Problem overview: Lung cancer is the most common cancer worldwide, with over 2.1 million new cases and 1.8 million deaths in 2018 [1]; and it is the leading cause of cancer deaths in the United States, accounting for over 25% of all cancer deaths [2] and estimated to have claimed the lives of over 150,000 Americans in 2018 [3]. The two primary forms are small-cell lung carcinoma (SCLC), a more aggressive form usually found in cigarette smokers, and non-small cell lung cancer (NSCLC), which tends to grow more slowly and takes longer to metastasize beyond the lungs [4].

The primary mode of lung cancer diagnosis is through imaging to detect suspicious nodules, including thoracic X-ray and computed tomography (CT) (**Figure 1**) [5-7]. Imaging alone, however, is insufficient to definitively classify the malignancy of a nodule, so it must be biopsied by inserting a needle into the lung space through the chest, known as CT-guided transthoracic needle aspiration of pulmonary nodules [8]. Radiologists identify the target location and confirm that the sample is extracted from the intended location, and pathologists analyze the sample by immunohistochemical approaches to determine the pathogenicity of the cells [9]. Reports from both are given to the physician team to analyze and develop a treatment plan. The radiologist selects images from the CT imaging stack that contain the biopsy needle, providing evidence as to whether it was targeted to the correct site. Correctly identifying images containing the biopsy needle, however, is time-consuming and susceptible to human error, fatigue, and reader variability [10, 11].

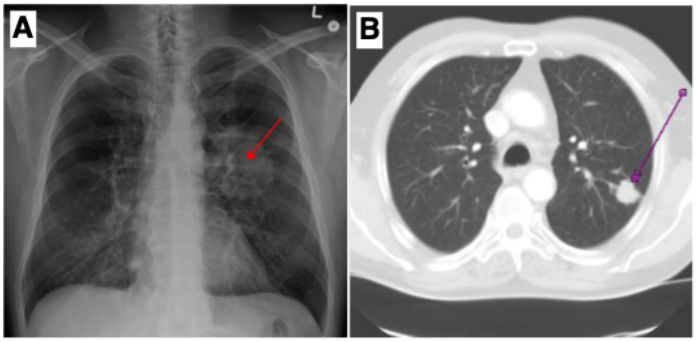

Figure 1: Pulmonary carcinoma visualization with **(A)** x-ray [6] and **(B)** CT imaging [7].

We set out to develop a computational platform to improve accuracy and efficiency in identifying needle-containing CT images from pulmonary nodule biopsy procedures. Provided with a collection of thoracic CT scans from patients undergoing lung biopsy for the evaluation of suspected nodules, we are tasked with developing a machine learning algorithm to automate the process of identifying images containing biopsy needles. The envisioned use case for this approach is to identify needle-containing images from the full stack of images comprising a CT scan from a given patient, raw and unprocessed. The images identified by this algorithm would then be passed on to a radiologist for confirmation and integration into the patient's oncology report, greatly reducing the workload and potential for fatigue and human error. With this use case in mind, the current development has focused on the subset of images provided that have not undergone any preprocessing by a radiologist, as we envision this to be a different task altogether. Further development of a different pipeline could also be undertaken to establish a system to check and confirm the manual work done

by radiologists to identify needle-containing images, focusing on preprocessed, zoomed, cropped, and otherwise altered images established for patients' oncology reports. This report and pipeline, however, will focus on unaltered and unprocessed images, with the aim of developing an approach to aid in the initial identification of needle-containing images from full pulmonary CT scans.

