Breast Cancer Prediction and Detection through Machine Learning and Deep Learning

David Kinney

Bellevue University

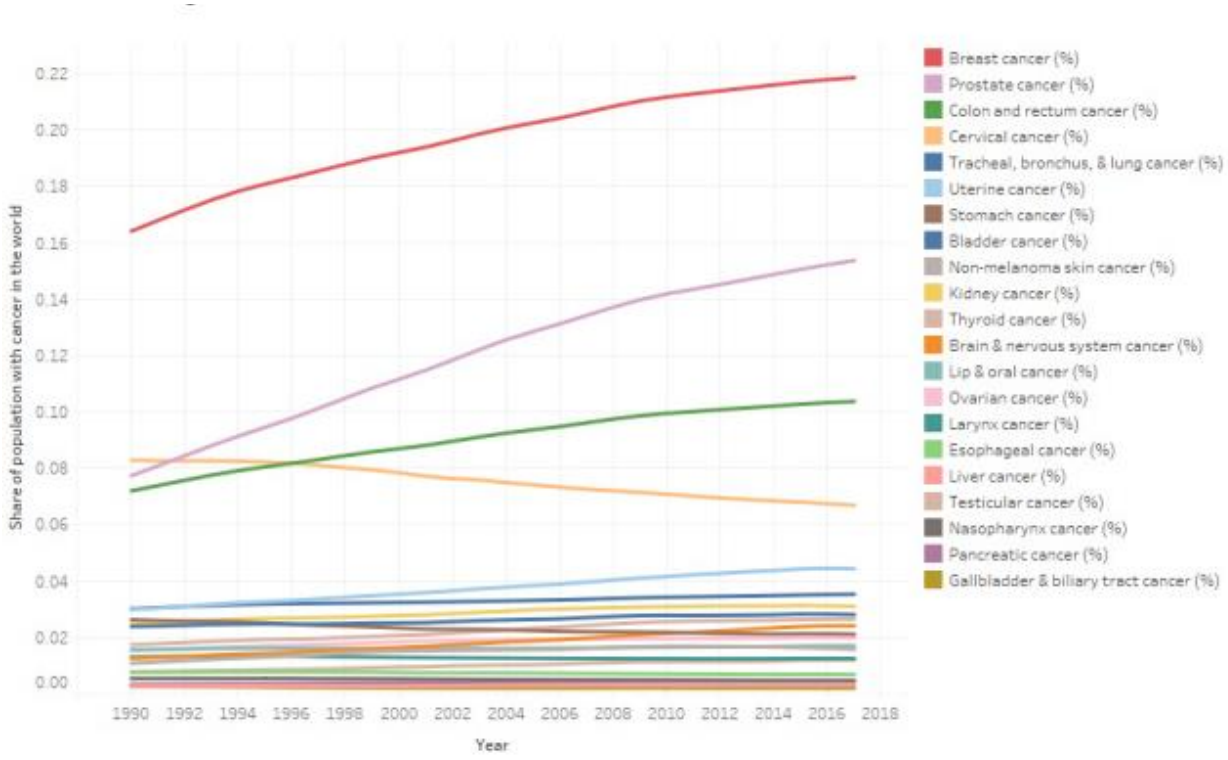
DSC 680 Applied Data Science

Professor Catherine Williams

May 30, 2021

Introduction and Hypothesis

Cancer is a serious public health issue worldwide and the second leading cause of death in the United States. According to the International Agency for Research on Cancer (IARC), about 18.1 million new cases and 9.6 million deaths caused by cancer were reported in 2018. [1] Breast cancer accounts for the largest share of cancer types world-wide and is on the rise.



Source: https://www.kaggle.com/midouazerty/breast-cancer-images-classification

False Negative Results

In cancer screening, a negative result means no abnormality is present. False-negative results occur when mammograms appear normal even though breast cancer is present. Overall, screening mammograms miss about 20% of breast cancers that are present at the time of screening. [2] While a false positive result may lead to undue stress and worry, the end result is no cancer. False negatives are far more alarming, as the result in this case is a woman who believes she is cancer-free when she is not.

Hypothesis

Great strides have been made in both Machine and Deep Learning in various medical fields regarding the prediction and detection of certain diseases. One of these areas showing promising results is breast cancer; both prediction based on observable measurements and detection regarding whether a tumor is benign or malignant. In this paper I hope to show two examples of models I trained that offer impressive results in this field.

