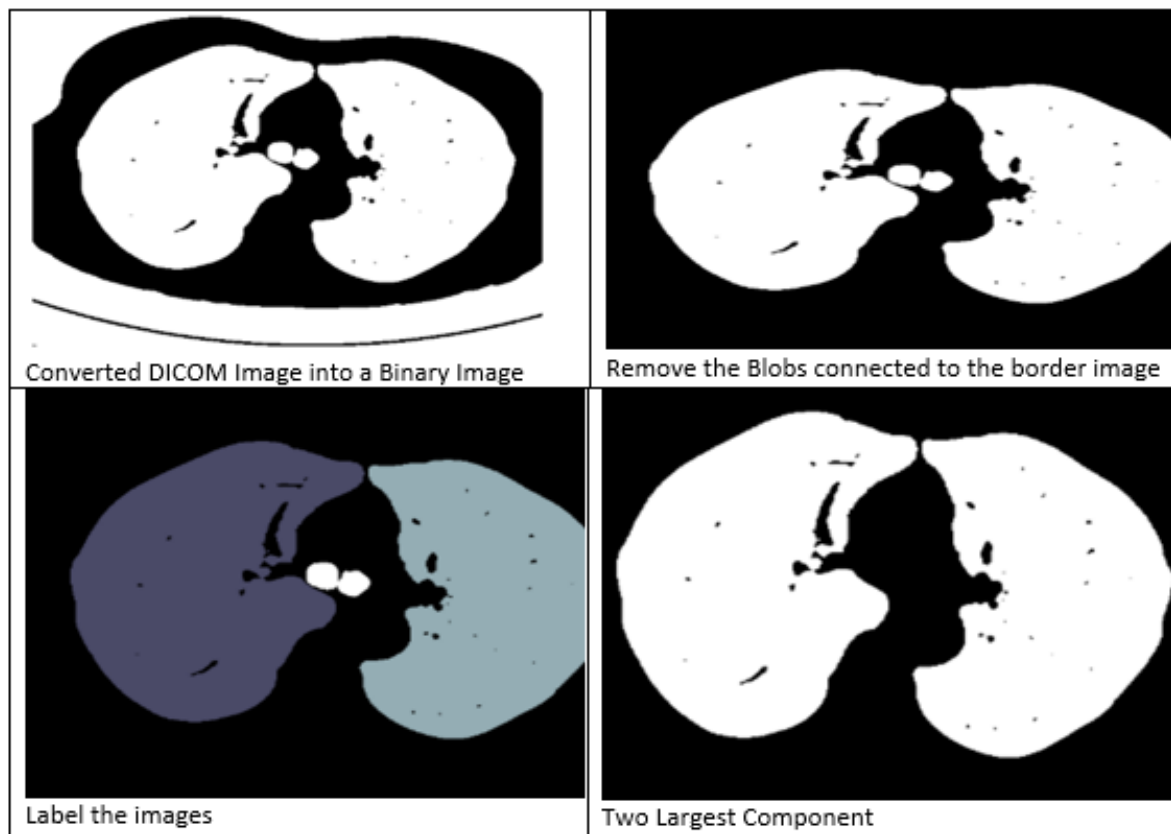
We will perform data augmentation to get more training data. Dropout will be the important part of model to decrease overfitting. Apart from trying different convolution networks, we will try to include features such as average size, morphology and position within the image for model building purpose.

