Cancers are the second highest cause of death in the United States (Kochanek et al. 2020). The factors that lead cells to become cancerous are still largely unknown, but ultimately cancerous cells start dividing rapidly creating tumors. These tumors can remain isolated, but many eventually spread throughout the body via the cardiovascular system. One underlying trend across all cancers is that they are much more treatable and have a much higher survival rate when caught early; colon cancer, for example, has a 91% survival rate when caught early, but only 11% survival rate if it has spread to other organs (Canary Foundation 2020). Breast and prostate cancers are the best examples of this, these have a 98 and 100% survival rate respectively when caught early (Canary Foundation 2020). Breast cancer is one of the most common cancers in the US, and about one in eight women (or 13%) will develop breast cancer.

Invasive ductal carcinoma (IDC) is the most common form of breast cancer, accounting for about 80% of cases (breastcancer.org 2021). IDC develops, as the name suggests, in the mammary ducts, and is found in around 180,000 women every year in the US (breastcancer.org 2021). Because of its pervasiveness in the population, mammograms are often performed as part of a woman's annual physical, especially after the age of 55. Mammograms are a key procedure in detecting breast cancers (including IDC), but usually a biopsy (or tissue sample) is needed for final diagnosis. These tissue samples are placed on microscope slides and the images are reviewed by specialists to determine the presence of cancerous tumors. This is a rigorous and time consuming endeavor that could cost the patient and doctor valuable time in deciding on a prognosis and treatment. As these images are often digitized, I propose using a convolutional neural network (CNN) to identify potentially cancerous tumors automatically, freeing the practitioner to focus on prognosis and treatments.

CNNs are a well-known deep-learning model in data science as an effective way of training a categorizing AI to identify aspects of images. CNNs work by mapping some (but not all) interactions of known inputs and estimated convolutions. The CNN then "infers" (more accurately estimates) other interactions not mapped. This reduced mapping saves on processing power required to run the CNN (critical for large data sets and image data).

For my project, I investigated the use of a CNN to identify cancerous regions of IDC positive tissue sample images from 280 patients. This is an existing data set available on Kaggle (Kaggle 2017). These tissue sample images were broken up into 277,524 slices (about 1,000 slices per patient) and labeled as either IDC positive or negative. There is no associated csv file with file names to be used as a data frame, so I made one for easier reference. The images are named with patient IDs, grid locations (for reconstructing the entire biopsy image), and IDC class (0 for negative, 1 for positive). They are redundantly organized in folders by patient ID and subfolders by IDC class. This made it easier for building a data frame with critical information such as patient ID, image ID, and IDC class. Table 1 gives a brief overview of the data frame I constructed from the file names.

Table 1: First five lines of my IDC data frame containing all pertinent data for analysis and modeling.

index	**patient_ID**	**path**	**image_ID**	**IDC**	**label**
0	10253	.../breast_histology mages/IDC_reg_ps _idx5	10253_idx5_x1001_y1001_class0.png	0	neg
1	10253	.../breast_histology mages/IDC_reg_ps _idx5	10253_idx5_x1001_y1051_class0.png	0	neg
2	10253	.../breast_histology mages/IDC_reg_ps _idx5	10253_idx5_x1001_y1101_class0.png	0	neg
3	10253	.../breast_histology mages/IDC_reg_ps _idx5	10253_idx5_x1001_y1151_class0.png	0	neg
4	10253	.../breast_histology mages/IDC_reg_ps _idx5	10253_idx5_x1001_y1201_class0.png	0	neg

After creating the data frame, I explored the data, ensuring no null values were present, and inspecting the ranges of data. One of the first things I noticed was that while the average patient would have around 1,000 patches per image, the actual number ranged dramatically from about 100 to 2,400 (Figure 1a). This indicated that for some of the tissue samples, we were working with incomplete biopsy images, this isn't an issue given that our data is really composed of these slices and each slice is denoted as having cancerous cells or not. The second discrepancy I noticed is the vast difference in counts of images with and without cancer: there were 198,738 images without cancer and 78,786 images with cancer (Figure 1b). When investigating use of CNN models I found that some people like to even out the total number of classes so that you're always working with the same number of images in each class. I prefer to work with as much data as possible and further research showed that one feature of CNNs is that they will automatically attribute weights to differing counts in classes to accommodate these differences. Lastly, I noticed that there was a significant range in number of slices containing cancerous cells, from zero (having no cancer) to over 80% of the slices containing cancerous cells (Figure 1c). None of these factors are worrying as I prepared to build the CNN, but it was important to evaluate these features.

