## Proposed by: Michael Brewer

## Goals

The goal of this project is to is to explore classification methods for image classification for biomedical imaging. More specifically, goals are to

1. Gain experience using data augmentation and data balancing techniques for image classification,
2. Understand the impact of having limited data sets, for at least one of the classes, and how well data augmentation can offset any negative impact, and
3. Visualize the impact of the augmentation as in relates to the areas of interest utilized by the models to determine classification.

## Data

The Proposed Dataset (Al-Dhabyani W, 2020) is available in the public domain and accessible via Kaggle:

<https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset>

The dataset includes ultrasound images with groundtruth labeling for three classes: normal, benign, and malignant. The groundtruth includes the classification as well as the mask image for the region identified as benign or malignant. For this project, the mask will not be used.

## Proposed Test

1. Identify a successful classification model that has been utilized in the literature or the competitions for this or similar datasets. The model will be one that can be trained without excessive compute resources.
2. Train three variations of the model
   1. A model with data oversampled using data augmentation to balance the three classes,
   2. A model with data undersampled to balance the three classes, and
   3. A model with with the ‘benign’ class reduced to only a few samples, and then augmented using data augmentation techniques.
3. Then examine the confusion matrices for the three models
4. For a sample of the benign class, visualize the areas of focus for each of the three models.

## Example

There are several source notebooks which have utilized this or similar datasets, and I propose utilizing one of them as a starting point. For example,

[🩺Cancer Detection with PyTorch | ResNet | CNN | Kaggle](https://www.kaggle.com/code/anityagangurde/cancer-detection-with-pytorch-resnet-cnn/notebook)

While this example does not provide good accuracy, it gives a starting point upon which we can build.

# Bibliography

Al-Dhabyani W, G. M. (2020). *Dataset of breast ultrasound images. Data in Brief.* doi:10.1016/j.dib.2019.104863