Using computer vision to detect if there is lung cancer in a CT scan

Papers:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10453592/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419420/

Possible datasets:

- https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction/notebook
- https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
- https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer

- https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70224216

Title: *Summarizes the main idea of your project.*

- Deep Learning for Lung Cancer Detection in CT Imagining

Who:

- Maxwell Ebersman (mebersma), Yen Chu (ychu12), Brian Delgado (bdelgad1), Christian Armstrong (carmstr8)

Introduction: *What problem are you trying to solve and why?*
If you are implementing an existing paper, describe the paper's objectives and why you chose this paper.
If you are doing something new, detail how you arrived at this topic and what motivated you.
What kind of problem is this? Classification? Regression? Structured prediction? Reinforcement Learning? Unsupervised Learning? Etc.

Lung cancer is one of the most common causes of cancer-related deaths, and using deep learning on CT images might help with early detection of the disease. As a disease dependent on

growth, early detection of lung cancer can significantly improve chances of survival. Traditional diagnostic methods may be time consuming and inaccurate, so deep learning developments have potential for improvement in this area.

We are using a paper, "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images" as a general basis for the project. The paper evaluates and summarizes multiple deep learning related methods of screening for and diagnosing lung cancer. The objectives of the paper are to provide a broad overview of the advantages and challenges of specific deep learning techniques.

The motivation for this project is somewhat self-explanatory, because improving diagnostic methods in medicine has the potential to improve management of cancer.

This problem is mostly binary classification, since the task is to classify CT images into categories based on the presence or absence of lung cancer. It should mostly be supervised learning, as it requires labeled training data to learn the mapping between input images and their corresponding labels/masks.

Related Work: *Are you aware of any, or is there any prior work that you drew on to do your project? Please read and briefly summarize (no more than one paragraph) at least one paper/article/blog relevant to your topic beyond the paper you are re-implementing/novel idea you are researching.*
In this section, also include URLs to any public implementations you find of the paper you're trying to implement. Please keep this as a "living list"–if you stumble across a new implementation later down the line, add it to this list.

There have certainly been several prior works which attempted to demonstrate the potential of deep learning models in performing high accuracy lung cancer detection. There are many examples of using 3D CNNs to capture volumetric data and spatial dimensionality.
Here is one example. They attempt to use two separate 3D learning models, with a focus on candidate generation and false positive reduction. The candidate generation model is basically the initial step, as it scans the entire lung volume and generates a list of candidate regions that have a high probability to

contain regions of interest. These candidates regions are then further analyzed to determine the likelihood of containing cancerous lumps, or nodules. The model which focuses on false positive reduction utilizes a 3D CNN for binary classification. The initial system was trained and evaluated using 888 scans, and achieved a detection rate of 94.77% with 30.4 false positives per scan. After combining with the false positive model, the detection rate was 89% with 1.78 false positives per scan.
https://www.sciencedirect.com/science/article/pii/S0957417422017079
https://www.nature.com/articles/s41598-020-70629-3

Data: *What data are you using (if any)?*
If you're using a standard dataset (e.g. MNIST), you can just mention that briefly. Otherwise, say something more about where your data come from (especially if there's anything interesting about how you will gather it).
How big is it? Will you need to do significant preprocessing?
Our current dataset is a 127GB set of images containing CT Scans of lungs from people with suspicion of lung cancer from the Cancer Image Archive. This dataset contains over 250,000 images with XML annotations marking the locations of tumors in the scans. It is contained in one file, but we may be able to extract it using a script with tf.data.Dataset. There shouldn't be a huge need for much preprocessing, other than normalizing data and extracting labels from XML data.

Methodology: *What is the architecture of your model?*
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10453592/

- This paper details a number of various methods used to collect patterns in images using deep learning.

How are you training the model?
Multiple options:

- *Train based on distance of believed location of a tumor (use MSE)*
- *Train based on whether or not the model believes a tumor is present in the image (use Cross-Entropy)*

If you are implementing an existing paper, detail what you think will be the hardest part about implementing the model here.

