# Capstone Project Proposal Lung CT Cancer Detection

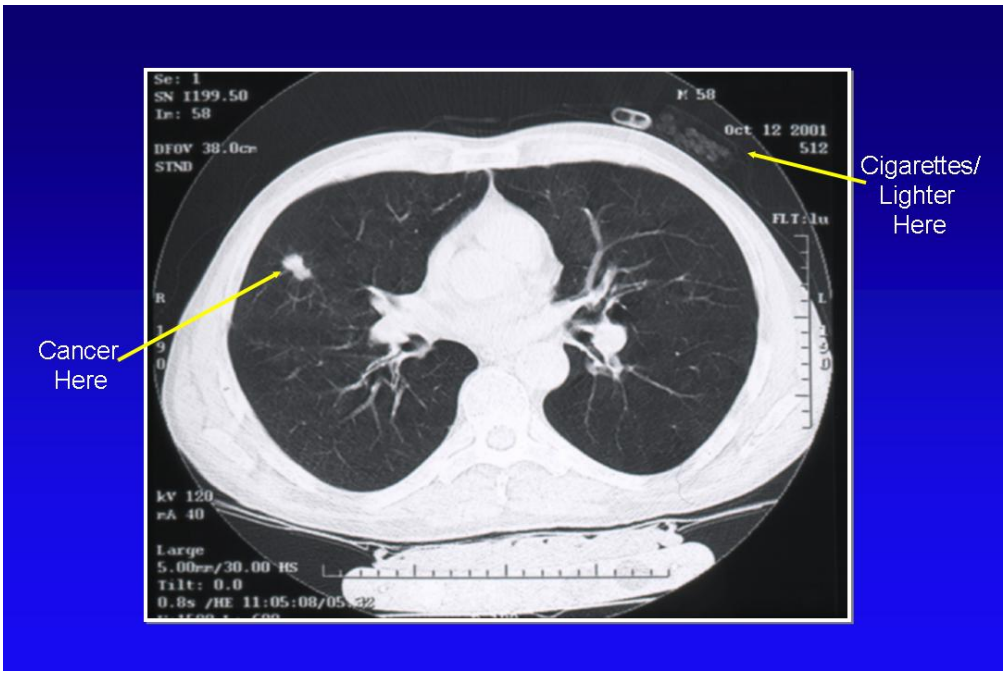
August 15, 2017

Mike Huang

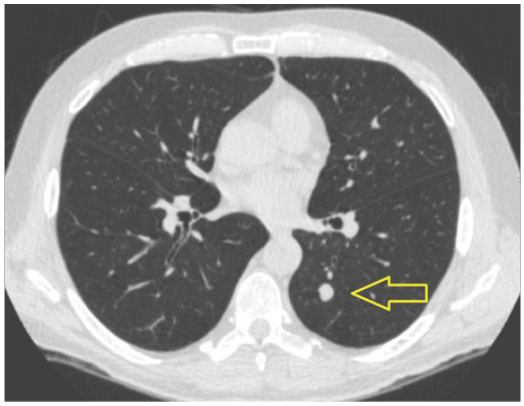
## 1 Domain Background

Lung cancer is the second most common cancer in both men and women that afflicts 225,500 people a year in the United States. About 1 out of 4 cancer deaths are from lung cancer, more than colon, breast, and prostate cancers combined [1]. Early detection of the cancer can allow for early treatment which significantly increases the chances of survival [2].

Lung cancer screening is performed with a CT scan that collects hundreds of images to build a full 3D image of the lung. Next, small growths called pulmonary nodules need to be detected. These nodules show up as small, circular structures on the CT scans.