Method

Data

I am using two datasets for this project; one is tabular, one is images. The tabular dataset (Breast Cancer Wisconsin (Diagnostic) dataset) [3] consists of 33 measurements; there are no categorical variables, therefore no encoding was required.

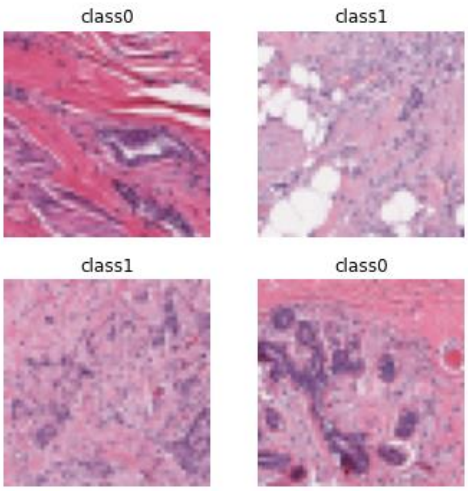
Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. A predictor label is provided: M for a malignant tumor, B for benign. Further, there are no missing attribute values. Class distribution: 357 benign, 212 malignant. [4]



The images dataset [4] consists of 162 whole mount slide images of breast cancer specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 negative and 78,786 positive). Class 0 represents benign tumors while Class 1 represents malignant tumors.

Machine Learning Model

The features for the Breast Cancer Wisconsin dataset Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. [3] The **PyCaret** library was used to select the best model for this application, which turned out to be *Extreme Gradient Boosting.*

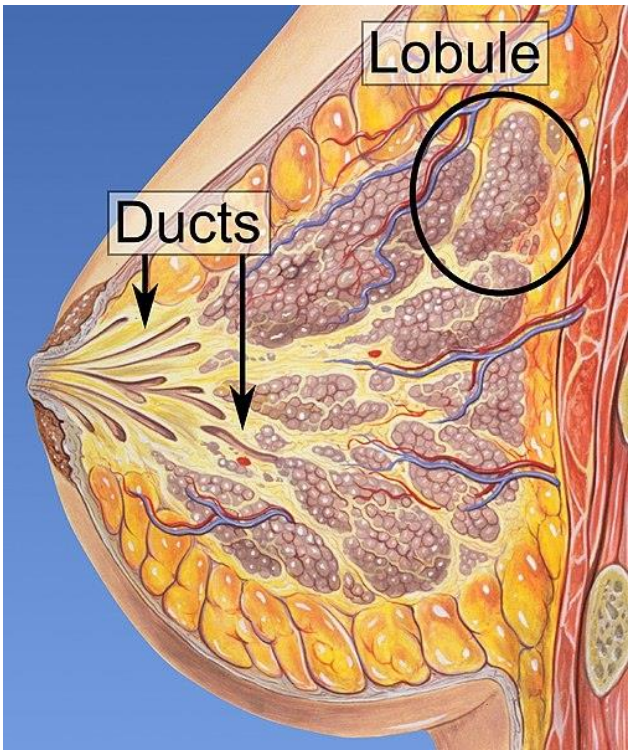
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.9624	0.9883	0.9500	0.9571	0.9521	0.9213	0.9231	0.1190
lda	Linear Discriminant Analysis	0.9599	0.9917	0.9058	0.9929	0.9457	0.9143	0.9185	0.0040
lightgbm	Light Gradient Boosting Machine	0.9599	0.9891	0.9438	0.9569	0.9480	0.9155	0.9183	0.3720
qda	Quadratic Discriminant Analysis	0.9574	0.9846	0.9433	0.9535	0.9474	0.9117	0.9130	0.0040
et	Extra Trees Classifier	0.9549	0.9920	0.9438	0.9440	0.9429	0.9057	0.9069	0.0650
catboost	CatBoost Classifier	0.9549	0.9911	0.9375	0.9508	0.9425	0.9054	0.9076	3.4790
ridge	Ridge Classifier	0.9525	0.0000	0.9000	0.9790	0.9360	0.8986	0.9025	0.0040
rf	Random Forest Classifier	0.9498	0.9913	0.9375	0.9405	0.9368	0.8953	0.8982	0.0880
gbc	Gradient Boosting Classifier	0.9474	0.9885	0.9250	0.9446	0.9319	0.8892	0.8927	0.0530
ada	Ada Boost Classifier	0.9448	0.9805	0.9188	0.9448	0.9278	0.8835	0.8877	0.0260
dt	Decision Tree Classifier	0.9146	0.9104	0.8875	0.9064	0.8911	0.8214	0.8285	0.0030
knn	K Neighbors Classifier	0.7390	0.7568	0.5071	0.7683	0.6037	0.4219	0.4463	0.0130
nb	Naive Bayes	0.6056	0.8843	0.0250	0.4000	0.0471	0.0148	0.0392	0.0030
svm	SVM - Linear Kernel	0.5169	0.0000	0.4000	0.1569	0.2254	0.0000	0.0000	0.0060
lr	Logistic Regression	0.4369	0.4363	0.8000	0.3169	0.4540	0.0000	0.0000	0.7900