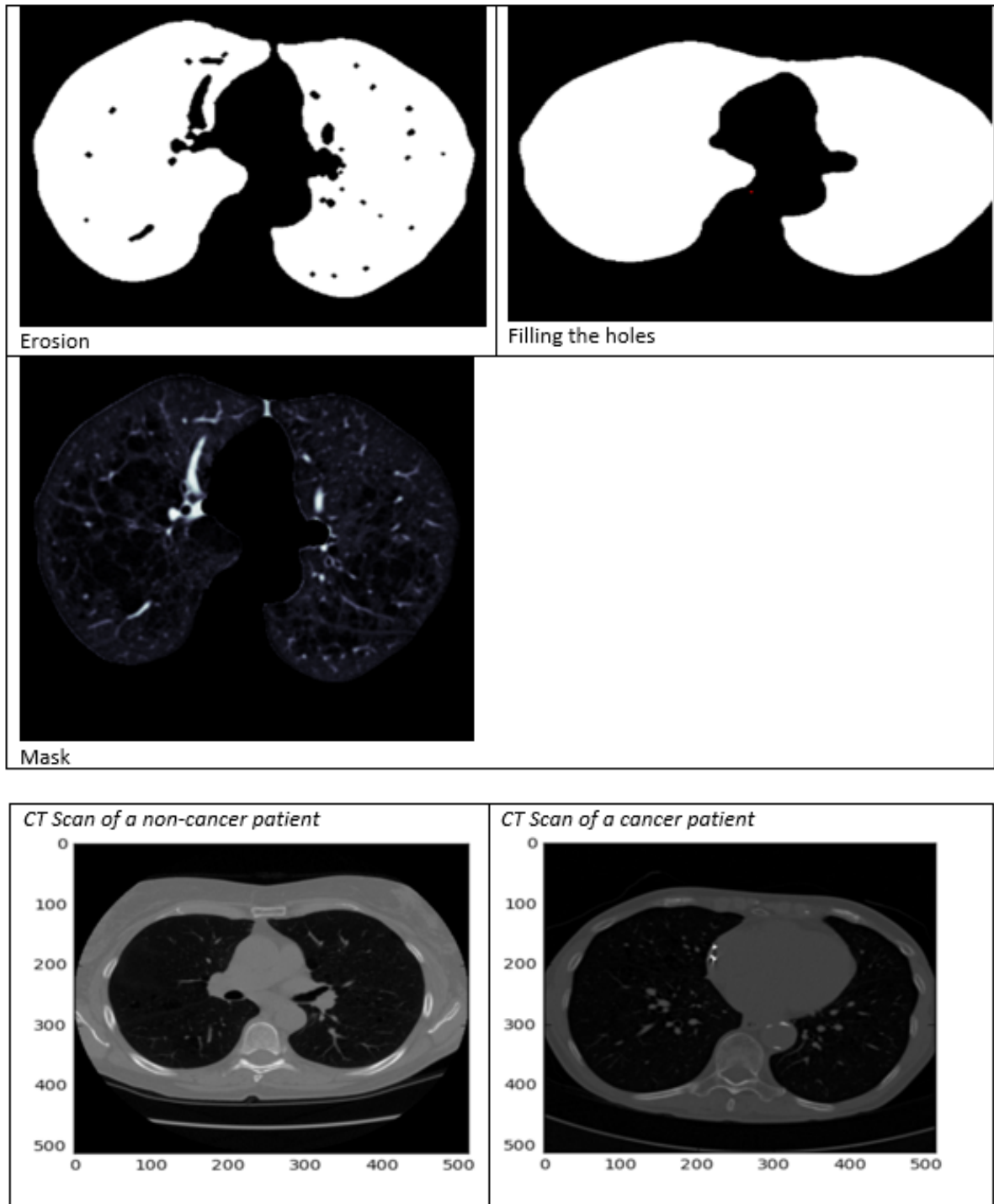
4.1 Results obtained so far

We started visualizing the sample data set and each CT scan consists of multiple slices in a DICOM format. After reading the image file, we will update the intensity values of -2000 with 0 because they are the pixels that fall outside of the scanner bounds.

We plotted image of a patient with cancer and another image without cancer:

We performed some image processing techniques on a single file of a cancer patient.





5 Possible Conclusions

After training the models for first segmentation of nodules and then classification of those nodules for lung cancer diagnosis, we generate a score corresponding to each CT scan. This score represents the probability of that patient to either have or will develop cancer within the next one year of the date the scan was taken. The final output will be like

follows:

```
id, cancer
01e349d34c02410e1da273add27be25c, 0.5
05a20caf6ab6df4643644c923f06a5eb, 0.5
0d12f1c627df49eb223771c28548350e, 0.5
...
```

6 References

1. Data Science Bowl 2017 (Can you improve lung cancer detection?)
<https://kaggle.com/c/data-science-bowl-2017>
2. LUNA16 Challenge (LUng Nodule Analysis 2016)
<https://www.grand-challenge.org/site/luna16/home>
3. <http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/>
4. http://scikit-image.org/docs/dev/auto_examples/xx_applications/plot_morphology.html
5. http://cs231n.stanford.edu/reports2016/313_Report.pdf
6. <https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-41>