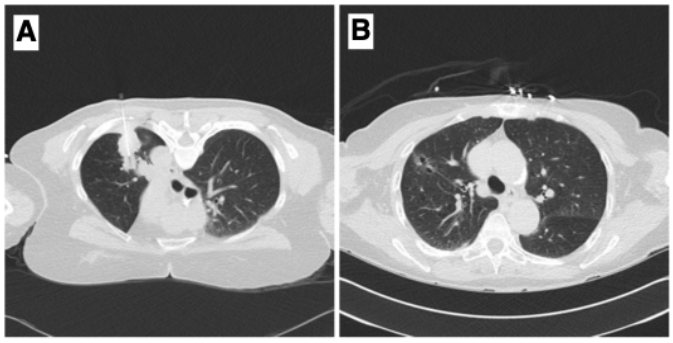

Figure 2: Example CT images from provided dataset demonstrating **(A)** a needle-containing image and **(B)** an image lacking a needle.

Dataset description: The dataset contains a collection of 632 CT scans labeled for the presence (240) or absence (392) of biopsy needles. Example images are shown in **Figure 2**. The images appear to be in greyscale, but the image arrays indicate 3 channels for RGB color, requiring preprocessing. Furthermore, there are a number of non-lung images, poor quality images, and heavily preprocessed images in the original collection. The implications of these variations are discussed in the data preprocessing section below.

Data preprocessing: We first manually inspected the dataset to confirm the original labeling and found a significant proportion of incorrectly labeled images. After manually re-labeling, 422 were found to contain biopsy needles and 209 do not, essentially the reverse of the original balance of image classes. Furthermore, one of the images is duplicated, many images are of very poor quality (33 images identified), and 12 of the images are not thoracic CTs but are instead focused on lower visceral regions, presumably for biopsy of tumors in the colon, kidney, liver, or otherwise. After removing duplicates, non-lung, and poor quality images, the main dataset comprised a total of 587 images, with 385 needle-containing images and 202 without needles.

The dataset also contains a significant number of preprocessed images, as alluded to in the introduction and problem overview. We decided to create a subset of the image collection that would align better with the envisioned use case for this pipeline by removing all zoomed, cropped, and otherwise preprocessed images. The remaining "culled dataset" total rests at 376 images, with 188 in each the needle and non-needle collections.

Additionally, we developed a lung segmentation algorithm to create image masks and exclude the body wall and ribs from the images, effectively reducing the number of features in the images and limiting the focus to the relevant area, i.e., the inner lung space where needle tips are found [16]. The algorithm first binarizes the images, generates borders between the binaries, identifies and labels key features in the images, selects the two largest areas, then does binary dilation and small hole filling to return masks of only the inner lung space. These masks were then applied to the original images and a new dataset of images with only the

inner lung space was used for training our models. The segmentation algorithm was only applied to the subset of 376 full, unprocessed lung images.

Applying the masks, in some cases, resulted in the loss of the needle-containing component of the images due to the proximity of the target nodule to the ribs, or bisection of the lung-space by the needle itself. Further development will be necessary to establish a reliable lung segmentation algorithm that can contend with these edge cases. In the envisioned use case, this pipeline and a 3D lung segmentation algorithm that allows proper resampling, segmentation, and normalization [19] will be applied to full CT image stacks in DICOM format. This approach is expected to work much more effectively and with less sensitivity to the issues mentioned.

After manual re-labeling and lung segmentation, the images were converted to greyscale from the RGB color scheme and downsampled to 256 by 256 pixels from 512 by 512 pixels using the scikit-image Python package for image

