Diagnosing the Presence of Cancer Using a Convolutional Neural Network

FINAL REPORT
Gabe Gibitz

Project Take-Aways

While we hoped to diagnose cancerous cells with an accuracy of 90%, we were not able to break 85% with our three CNN models we created. Moving forward, we will need to experiment with:

- Increasing the pixel size and amount of original images provided
- Increasing, decreasing and tweaking our convolution layers
- Using color images instead of grayscale images
- Balancing the amount of cancerous vs. non-cancerous images either artificially or in more cancerous images provided to the model

1. Problem Statement

Can we recognize and diagnose the presence of cancer using machine learning with **90% accuracy**?

2. 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.

3. Our Approach

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). We accomplished the following:

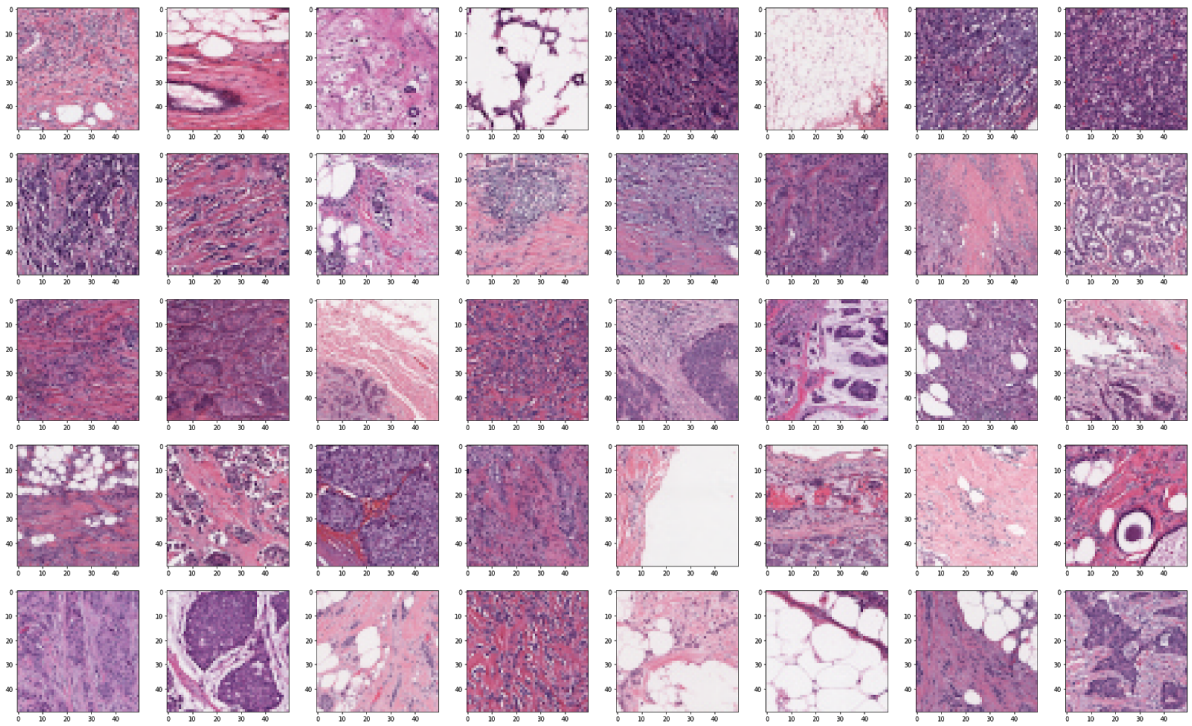
- Wrangled the data and created a dataframe for the entire 250,000+ images.
- Explored the images and data.
- Turned all images into grayscale images.
- Split the data into training and validation.
- Set up three variations of a Convolutional Neural Network.
- Compared the results.