Hyperparameter tuning was then performed, followed by model prediction calculations on the holdout dataset results:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extreme Gradient Boosting	0.9649	0.9916	0.9444	0.9444	0.9444	0.9188	0.9188

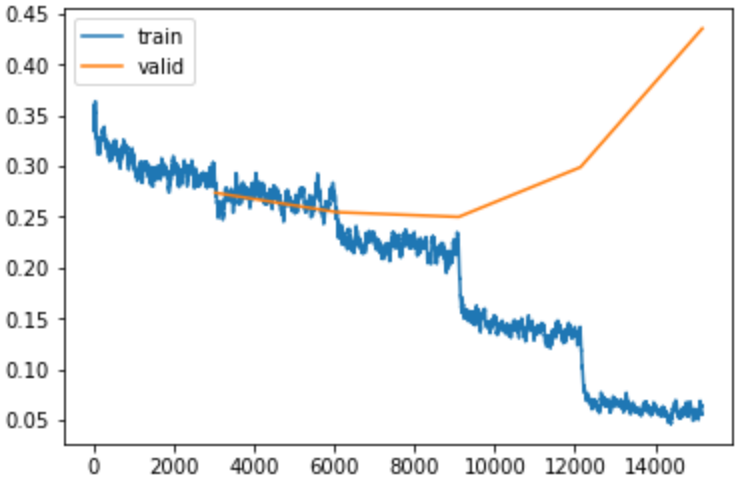
Deep Learning Model

This inspiration and figure were provided by Salah Sammari on his Kaggle page, which offers a concise and informative summary regarding Abstract Breast Cancer:



"Abstract Breast cancer is a common cancer in women, and one of the major causes of death among women around the world. Invasive ductal carcinoma (IDC) is the most widespread type of breast cancer with about 80% of all diagnosed cases. Early accurate diagnosis plays an important role in choosing the right treatment plan and improving survival rate among the patients. In recent years, efforts have been made to predict and detect all types of cancers by employing artificial intelligence." [1]

For this model, I chose the **fastai** library, which leverages **PyTorch** by wrapping many of the more tedious tasks in user-friendly wrapper methods. I first trained a *Resnet18* model to establish a baseline and trained for 5 epochs. The learning rate between the training and validation sets began to diverge after 3 epochs. For the next iteration, I trained for only 3 epochs, using a *ResNet34* model.



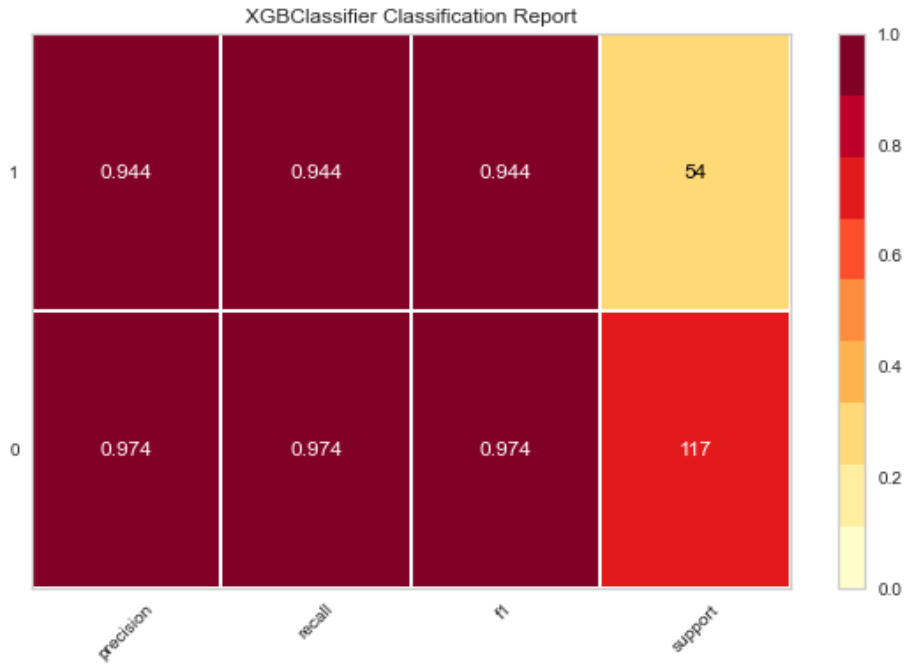
A good explanation of ResNets (shorthand for Residual Networks) and their many variants can be found here:

