Breast Cancer Detection and Prediction with Machine and Deep Learning

# **David Kinney Spring 2021 https://github.com/dkinneyBU/breast-cancer**

# Domain

Breast cancer detection and prediction, based on both a tabular dataset (created by Dr. William H. Wolberg, using fluid samples taken from patients with solid breast masses) and histopathology images (digitized images of fine needle aspirates (FNA) of a breast mass).

**References**

[1] “Building a Simple Machine Learning Model on Breast Cancer Data” vishabh goel

vishabh goel Sep 29, 2018 <https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3>

[2] “Breast Cancer Images Classification” Salah Sammari n.d. <https://www.kaggle.com/midouazerty/breast-cancer-images-classification>

[3] “Breast Histopathology Images” Paul Mooney n.d. <https://www.kaggle.com/paultimothymooney/breast-histopathology-images>

[4] “Breast Cancer Wisconsin Data Set” <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>

# Data

* Breast Cancer Wisconsin (Diagnostic) Data Set <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>
  + Attribute Information:
    - ID number
    - Diagnosis (M = malignant, B = benign)
    - 3-32 Ten real-valued features are computed for each cell nucleus:
    - radius (mean of distances from center to points on the perimeter)
    - texture (standard deviation of gray-scale values)
    - perimeter
    - area
    - smoothness (local variation in radius lengths)
    - compactness (perimeter^2 / area - 1.0)
    - concavity (severity of concave portions of the contour)
    - concave points (number of concave portions of the contour)
    - symmetry
    - fractal dimension ("coastline approximation" - 1) [4]
* Breast Histopathology Images <https://www.kaggle.com/paultimothymooney/breast-histopathology-images>
  + Content - 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). Each patch’s file name is of the format: uxXyYclassC.png — > example 10253idx5x1351y1101class0.png . Where u is the patient ID (10253idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC. [2]

# Research Questions

1. What, if any, benefits can machine learning bring to the detection of breast cancer?
2. What, if any, benefits can machine learning bring to the prediction of breast cancer?
3. What is an acceptable level of false positives?
4. What is an acceptable level of false negatives?
5. Which machine learning classification models provide the most promising results?
6. Do deep learning neural networks offer any benefits over “traditional” machine learning models?
7. For the Images Classification model, the sample size is relatively small. Is overfitting an issue?
8. Why does this research matter?

# Method

I plan to explore this project by leveraging both Machine Learning for prediction and Deep Learning for detection.

* Machine Learning – for this effort I will leverage the Breast Cancer Wisconsin dataset, initially applying exploratory data analysis (EDA) to get a feel for the data schema, and then various Machine Learning Classification algorithms to train a model to predict breast cancer detection.
* Deep Learning – while also a Classification application, this approach will be based on image analysis.

# Potential Issues

At this stage I do not anticipate any major issues, given the wide availability of both data and research on this subject.

# Concluding Remarks

I originally planned to do my final project on predicting mortality rates based on how much someone abused their body during their lifetime—smoking, alcohol, obesity, etc. But much to my chagrin, finding relevant datasets became a lost cause. I had taken a day off from work and spent a fair amount of time in my (newly renovated) back yard, just thinking. Cancer in general is devastating, but for some reason, breast cancer has always struck a deeper emotional chord with me. While we humans hurtle forth and evolve towards gender equality, even so, on a visceral level I feel that a woman recovering from a mastectomy must be particularly devastating, when that part of a woman’s body speaks to so many things, perhaps the greatest being motherhood. Again, not a woman, so all I can do is imagine how I would feel in that situation.

And if I now switch over to the Data Scientist, there is a veritable plethora of data as well as activity on Kaggle and Towards Data Science regarding this subject. So, unlike my original idea, there is ample data available, as well as research to review.

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society. [1]

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling. [1]