LCDetect: A Lung Cancer Prediction and Detection System using Nodule Based Methods and Machine Learning Algorithms

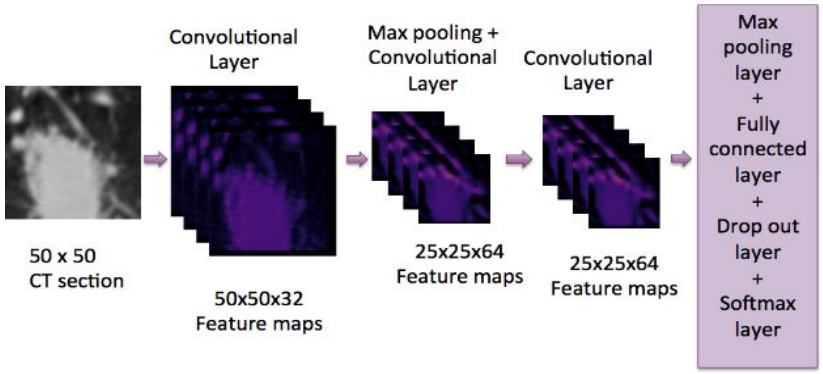
As of 2018, Lung Cancer has been the leading cause of cancer related deaths in the world. The need for human intervention to analyze CT scans and perform tests is a problem and must be addressed. Especially in underdeveloped countries who do not have wide access to diagnostic equipment and advanced devices. Therefore, the need exists to develop an automated and accurate system to detect lung cancer using nodule-based localization using machine learning algorithms to increase the efficiency. Ubiquity of smartphones and laptops with advanced software make them the best candidates to cater to such needs. Current systems that are used do not provide all the necessary information to complete a diagnosis. It includes an initial CT scan, a radiologist report, and if cancer is suspected, then a PET scan and biopsy necessary to be able to determine the type and stage of cancer. However, using machine learning and image processing techniques, a smarter diagnostic tool can be used to increase the number of early detection of cancer and become an assistant tool to the radiologist.

This led me to the realization that the best way to solve this growing epidemic was to develop an accurate and portable algorithm to be able to detect all types and stages of lung cancer. In high level, the system I created is capable of detecting stages of adenocarcinoma, squamous cell carcinoma, small cell lung cancer, and non small cell lung cancer and provide a percentage prediction by analyzing solitary pulmonary nodules. This is a three part algorithm that consists of **data preprocessing, image detection, and a Convolutional Neural Network (CNN)**. My model is trained to segment and locate potential nodules and differentiate them as malign and benign based on **features such as size, contrast, lobulation, calcification, sphericity, spiculation, and distance.** My entire end to end design was developed in Python and after going through several iterations and changes, I performed user testing on my system with 50,000 test datasets and achieved a 97% success rate thus achieving my engineering goal and solving the original problem.

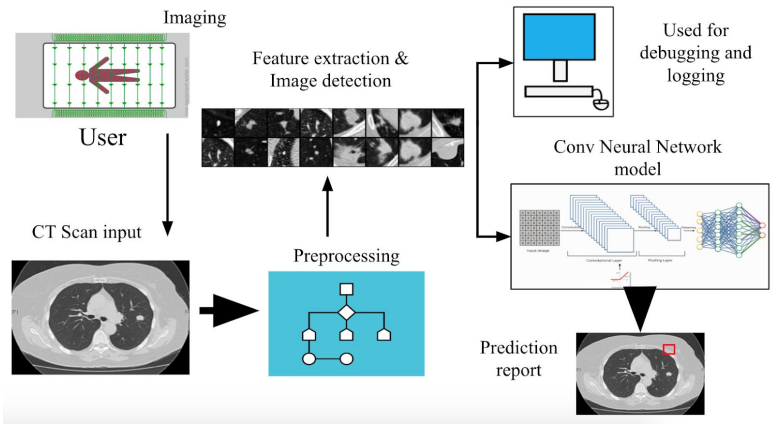
The **first** part in my algorithm was the data preprocessing. I began by collecting datasets from the Lung Image Database Consortium (LIDC) to compile 50,000 datasets and then performed a 80-20 split of test:train. To clean the data, I performed a noise removal method that reduced the spatial size of the CT scan and also normalize the data by converting the RGB pixels to grayscale colors. To train my model, I was able to compile annotations in .xml files that provided coordinates of the nodules. Then I developed a segmentation script that allowed to reduce the spatial size of the original CT scan to just the regions of interest that were used for training the CNN model. To increase the amount of training data, I performed a data augmentation script that allowed me to generate more data to work with. From here, the model

was trained to differentiate between malignant and benign nodule based on the feature analysis that is conducted. This model was then optimized with the training data using methods such as Adam regression and backpropagation.

