Classifying Breast Cancer with Deep Learning Methods in Python

# **Augustus Madsen Summer 2020 https://github.com/AMadsen32/DSC680.git**

# Which Domain?

The dataset will come directly from the website Kaggle where the initial analysis was performed – there are various other links describing some more in-depth research behind the dataset and the cancer imaging. These provided more of a background for how the samples were collected. It is primarily Kaggle where the model selection and evaluation are performed based on this specific dataset, with other branching models from similar data in other locations.

<https://data-flair.training/blogs/project-in-python-breast-cancer-classification/>

<https://www.kaggle.com/lbronchal/breast-cancer-dataset-analysis>

<https://www.kaggle.com/paultimothymooney/predict-idc-in-breast-cancer-histology-images>

<https://pubmed.ncbi.nlm.nih.gov/27563488/>

<http://spie.org/Publications/Proceedings/Paper/10.1117/12.2043872>

<https://www.ncbi.nlm.nih.gov/pubmed/27864452>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3893344/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4952020/>

<https://github.com/sfikas/medical-imaging-datasets>

<https://data.world/datasets/breast-cancer>

# Which Data?

The dataset will be the IDC\_regular dataset through the Kaggle site. It is made up of breast cancer histology image data. There are over two hundred thousand patches of specific sizes extracted from different cancer specimens on slides collected from imaging. Roughly 198,000 of the slides were negative results, with 78,000 testing positive for IDC. Here is a direct link to the dataset:

<https://www.kaggle.com/paultimothymooney/breast-histopathology-images>

Additionally, here is a link to the codebook for a similar study done with Wisconsin data and the variables associated with the characteristics of the images collected:

<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

# Research Questions? Benefits? Why analyze these data?

The ultimate goal is to detect whether or not IDC exists in images that are initially unlabeled. I hope to predict, with a greater accuracy and on a larger scale, the presence of this type of cancer in these images. The research question is centered around whether or not a breast cancer subtype is present in an image that has not been analyzed. The benefit here is to reduce error associated with breast cancer classification and also reduce the amount of time it takes to determine if the cancer is present, with an automated task running to detect key characteristics in the images provided.

# What Method?

I will be working in Python and importing several modules, along with the initial dataset, to analyze the images. I will use multiple arrays of the imported images to assign labels based on a classification algorithm. Random samples will be used for testing and compared along histograms the count of labels, whether IDC is present or not, in the distribution. These methods will allow a model to be created with a certain amount of accuracy as to which images contain an IDC sample. This is useful for comparing the training and testing sets alongside a matrix to see how well the model performed. I plan to expand on the original method with several models to compare, applying more imaging techniques on different Python shape and color identifiers, and attempting for a higher accuracy in the output.

# Potential Issues?

One of the potential issues here is that there could be bias associated with several different steps in the model creation. The creation of a matrix could predict the presence of IDC in too many samples based on how the parameters are set, and compared to the curve of the machine learning it might result in a low score than the training data. The author found a similar issue when creating a “confusion matrix” – I will need to make sure that I take the time to really understand how the model is working so that I avoid the same issues.

# Concluding Remarks

Invasive ductal carcinoma (IDC) is the most common type of breast cancer – and breast cancer is the most common type of cancer in women. My stepmother, along with two aunts and a grandmother, were affected by this disease. It is crucial that available resources be applied to classifying images and identifying at-risk women, along with cases where IDC is present and can be treated as soon as possible. With the use of Python, images can be imported and models applied to determine which samples contain IDC and how those numbers compare to the set, along with graphs, visualizations, and distributions detailing where this subset is present.