**A Portable Machine-Learning Based Detection System of Prevalent Chronic Respiratory Illnesses and Lung Cancer**

Sathvik Nallamalli

**Section 1: Project Background**

Growing up and living in my hometown of Olympia, not a day goes by where I don't see a homeless person shivering on the street corner and being harmed by a disease. From my early childhood, my parents engaged me in numerous community-service activities. I would spend many of my weekends at the Salvation Army cooking and serving food for the homeless community. The more I encountered people of lower socio-economic status, I began to realize the direct correlations between access to money and proper health maintenance.

Beyond the United States, I began to notice this problem prevail in other nations, especially those that are underdeveloped. On one of my trips to India, I visited my parents' villages and when talking to some of my relatives, they would always mention how their health is deteriorating and how they need to be extra careful. Often, I would follow up with "What's wrong with your health?" or "What are you having to deal with?" or "Do you know what's going on?" The response I got from all these questions was "I don't know." Beyond being confused and astonished as to why my relatives didn't even know what they were suffering from, I questioned "How?" I got my answer after visiting the local hospital and clinics where I was awestruck. There were not sufficient medical tools to even diagnose a fever, let alone a full body illness. The larger and more powerful tools that I saw in the United States were nowhere to be found in that clinic, only in larger cities in India.

After going back home, I began to delve deeper into this disparity and how accessibility to these essential and accurate tools are not widespread. As I invested countless days into learning about this, I began to uncover the related topics of portability and cost that go into the development of these tools. With diagnostic devices increasing in expense, limiting in portability, and alternatives decreasing in accuracy, many are unable to get timely treatment to protect their health. It's this

immense problem that inspired me to begin my research on developing novel medical devices. I've narrowed down on specific diseases that have extensive screening procedures, require large and expensive tools, and that are prevalent - specifically respiratory illnesses and lung cancer.

Respiratory illnesses often require the use of a spirometer, only available in a clinician's office, and can't be used periodically to monitor chronic conditions. Lung cancer necessitates several CT and PET scans and biopsies before undergoing treatment which can delay early detection and lead to high costs. I conducted hours of initial research pouring over literature on these current medical processes as well as the advances in medical technology. I made it my goal to make a system that is portable, accurate, and accessible to use.

Most of my research was performed independently at home. I developed the code for my system and algorithm using open source platforms at home. I seeked guidance with medical professionals and doctors in understanding what features I need to focus on as well as the specifics of lung cancer and respiratory illnesses.

The mathematical concepts that I needed to carry out my project were ones that I had already learned - calculus, integration, linear algebra, and advanced statistics. In addition, I learned specific physics concepts when building my hardware component to ensure proper accuracy and validation. My project allowed me to explore the world of interdisciplinary research and combine mathematics with physics with biology with computer science. So by translating what I learned inside the classroom and applying it in my research endeavors, I was able to see the beauty of how everything is connected in the world. When encountering obstacles in my work, sometimes I used knowledge from one subject to solve a problem in another subject.

My advice to fellow students who aspire to make bounds in research and create a positive impact in society is to think big. It only makes sense that big problems require big solutions, so don't be afraid to have bold ideas. Chase after those ideas and continue experimenting by doing the appropriate research beforehand. In addition, a great lesson I learned from my experience is that

when you encounter obstacles, solve them with an open mindset. You never know where the answer may be. It can be in another subject where you have to use that content in a different place, and therefore allow yourself to embrace interdisciplinary work. The basics that you learn in math and science classes can take you a long way when embracing complicated research problems.

**Section 2: Project Research**

Lung cancer is the leading cause of cancer-related deaths in the world and succeeding melanoma skin cancer, it is the second leading cancer in men and women. Accessibility to the expensive imaging technology and accurate diagnostic tools for this disease is limited and therefore leads to death at an early stage. The need exists to develop an innovative diagnostic and detection measure to promptly detect the presence and type of cancer. This sparked the interest to develop an innovative lung cancer detection algorithm, LCDetect. It uses Python scripts, machine learning models, segmentation algorithms, and localization of solitary pulmonary nodules (SPN's), tumors, and lesions to detect the presence of cancer, malignity, type and stage. It uses the lung CT scan to provide a thorough diagnosis and reduce the need for PET and biopsy to prevent late detection. The algorithm works in three modules; image preprocessing, image detection, and a convolutional neural network model (CNN). Preprocessing includes noise removal, normalization, image filters, segmentation and augmentation. Extracted regions of interest, that are based on the location of nodules, are passed to image detection for feature extraction through segmentation and pixel calculations. Based on the extracted features, the layers of the CNN categorize the lung cancer using location-based analysis. Through the complex image detection and preprocessing techniques that feed through the layers of CNN model, the generated feature maps perform this classification. After several optimization techniques such as backpropagation and regression, a thorough statistical analysis was performed. Then, the algorithm was deployed on Azure Web Services. The system was trained and tested across 50,000 datasets and patient CT scans from local radiologists. The final

algorithm passed with an accuracy of 98%. LCDetect is an innovative and fully functional solution to accurately detect for adenocarcinoma, squamous cell, non-small cell, and small cell lung cancer and predict the stage. Using low-level computational techniques and the input CT scan, LCDetect satisfies the goals of accuracy, portability, and performance to reduce detection time.