Second, I developed an image detection module that was able to segment nodules automatically when new data was given. This detection phase was able to localize the scan as a lung by detecting the lobes of the lung and lymph nodes to be able to determine the type of lung cancer. If the nodules are found in the outer lining, it is detected as adenocarcinoma, if in the inner lining, then it is squamous cell. Based on the distance from the lymph nodes of the lungs, the stage of the cancer is also determined through the percentage prediction. This location based analysis is then based on the CNN model to be able to determine the malignancy and perform the end classification.

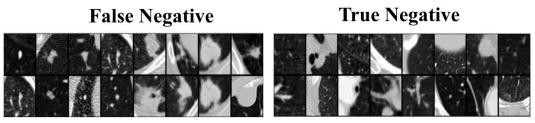


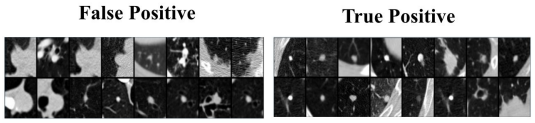
The **third** part of the algorithm is the CNN model. This consisted of layers such as the pooling, convolution, dropout, and softmax layer. The cleaned CT scan with regions of interest is passed to this model to classify for the type of cancer and stage and present the final diagnosis report. This model works by extracting the relevant features to the scan and weights each one with a significance value. This weights were altered during the optimization phase through the back propagation method.

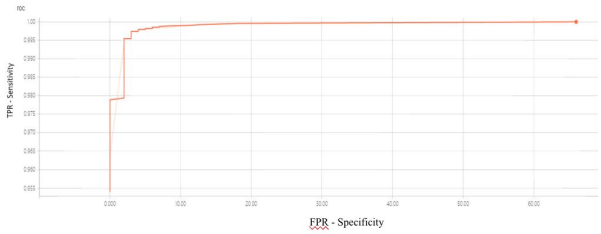


After the algorithm was performed, I performed a **testing** procedure with the initial testing datasets that I had. From here, I optimized my algorithm by altering the feature weights, the regression model that was used, and apply a new image filter for data preprocessing. Then I performed another validation test using sample patient CT scans from South Sound Radiology to achieve 98% success rate.

To improve the performance of the algorithm, I performed a **statistical analysis** of my algorithm and it's data output. I calculated the sensitivity, specificity, accuracy, train loss, confusion matrix, and graphed the ROC curve to evaluate the algorithm. The False positive, True positive, True negative, and False negative rates were used in the graph curve to achieve an ideal of close to 1 and horizontal to the right. The values of these statistics are what guided me to perform the necessary optimization techniques.







The entire end to end algorithm was developed in PyCharm using Python and a TensorFlow backend. I titled my system LCDetect and am continuing to enhance it by applying a generative algorithm to be able to predict how a future CT scan and show the progression of the cancer for radiologists to get a better understanding of the time vs growth relationship. I am also working with a professor from the University of Washington to publish my Research Paper on my algorithm.

Overall, LCDetect is an end to end three part algorithm capable of detecting 4 types of lung cancer and the stage in an accurate and portable way. The algorithm I created revolutionizes diagnostic procedures by allowing radiolists to simply input the CT scan and avoid the necessary follow up PET, CT, and biopsies. I have deployed this algorithm as an Azure Web Service and have tested this side by side with doctors and shown how it is able to pick up more significant nodules than the human eye and classify the lung cancer. LCDetect is capable of detecting adenocarcinoma, squamous cell, small cell lung cancer, and non-small cell lung cancer using an image detection module and machine learning algorithm.