4. Our Findings

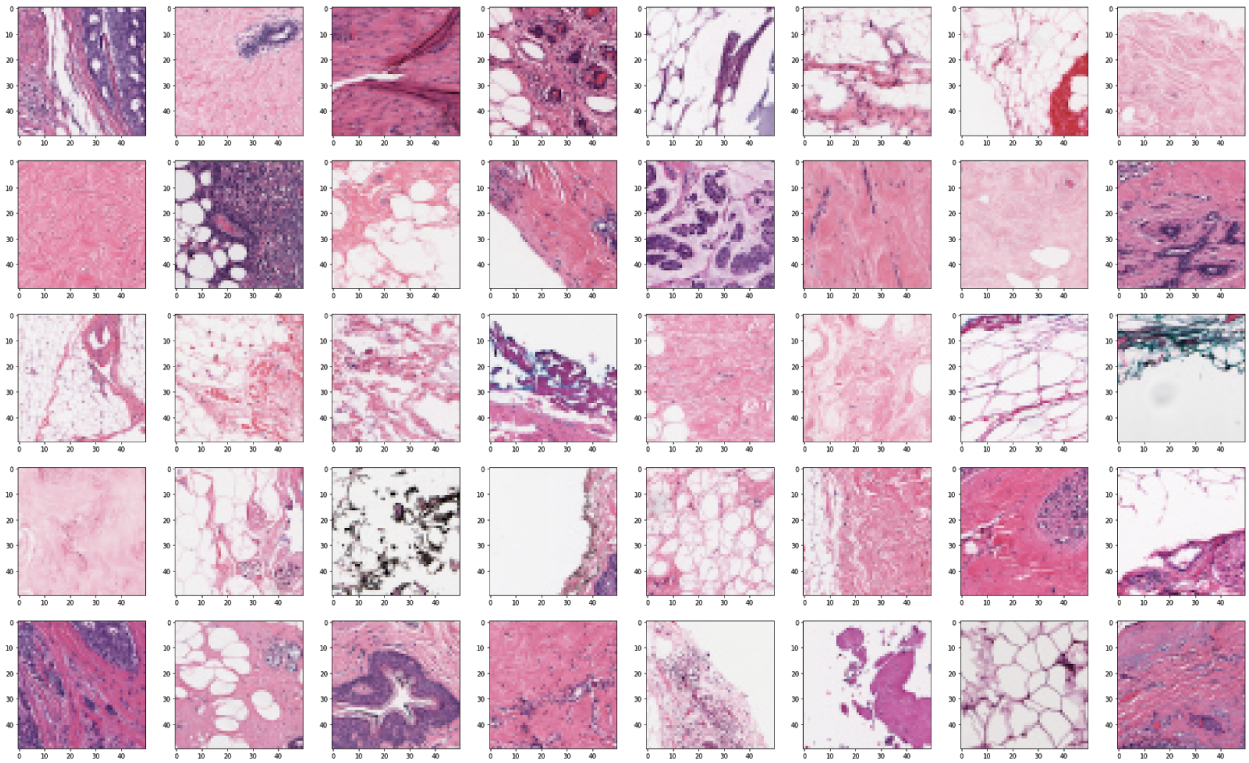
We were able to diagnose the presence of cancer in a given image with nearly 85% accuracy. This was short of our goal of 90% but definitely a solid step in the right direction. Let's take a look at the results of each model.

OBSERVATIONS OF THE IMAGES

Let's turn our attention first to cancerous cells. The next set of images will be non-cancerous cells. To the human eye, we see some key differences. First, cancerous:



Next, non-cancerous tissue:



Looking at this from a data scientist's perspective, I see that:

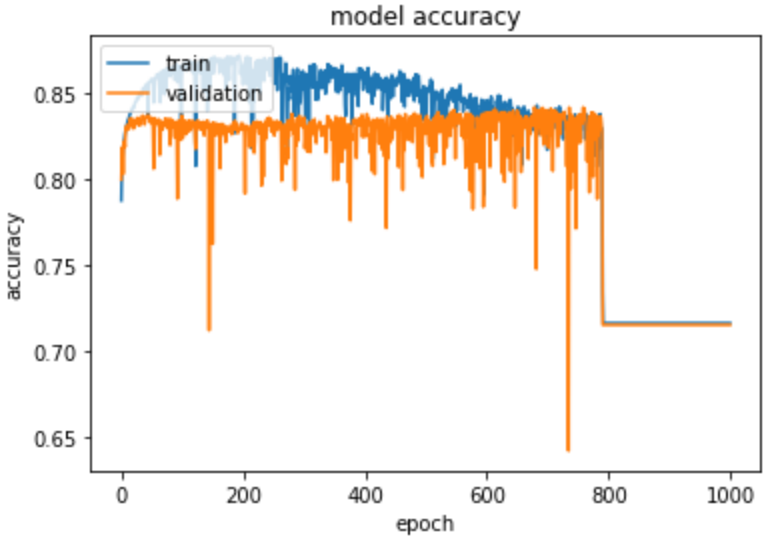
1. The overall color of the slides darkens drastically when you compare the slide samples. Non-cancerous tissue is, on the whole, lighter, a more vibrant pink and more consistent throughout.
2. There is more consistent white space in the non-cancerous images overall.
3. The neural network will most likely have the hardest time categorizing the images that are darker purple and don't contain any white space.

