DIAGNOSING THE PRESENCE OF CANCER USING A CONVOLUTIONAL NEURAL NETWORK

CAPSTONE THREE: PROJECT PROPOSAL
Gabe Gibitz

Problem Statement Formation

Can we recognize and diagnose the presence of cancer using machine learning?

Context

Nearly everyone knows a family who has been directly impacted by the effects of breast cancer. It is the most common form of cancer in women. It is a devastating disease, and early detection is an important factor in getting the proper treatment soon enough.

Much work has been done in using deep neural networks to detect cancer, and there is still much more to do.

This particular project will examine slides of patients fighting Invasive Ductal Carcinoma (IDC), which is the most common form of breast cancer.

Criteria for Success

In this project, I would like to see the algorithm reach an accuracy of 90%.

However, an accuracy rating closer to 99% is ideal. This is beyond the scope of this project, but accuracy scores to this degree will help advance progress of cancer detection even faster!

Scope of Solution Space

The original dataset consisted of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive). Here are a few specific items we will focus on to narrow down the scope of this project:

- We will be working with these 162 women.
- We do not know who these women are, their nationality or their background.
- We will be working with this specific form of cancer (IDC).
- We are also assuming these cells have all been diagnosed correctly and aggregated correctly.

Constraints

- I am working with more than 250,000 images, so building a model for this will take several hours with the equipment available.
- Because of this, I will need to turn these images into grayscale images (decreases the complexity by 66%).
- These images are already small, 50 px. by 50 px. We won't know without running other tests, but running these images at 250 x 250 could improve results in the future.
- We are only working with data from less than 200 patients. Increasing the number of patients by 200% could increase our accuracy.

Stakeholders

1. Local hospitals
2. State-wide oncologists

Data Sources

1. The original data set can be found here: http://gleason.case.edu/webdata/jpi-dl-tutorial/IDC_regular_ps50_idx5.zip
    a. These are, again, from 162 patients and downsized into 50x50 color images.