**CNN:**

The convolutional neural network is a deep learning algorithm that consists of multiple layers and perceptrons used for visual imaging. By using CT scans and analyzing the presence of nodules on them, the CNN assigns 'weights' for each significant feature and characteristic of the nodule to accurately determine the presence of a cancerous nodule. The network was developed using Python scripts and Python packages. The different layers of the CNN include locating the nodule and abnormal cell masses, determining their sizes and features such as color and contrast, density, and opulence to determine if cancerous. After training the model with a valid dataset and the optimal weights are used, the model is then fitted and then used to test on another dataset.

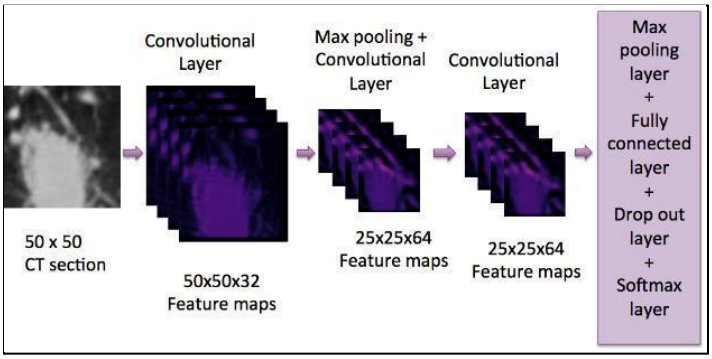



**Image Detection Module:**

TensorFlow was used as the backend for image detection and part of the trained CNN model. The goal of image detection is to localize and detect nodules, differentiate between malignant and benign nodules, and based on the location on the lung scan, predict the form of lung cancer

- Data Augmentation

    - To reduce computational power and time for training the model, the annotation provided regions of interest to resize. Hounsfield conversion units had to be done to scale from the cartesian -> voxel coordinates
    - To overcome imbalance in the training dataset from LIDC, by performing transformations on an image and augmenting the data, more data was created
- Localizing the pulmonary nodules based on specifications of size, calcification, sphericity, etc.

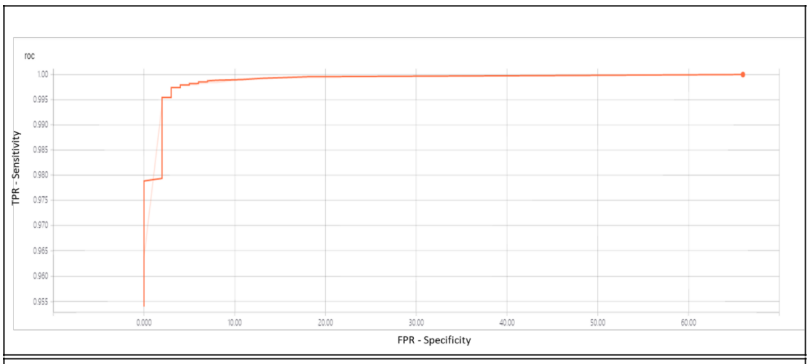
**Datasets:**

- The obtained datasets from the LIDC were split using 80:20 for train:test. Sample patient CT scans from South Sound Radiology were used for the test phase
- The training data was labeled from annotations by radiologists with the classification labels and regions of interest specified for nodule presence
- The raw CT scans underwent the image preprocessing for noise extraction, binary normalization using image filters, and extraction of regions of interest for the train data
- This output is fed to the image detection to classify as a nodule or no nodule based on the image features using pixel method techniques
- Further classification of the original datasets by the CNN provide final diagnosis report
- Test datasets processed through algorithm and validated. Sample patient CT scans fed through model for radiologist confirmation

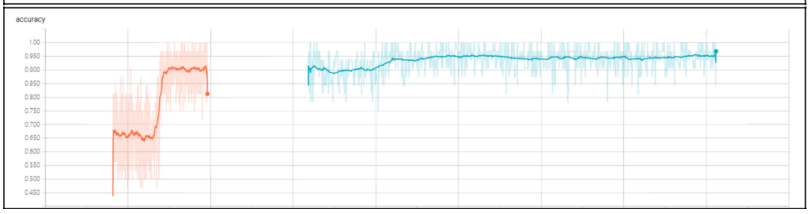
**Algorithm Graph Analysis:**

Below are some graphs that illustrate the computational accuracy and efficiency of the algorithm based on prediction sorts, training accuracy, and loss.

ROC:




Accuracy:




**Statistical Analysis:**

To ensure validity and accuracy of the algorithm, several statistical metrics are measured.

| Test | Purpose | Results |
|---|---|---|
| Specificity | Measures the proportion of actual negatives that are correctly identified | 90.1% |
| Sensitivity | Measures the proportion of actual positives that are correctly identified as the probability of detection | 98.2% |
| Confusion matrix | Categorization of the results into the true and false variants | See journal for confusion matrix image |

| | | |
|---|---|---|
| Train loss | Indicates the false predictions of the model on a single example, penalty for bad prediction | 0.15% |
| Accuracy | The fraction of predictions correctly determined by the model | 97% |

**Block Diagram:**