- We intend on testing a few of them from papers, and to add small variations, including:
- Use CNNs to capture patterns in the image, and then using Transformers to extract whether or not a tumor is present from a CNN encoding
- Use clustering algorithms to collect based on location of the tumor

Metrics: *What constitutes "success?"*
What experiments do you plan to run?
For most of our assignments, we have looked at the accuracy of the model. Does the notion of "accuracy" apply for your project, or is some other metric more appropriate?
If you are implementing an existing project, detail what the authors of that paper were hoping to find and how they quantified the results of their model.
If you are doing something new, explain how you will assess your model's performance.
What are your base, target, and stretch goals?

Our experiments will basically be training and validating the models on the dataset, and differentiating between nodules and non-nodules (as well as benign vs. malignant nodules).

Accuracy is a typical metric but probably isn't plainly the best for this project. Medical datasets are often imbalanced, where the number of non-modules might be much greater than the number of modules. So, we might choose to focus on something like true/false positives/negatives. There are also specific metrics within some of the papers that we might be able to use such as precision and F1 score.

Our base goal might be to, say, detect true positives and true negatives above a threshold like 80%. This would probably indicate we can reliably detect nodules. The target goal might be to achieve a sensitivity of 85% while maintaining a low false positive rate. Considering that the literature was reaching 90%, this might still be very difficult. A stretch goal would be to not only detect nodules with

high accuracy but to reliably classify between benign and malignant. This would require incorporating more features and data sources to improve the diagnostic capabilities.

Ethics: *Choose 2 of the following bullet points to discuss; not all questions will be relevant to all projects so try to pick questions where there's interesting engagement with your project. (Remember that there's not necessarily an ethical/unethical binary; rather, we want to encourage you to think critically about your problem setup.)*
What broader societal issues are relevant to your chosen problem space?
Why is Deep Learning a good approach to this problem?
What is your dataset? Are there any concerns about how it was collected, or labeled? Is it representative?
What kind of underlying historical or societal biases might it contain?
Who are the major "stakeholders" in this problem, and what are the consequences of mistakes made by your algorithm?
How are you planning to quantify or measure error or success? What implications does your quantification have?

Stakeholders and Consequences of Mistakes

The stakeholders in this problem are patients at risk of lung cancer, radiologists, oncologists, and just healthcare systems in general. The consequences of mistakes made by the algorithm are significant. False negatives, where the algorithm fails to detect a malignant nodule, can lead to a delay in (or lack of) diagnosis and treatment, which would worsen the patients chance of survival. False positives, where benign nodules are falsely labeled as suspicious/cancerous, can lead to unnecessary biopsies, a waste of resources, and overall just unnecessary stress. An algorithm with poor generalizability may do well on our test dataset but fail in the real world. When it comes to CT image screening, models must be validated on diverse and representative datasets to decrease the chances of poor real-world usage.

Dataset Concerns and Representativeness

We will be using publicly available repositories of CT scans specifically compiled for development of computer-aided detection systems. This dataset is certainly a great resource, but there may be concerns about representativeness/biases. The data is sourced from multiple institutions, so there may be variability in imaging standardization and patient demographics. The characteristics of a person's lung nodules may be affected by certain factors like ethnicity, age, and gender, so these are important to consider when we think about whether the data properly represents the population. Furthermore, the labeling process is likely accurate and consistent due to its high usage and popularity, but it is possible that somewhere in the labeling process some inconsistencies or errors occurred due to mistakes by radiologists.

Add your own: *if there is an issue about your algorithm you would like to discuss or explain further, feel free to do so.*

D**ivision of labor:** *Briefly outline who will be responsible for which part(s) of the project.*

Maxwell:

Yen: A bit of the outline methodology section, programming

Brian: data section, methodology section

Christian: zoom link

Meeting goals:

- Try MLP, then train autoencoder (use encoder to do classification on encoded bit) [stretch]. Then, play around with different architectures (like HW3).
- https://en.wikipedia.org/wiki/Object_detection

- Object detection could be a stretch goal.