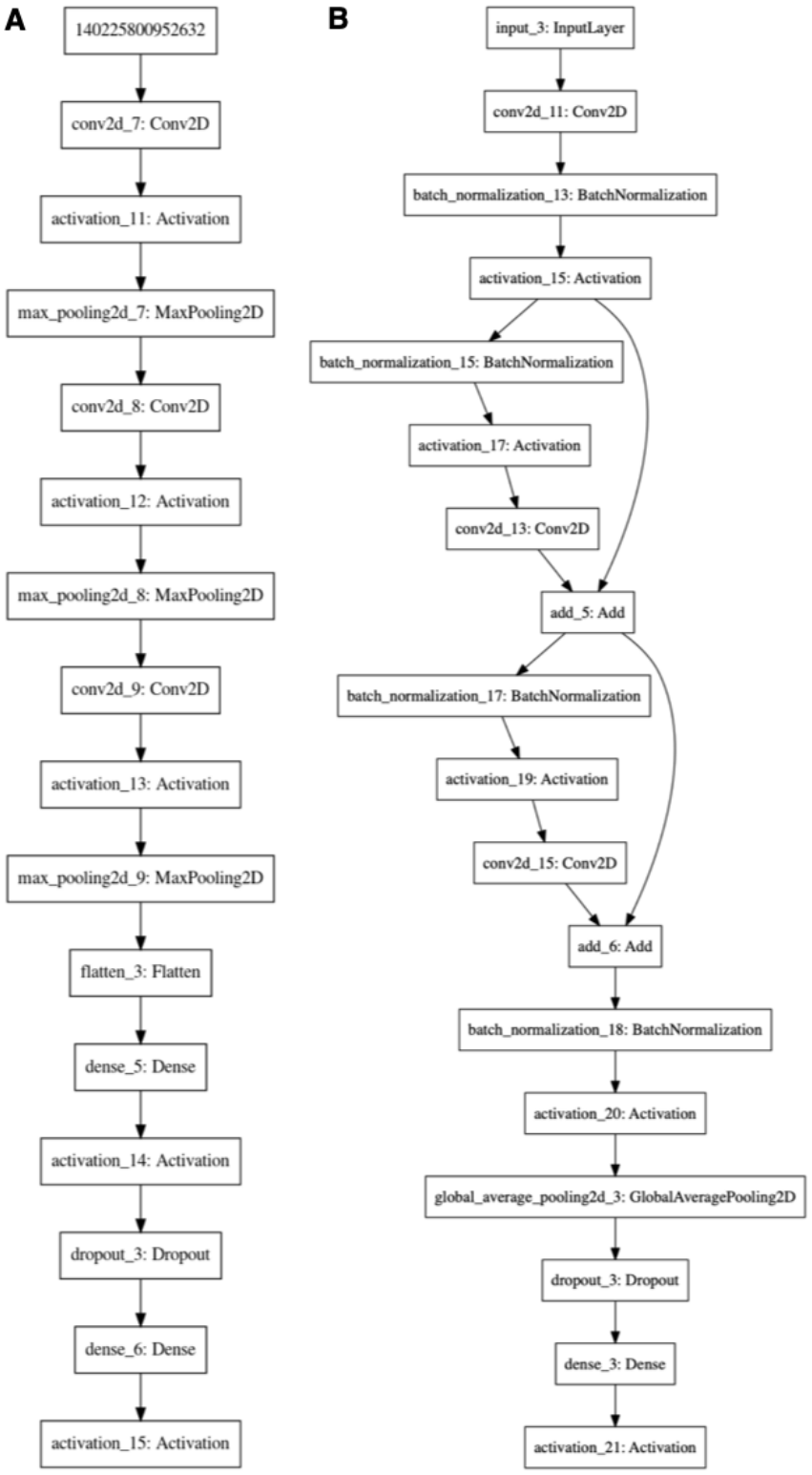

Figure 3: Machine learning models used in development efforts included **(A)** a 16-layer Convolutional Neural Network (CNN) model adapted from VGG and **(B)** a 16-layer Residual Network (ResNet)