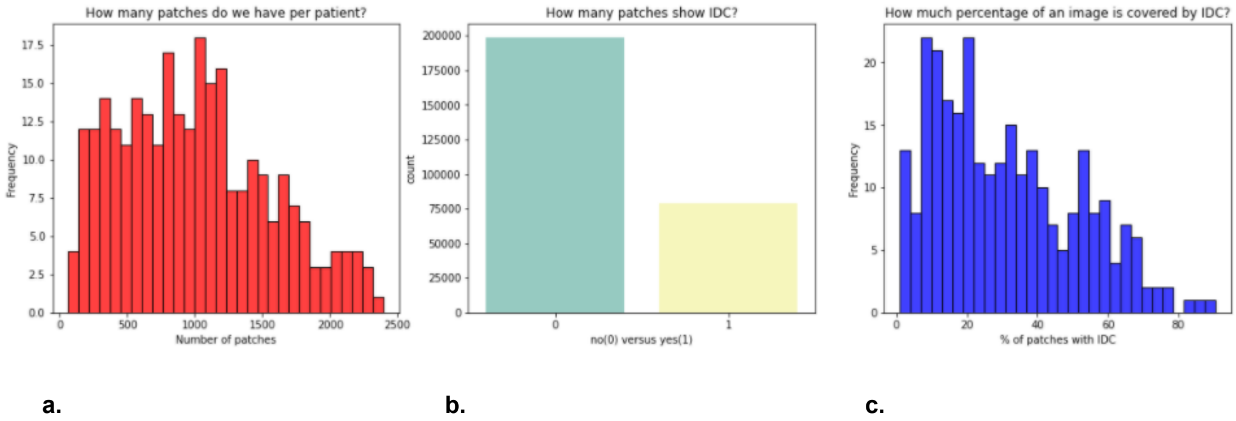

Figure 1: Distribution of number of patches, or image slices, per patient (**a**). Count of benign tissue slices (no) and cancerous tissue slices (yes) (**b**). Distribution of percentage of full tissue samples with cancer (**c**).

The next step in my exploratory data analysis was to construct some binary visualizations of full biopsy samples. I randomly selected 15 patients to do this for and generated a binary heat map of each biopsy, noting cancerous regions in red and benign regions in blue (Figure 2). Following this, I was curious to see what a full biopsy actually looked like, so I reconstructed one biopsy image from the image slices for that patient (Figure 3a). I then highlighted in red the regions containing cancer (Figure 3b).