In some cases, the nodules are not obvious and may take a trained eye and considerable amount of time to detect. ***Building a machine learning algorithm that can automatically detect the nodules can save considerable time and money.***



Unclear Nodules

Additionally, most pulmonary nodules are not cancerous as they can also be due to non-cancerous growths, scar tissue, or infections [1]. The task is then to determine the features of a nodule that are associated with malignancy. Current state-of-the-art methods yield a 25% false positive rate in CT lung cancer screenings [4]. ***A convolutional neural network may be used to determine the features associated with cancerous or non-cancerous pulmonary nodules, and may reduce the false positive rate of CT lung cancer screenings.***

## 2 Problem Statement

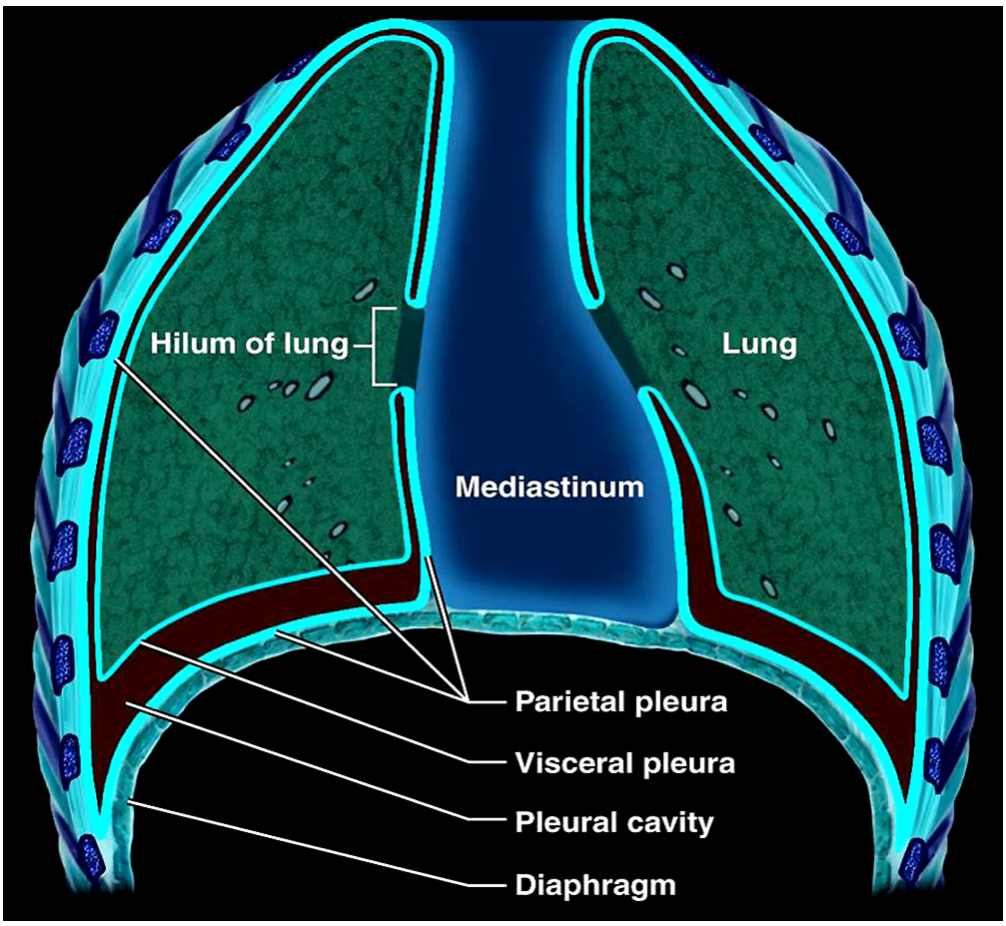
Use a convolutional neural network to detect the probability lesions in the lungs are cancerous with an automated algorithm. The neural network should be trained to maximize that probability when it is and minimize it when it doesn't. This optimization function can be defined by the log loss.

- **Task:** Analyze CT scan images of lungs to find nodules which are defined as a discrete, well-marginated, rounded opacity less than or equal to 3cm in diameter that is completely surround by lung parenchyma, does not touch the hilum or mediastinum [3]. Then create a classifier to calculate the probability a nodule is cancerous.