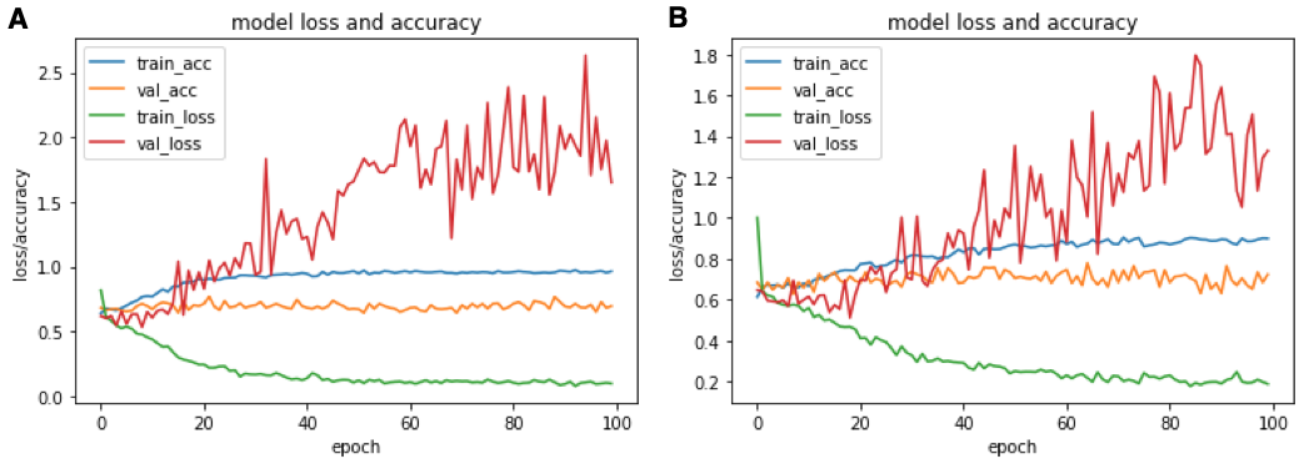

Figure 4: Initial loss and accuracy results on training and validation data on the VGG16 model with the final dropout layer of **(A)** 0.5 and **(B)** 0.75. The training loss function approached convergence with zero and training accuracy reached exceeded 90% in both cases, but validation loss rose well above 1 and continued to oscillate throughout training. Validation accuracy achieved 73%.

processing [12]. Basic *csv* and *os* packages were employed to separate images into two folders based on their labels, and further separated into training, testing, and validation collections in an 80:10:10 split. Using the *random* package, we used a deterministic seed value of 420 to randomly sort the images while maintaining reproducibility and applied the training, testing, and validation split. In an additional effort to improve model training approach, we attempted *k*-fold cross validation [20] to reduce the potential for overfitting to our training dataset; this component is still in development.

Modeling approach: We employed several different modeling approaches to compare their accuracy on both the full dataset and the subset of non-preprocessed images, including Convolutional Neural Networks (CNNs) and Residual Networks (ResNets). First, we constructed a CNN based on an approach used for the "Dogs vs. Cats" Kaggle competition task to classify images as containing either cats or dogs [13]. We began with a 16-layer network developed by the Oxford Visual Geometry Group (VGG) [14, 15]. The VGG-16 ConvNet achieved very good performance on the ImageNet dataset (http://www.image-net.org/) and has been since been implemented successfully in the aforementioned Dog vs. Cat image classification competition. In addition, we constructed a simpler CNN intended for small datasets, a 50-layer ResNet, and a 16-layer "baby"-ResNet [17, 18].

The 16-layer CNN shown in **Figure 3A** includes a sequential layer, 3 convolutional 2D (Conv2D) layers, 5 activation layers, 3 max pooling 2D (MaxPooling2D) layers, a flattening layer, 2 dense, fully connected layers, and a dropout layer (dropout rate of 0.5). After the sequential layer, three series of Conv2D-Activation-MaxPooling2D sets are followed by the flattening layer, the first fully connected layer, the fourth activation layer, the dropout layer,

and the second fully connected layer, ending with the fifth and final activation layer. The 16-layer "baby" ResNet is shown in **Figure 3B** and comprises the initial Input layer, Conv2D, and BatchNormalization layers, followed by 2 sets of Activation-BatchNormalization-Activation-Conv2D series with layer skipping enabled across them, then BatchNormalization, Activation, Global Average Pooling 2D, a Dropout layer, a Dense layer, and a final Activation layer. The ResNet50 network is displayed on the final page. We attempted modifying learning rates, dropout rates, and several different data splits for all models.