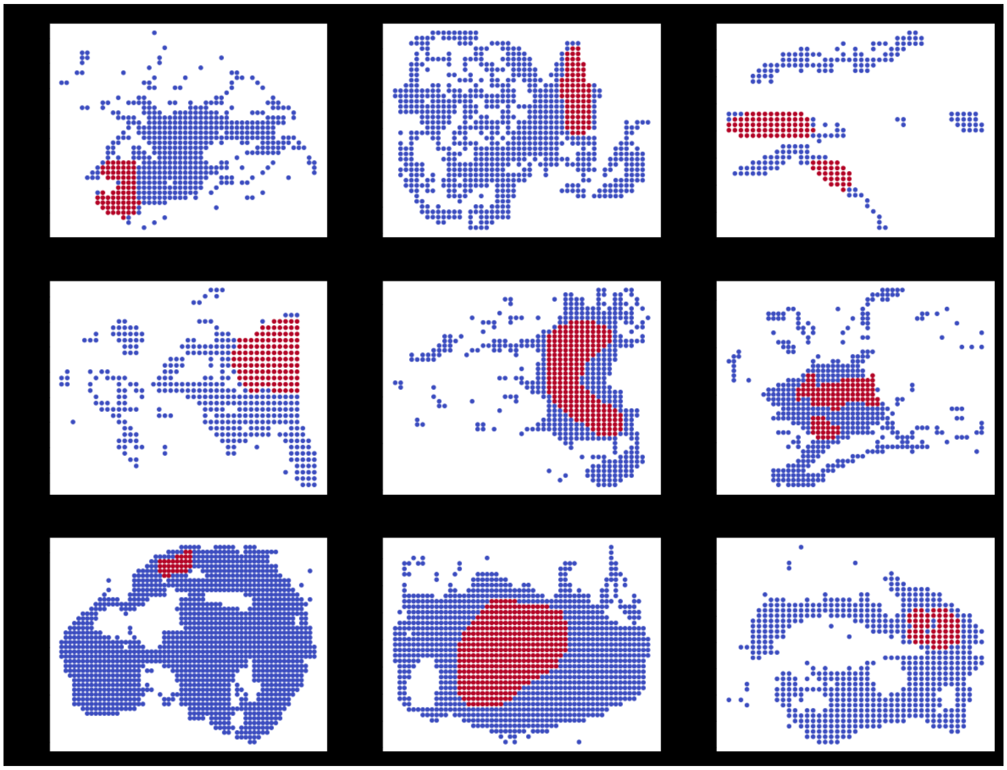

Figure 2: Randomly selected patient tissue biopsy samples, reconstructed as dot illustrations. Blue dots represent benign tissue, and red dots represent cancerous tissue.

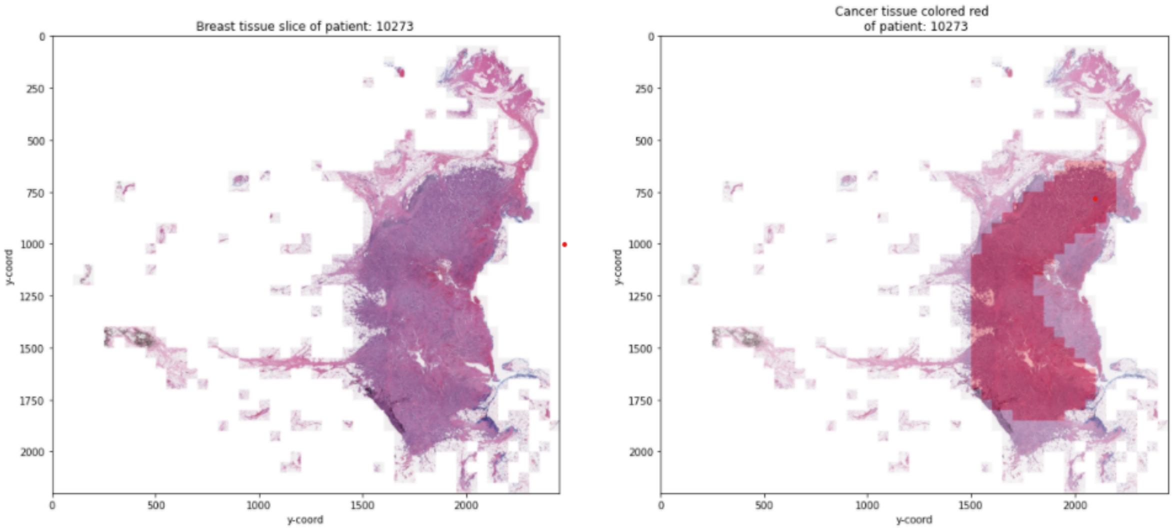

Figure 3: Example tissue sample biopsy of patient 10273 (**a**). Same tissue sample biopsy but with cancerous regions highlighted in red (**b**).

Following this, I began to build my convolutional neural network model. I used the Sci-Kit Learn Train_test_split function to split the image data to a 80% training batch and a 20% testing batch. I designed a function to resize all image slices in each set to be 50X50 pixels, then set a batch size of 10 to reduce the required processing power and memory for analyzing all 277,524 slices. I then began building the CNN model with kernel size of 3, pool size of 2 (for each category), and filters for each kernel of 32, 64, and 128 respectively. Next, I tested differing efficacies of epoch values from 1 to 20 to identify the correct hyperparameter to use (figures 4a and 4b). Based on the test accuracy and loss values I determined that an epoch value of 15 was optimal before the model began overfitting the training data (figures 4a and 4b).