- **Training set:** 3D CT scan slice images with a label for each patient as cancer or non-cancer. 1595 patient samples.
- **Performance:** Maximize the probability that the cancer is present when it is and minimize the probability when it isn't present
- **Target function:** Log Loss:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})],$$

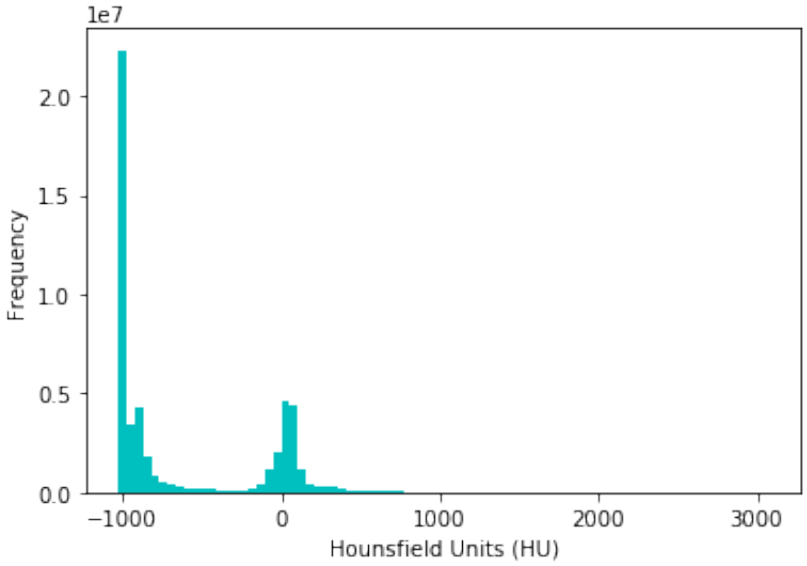
where
- $n$ is the number of patients in the test set
- $\hat{y_i}$ is the predicted probability of the image belonging to a patient with cancer
- $y_i$ is 1 if the diagnosis is cancer, 0 otherwise
- $log()$ is the natural (base e) logarithm


Hilum of lung
Lung
Mediastinum
Parietal pleura
Visceral pleura
Pleural cavity
Diaphragm

### 2.0.1 Datasets and Inputs

The dataset includes CT scan images of 1,595 patients collected from the Kaggle Data Science Bowl 2017 [5]. For each patient, there are 100-150 CT scan slices, forming a 3D composite of the lung. The samples are labeled as either cancerous or non-cancerous.

CT scan works by using x-rays to image structures deep within the tissue with high spatial resolution. The x-rays can also detect the type of substance it transmits and is quantitatively expressed as a unit of radiodensity, called the Hounsfield Unit (HU). For example, air has very high transmittance and has a radiodensity of -1000HU. Soft tissue has a radio density of 100-300HU. Bone has a radiodensity between 700-3000HU. Most of the space within the lung consists of air. Nodules consist of spherical tissues within the lung with a radiodensity between 100-300HU.




HU distribution of a lung scan

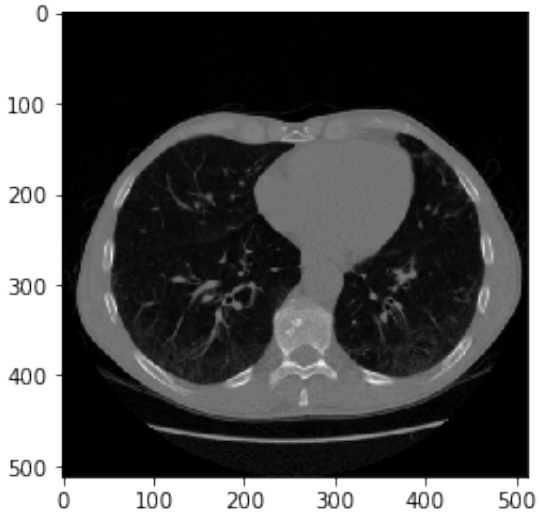
## 3 Solution Statement

A deep-learning algorithm with a convolutional neural network can first detect structures that resemble nodules. This can be done by using pre-defined convolutional filters that approximate the shape of nodules. These filters can be created by averaging the shape of X amount of known nodules. These convolutional filters can be then scanned throughout the entire lung to determine the positions likely for the shape to be present.