MODEL FINDINGS:

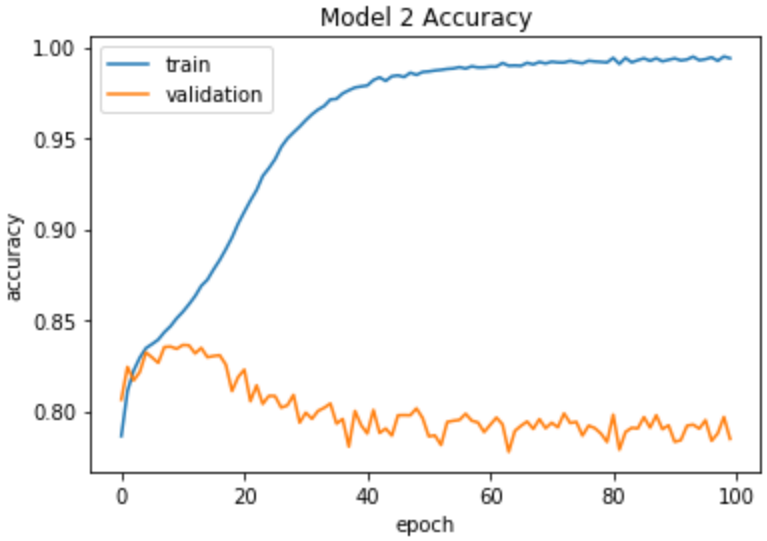
When we ran our models, we walked away with a few key findings.

First, our accuracy hovered right around 83-84% across the three CNNs we created. Whether we were using 3 convolution layers or 2 -- hidden layers present or not -- we could not break 84.5%.

In the first model, we added 3 convolutional layers with two hidden layers, focusing on a large set of epochs (see the graph to the right). The increased amount of epochs actually harmed the results. You can see the training data accuracy slump with a huge drop off around the 800th epoch. If we wanted to tweak these results, we'll have to find out where we seemed to have a neuron die on us.

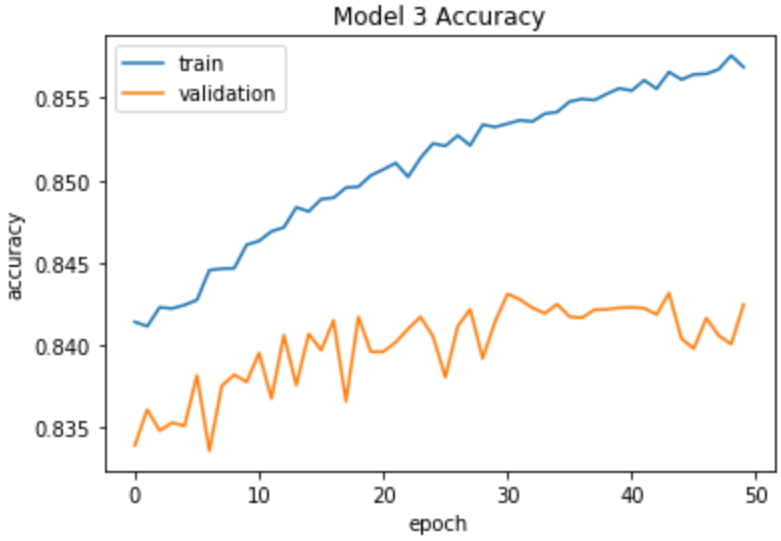


Our second model (see the graph on the right) seemed prone to overfitting. Ten epochs looked like a nice number before seeing the training data shoot up near 100% accuracy with the validation accuracy coasting downward.



Our third model (bottom right graph) showed promise, but nothing more than our first model. Within 50 epochs, we were holding steady at 84.3%. Two convolutional layers with absolutely no hidden layers was enough to give us results that competed with Model 1.

I am pleased with these results, but I know there is significant room for improvement. Let's look at where we can go from here.



5. Recommendations & Further Research

I mentioned this at the beginning of the report, and I will go into these points a bit more in depth. We have room for improvement with these models. In the medical space, this accuracy rating is not enough to make an impact in health care.

Here's where I recommend moving forward:

- Increasing the pixel size and amount of original images provided
    a. Though the model will take longer to fit, the increase in pixel size will allow potentially for more clarity for the CNN to process. Increasing the amount of images to 1-10 million would potentially lead to a significant improvement in accuracy, as well.
- Increasing, decreasing and tweaking our convolution layers
    a. We just scratched the surface of ways we could tweak the models. This was a starting point, and we can tweak this more.
- Using color images instead of grayscale images
    a. This would increase processing time, but there seems to be a strong correlation between color and categorization. The grayscale images account for this, of course, but it's possible that we could see an increase in accuracy with color images.
- Balancing the amount of cancerous vs. non-cancerous images either artificially or in more cancerous images provided to the model
    a. This is a glaring problem. We simply need about 400% more images of cancerous cells to catch up with our non-cancerous cells to allow the model to predict even more accurately. This could be a challenge as we increase our image set, but this is an important part of an accurate model.,

5. Conclusions

We hoped to diagnose cancerous cells with a 90% accuracy rate, but we were unable to break 85%.

This was a relatively small dataset, and we have a lot of enhancements we could make to this project moving forward. As we increase our image base from 250,000 to 10 million or more, I would anticipate our accuracy rate to reach well into the high 90s.

We would love to discuss with you how we can deploy this into your workflow once we increase our accuracy rating with more data. Thank you so much!

Sincerely,
Gabe Gibitz