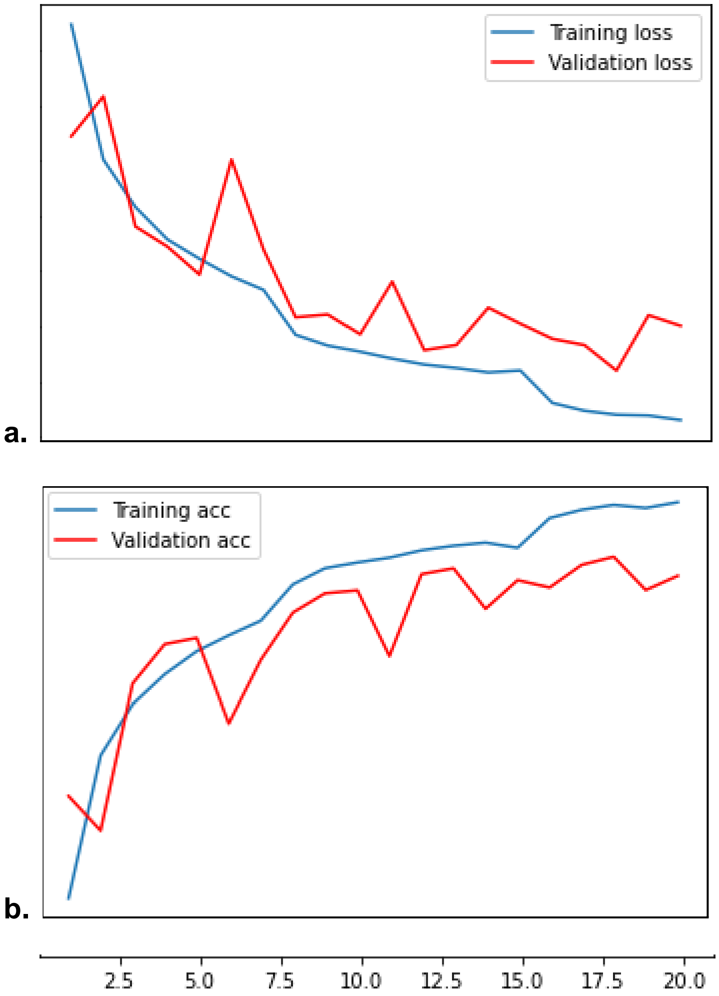

Figure 4: Training and validation loss per incremental increases in epoch values (**a**). Training and validation accuracy per incremental increases in epoch values (**b**).

When I ran the finalized convolutional neural network on the test data, it identified cancerous and non-cancerous tissue with 89% accuracy (Table 2). In addition to the raw accuracy metric, which is often a poor representation of a model's actual performance, I calculated precision, recall, and F1-scores, as well as the receiver operating characteristics (ROC) area under the curve (AUC) (Table 2). This last metric yielded a 95% accuracy for my model. The final metrics for this CNN are very promising, and by adjusting the train-test split values, and kernel and filter sizes I could potentially generate a more accurate model. However, as the famous George Box once said, "all models are wrong, but some are useful." Ultimately, there likely won’t be a perfect model for identifying cancerous tissues from biopsies, and actual review of the tissue samples by knowledgeable doctors will still be required, no matter how accurate the model is. My goal, though, is only to reduce the number of samples that a doctor would need to review by creating an accurate model that can identify with over 90% accuracy, patients that do or don’t have cancer.

Table 2: Final model summary outputs: accuracy, precision, recall, F1-scores, and ROC-AUC score.

	Precision	Recall	F1-score	ROC-AUC	Support
Benign	0.94	0.90	0.92	N/A	39866
Malignant	0.77	0.84	0.81	N/A	15639
Accuracy	N/A	N/A	0.89	0.95	55505
Macro Avg	0.85	0.87	0.86	N/A	55505
Weighted Avg	0.89	0.89	0.89	N/A	55505

Cancer isn’t the only application for a convolutional neural network image classification model in modern medicine either. During my research and development of this model, I found similar models used to identify pneumonia in patients from X-ray images. In addition to these, we could apply a similar approach and try to identify breast cancer earlier in a patient’s diagnosis, using X-ray images from mammograms. These are done more regularly, and could help identify cancers very early on, which, as mentioned above, is critical for improving a patient’s prognosis.

References:

Breastcancer.org. US Breast Cancer Statistics. 2021. Accessible here: https://www.breastcancer.org/symptoms/understand_bc/statistics

Canary Foundation. Early Detection Facts and Figures. 2020. Accessible here: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiQgZaL3KPvAhXDU80KHWt4Cn0QFjABegQIARAD&url=https%3A%2F%2Fwww.canaryfoundation.org%2Fwp-content%2Fuploads%2FEarlyDetectionFactSheet.pdf&usg=AOvVaw2Mm3jUoxeertSdmNVeT5zA

Kochanek, Kenneth D., Jiaquan Xu, and Elizabeth Arias. Mortality in the United States 2019. 2020. NCHS Data Brief 395. Accessible here: https://www.cdc.gov/nchs/products/databriefs/db395.htm

Data available at: https://www.kaggle.com/paultimothymooney/breast-histopathology-images

Code adapted from:

https://www.kaggle.com/angieashraf/eda-breast-histopathology

https://www.kaggle.com/allunia/breast-cancer

https://www.kaggle.com/souro12/breast-cancer-cnn