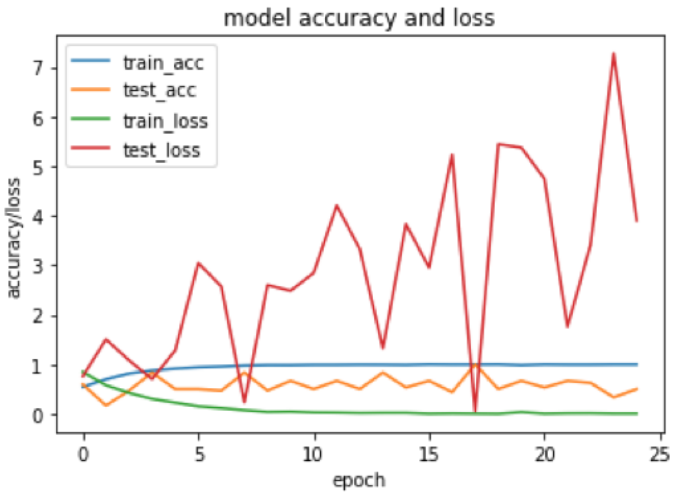

Figure 5: Training and validation loss and accuracy results on the culled dataset with the VGG16. The training loss function approached convergence with zero and training accuracy reached exceeded 90% in both cases, but validation loss rose well above 1 and continued to oscillate throughout training. Validation accuracy oscillated around 50%.

Results: We trained the VGG-16 ConvNet on 80% of the larger, more inclusive image dataset (total images = 587) and validated with 10%, amounting to 504 images for training, 63 for validation, and 64 for testing. The initial training results were not impressive, as shown in **Figure 4**. The training loss function converged to zero and training accuracy reached 99.79%, while validation loss rose well above 1 and validation accuracy hovered around 70%. These data likely indicate overfitting of the model to the training data and beg for optimization of the model, as well as further preprocessing efforts. We proceed with evaluating the test dataset using this model anyway, resulting in an overall prediction accuracy of 73%. We saw no measurable improvement when applying the same models to our smaller dataset comprising only the unprocessed lung images subjected to our lung segmentation algorithm (total images = 376). The results from one of the training attempts are shown in **Figure 5**.

Discussion: We envision this approach being utilized as an initial analytical tool to aid radiologists in their identification of needle containing images in their patients' lung CT imaging, effectively increasing their capacity for image processing by identifying the correct images and reducing the potential for fatigue and human error.

The results obtained from these efforts were not ideal and ultimately did not provide an effective pipeline for needle identification in the set of imaging data provided. However, with a sufficient training set including a reasonable collection of full, unprocessed CT scans from multiple patients, we envision that this approach will be quite powerful and could drastically

improve the capacity and efficiency of radiology departments to analyze their oncology patients' imaging data. Lung segmentation offers only limited utility for the current dataset, given that needles in the needle-containing images are frequently cut out of the image.

In both the full dataset and the smaller dataset with any preprocessed images removed, validation loss tended to spike well above 1, which may indicate overfitting or an issue with the input data. Because of the complexity of the CT images, it will likely be necessary to conduct a much more aggressive preprocessing approach. Annotating the images with bounding boxes to identify regions with and without needles in them could drastically improve training success and model performance. This is, however, a very time-consuming process and will further increase the processing time for model training efforts.

Training neural networks requires a large amount of processing power, and it is highly recommended that any implementation be conducted on a workstation with an available GPU for faster processing. Due to a lack of computational resources available on a MacBook Pro laptop, we rebuilt our Conda Python virtual environment on an Ubuntu workstation with a relatively powerful NVIDIA Quadro M4000 GPU, taking advantage of the 1664 CUDA cores and 8GB vRAM. This provided vastly better processing capacity and improved the training speed of our models by over 10-fold. The initial setup required a significant amount of effort and delayed further model development, but has made training and testing the additional models more efficient and less time consuming, ultimately allowing for a more rapid development process.