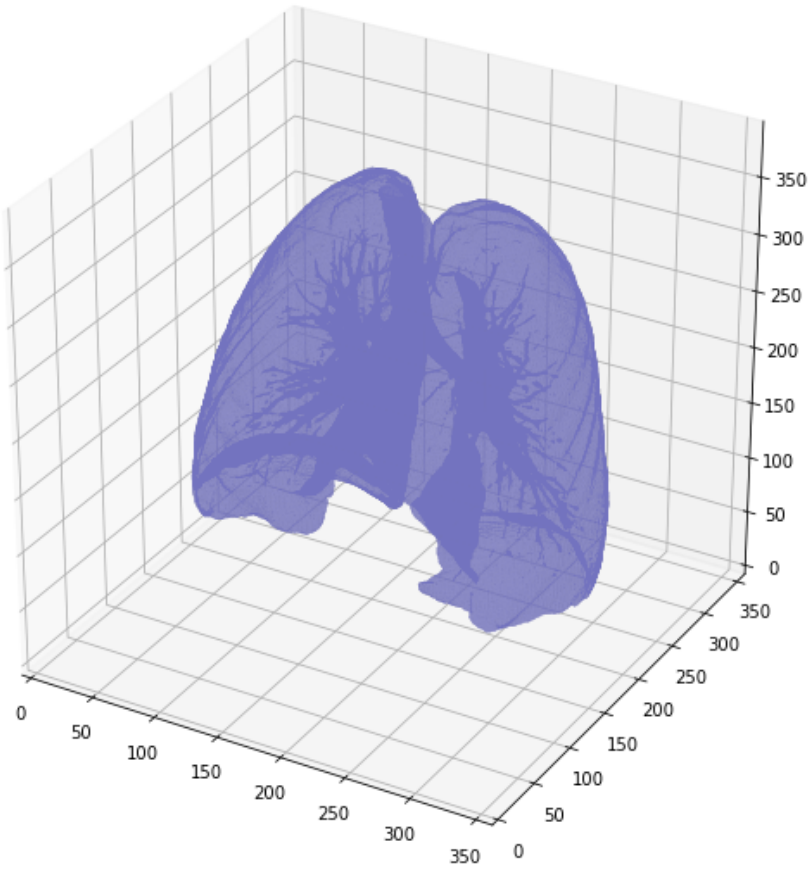
Next, the nodules need to be determined if they are cancerous or not. A convolutional neural network may be used to determine the shape features associated with cancerous or non-cancerous by training the convolutional network.

## 4 Benchmark Model

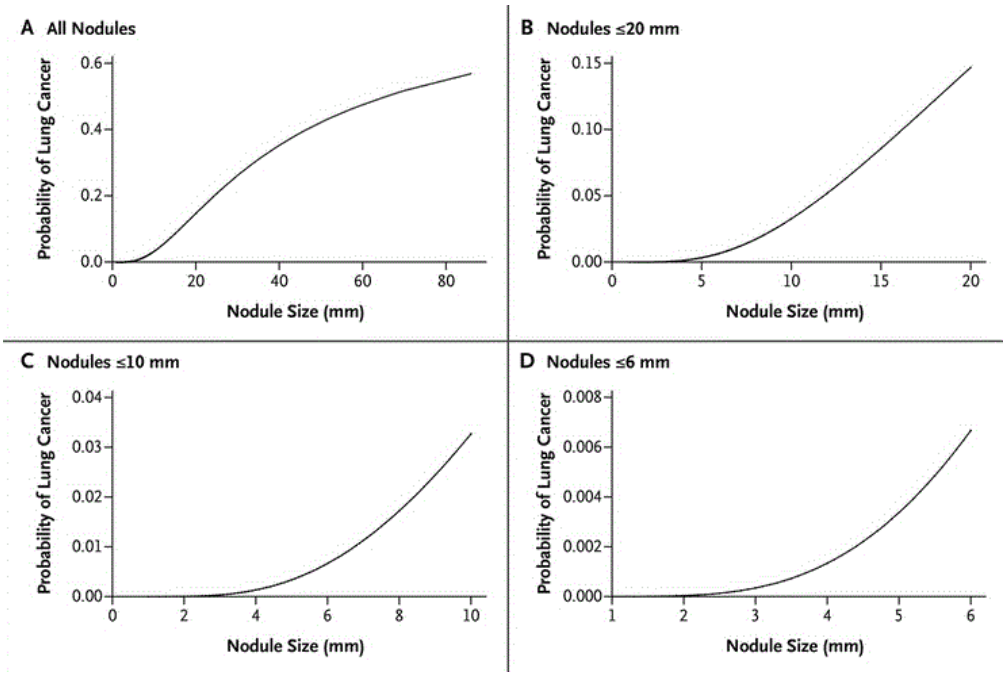
A previous study found nodule size to be the greatest predictor of lung cancer [6]. The probability of lung cancer with respect to nodule size is plotted on the graph below.




CT image of one lung slice




Composite 3D rendering of collection of lung slices for a single patient sample




By digitizing the plot and running a curve fit, I arrive at the following 6th order polynomial function.

$$CancerProbability = 6\text{e}{-12}x^6 - 2\text{e}{-09}x^5 + 4\text{e}{-07}x^4 - 3\text{e}{-05}x^3 + 0.0011x^2 - 0.0051x + 0.0051$$

$x = $*Nodule Size(mm)*

For each sample, the size of the nodule will be inputted into the function to return the probability. This probability can be inputted into the **Log Loss function** defined in the Problem Statement.

## 5 Evaluation Metrics

This model as well as the benchmark models all use the **Log Loss function** defined in the Problem Statement as the evaluation criteria. This function will maximize the probability that the cancer is present when it is and minimize the probability when it isn't present.

## 6 Project Design

**Process data**

The dataset consists of 66GB of Dicom files, the standard for medical images. These need to be processed to its ndarrays to be inputted into tensorflow. The following tasks for processing need to be performed:

```
1. Convert dicom to ndarray
2. Resample image so a pixel represents a fixed unit length of the lung, eg: 1px=1m
3. Mask the lung to minimize the search space of the convolutional filters. This ca
4. Regularize the data by centering it to zero and dividing by the mean.
```

**Defining nodule filter:** 1. Take 10 samples labeled as cancerous and find the nodule by manually extracting it by eye. Overlap the nodules and normalize their radii. Average them. Then create N number of convolutional filters with varying sizes.

**Extracting the nodules from samples:** 1. Input processed data into a 1-layer convolutional network and run through all the pre-defined nodule filters. Use either a softmax or a ReLU function as the activation function. For positions that cross X threshold of the activation function, capture the input at that position and output it to an isolated ndarray.

**Classifying nodules as cancerous vs non-cancerous** 1. Input nodules into a convolutional neural network with two to three layers and an output layer with two neurons, cancerous and non-cancerous. The weights of the network can be trained with gradient descent or adam optimizer. Ideally, the neural network will determine the features in the shapes that are associated with cancerous and non-cancerous.

# 7 References


1. https://www.cancer.org/cancer/lung-cancer/prevention-and-early-detection/exams-and-tests.html
2. http://www.mayoclinic.org/diseases-conditions/lung-cancer/basics/tests-diagnosis/con-20025531
3. http://www.radiologyassistant.nl/en/p460f9fcd50637/solitary-pulmonary-nodule-benign-versus-malignant.html
4. https://biometry.nci.nih.gov/cdas/approved-projects/531/
5. https://www.kaggle.com/c/data-science-bowl-2017
6. http://www.nejm.org/doi/full/10.1056/NEJMoa1214726#t=article