Future directions: Further development of this pipeline on a larger dataset of original, unprocessed images is necessary to achieve the desired results. Incorporating a more powerful lung segmentation algorithm that utilizes full CT scans from a collection of 100 or so patients would be ideal. Constructing a 3D lung mask from these datasets would greatly improve our ability to deal with edge cases where needles bisect the lungs and are thus excluded from the inner lung space, or where the needles are targeted to nodules proximal to the body wall. Additional models and optimization efforts are also recommended. An additional pipeline could also be developed that is specifically designed as a sort of quality control mechanism to confirm the identification of needle-containing images by radiologists, and flag those that may not be correctly identified.

Reference:

1. World Health Organization. Cancer Fact Sheet, 2018.
2. Centers For Disease Control And Prevention. National Center For Health Statistics. CDC WONDER On-Line Database, Compiled from Compressed Mortality File 1999-2016 Series 20 No. 2V, 2017.
3. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2018. CA: A Cancer Journal for Clinicians. 2018; 68:7-30. doi:10.3322/caac.21442.
4. Lung Cancer (small-cell and Non-small-cell) - Diagnosis and Treatment. Radiological Society of North America (RSNA), American College of Radiology. Reviewed: Feb 1, 2017. Access date: Feb 18, 2019. radiologyinfo.org/en/info.cfm?pg=lung-cancer.
5. Robert L. Keith, MD. "Lung Carcinoma: Tumors of the Lungs". Merck Manual Professional Edition, Online edition. Updated: Mar 2018. Access date: Feb 18, 2019. https://www.merckmanuals.com/professional/pulmonary-disorders/tumors-of-the-lungs/lung-carcinoma.
6. James Heilman, MD. “Lung CA seen on CXR.” Own work, CC BY-SA 3.0. Created: Jan 17, 2012. commons.wikimedia.org/w/index.php?curid=18076961.
7. Lange123. “CT scan showing a cancerous tumor in the left lung.” CC BY-SA 3.0. Created: Nov 10, 2004. commons.wikimedia.org/w/index.php?curid=2110754.
8. Collins LG, Haines C, Perkel R, Enck RE. "Lung cancer: diagnosis and management". *Am Fam Physician*. 75 (1): 56–63 (2007). PMID 17225705.
9. Anzidei, M. et al. “Imaging-guided chest biopsies: techniques and clinical results.” *Insights Imaging*. 8, 419–428. 2017. PMID: 28639114.
10. Del Ciello, Annemilia et al. “Missed lung cancer: when, where, and why?” Diagnostic and interventional radiology (Ankara, Turkey) vol. 23,2 (2017): 118-126. PMID: 28206951.
11. Singh, Satinder et al. “Evaluation of reader variability in the interpretation of follow-up CT scans at lung cancer screening” Radiology vol. 259,1 (2011): 263-70. PMID: 21248232.
12. Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu and the scikit-image contributors. scikit-image: Image processing in Python. PeerJ 2:e453 (2014) https://doi.org/10.7717/peerj.453.
13. Dogs vs. Cats. Kaggle. https://www.kaggle.com/c/dogs-vs-cats/.
14. Visual Geometry Group. Oxford, UK. http://www.robots.ox.ac.uk/~vgg/.
15. Karen Simonyan, Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” 4 Sep 2014. Revised: 10 Apr 2015. arXiv:1409.1556v6.
16. Lung segmentation. Kaggle. https://www.kaggle.com/kmader/dsb-lung-segmentation-algorithm.
17. Kaiming He, et al. “Deep Residual Learning for Image Recognition.” 2015. arXiv:1512.03385.
18. ResNet16 development.
19. Full lung segmentation and preprocessing. https://www.kaggle.com/akh64bit/full-preprocessing-tutorial
20. K-fold cross-validation. https://machinelearningmastery.com/k-fold-cross-validation/