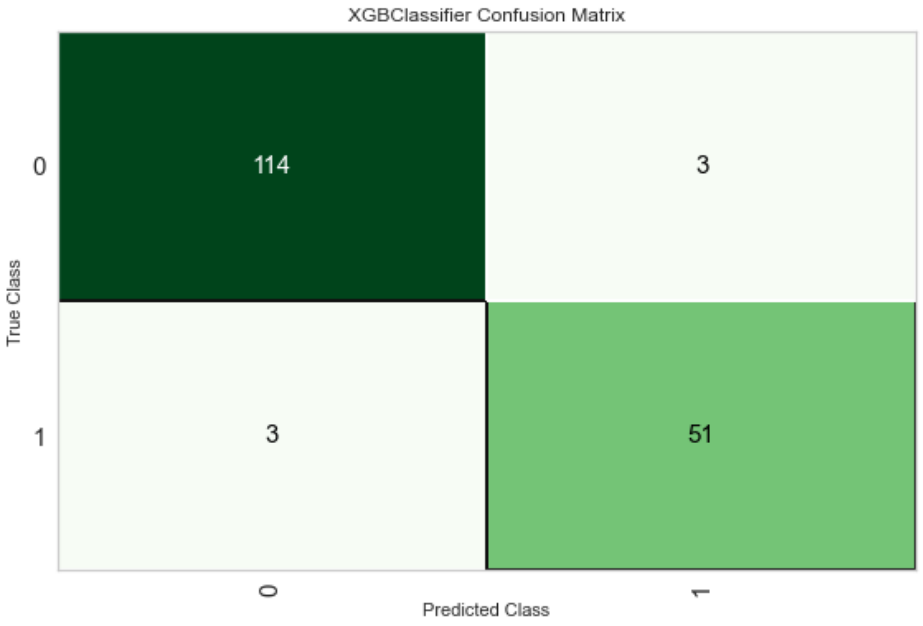
https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

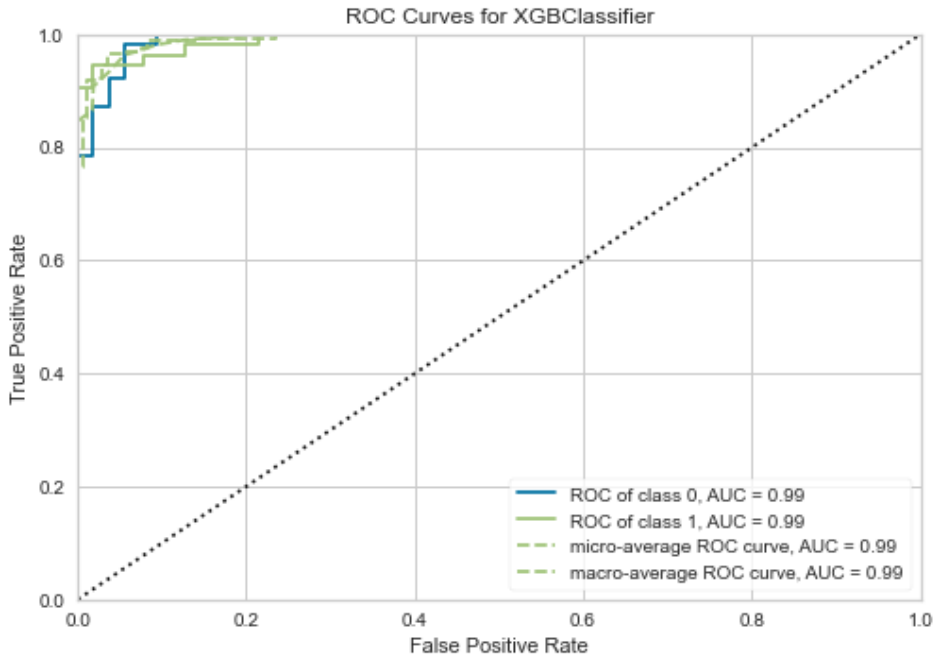
Results

Machine Learning

To reiterate, the purpose of training the Machine Learning model was to be able to predict, based on the measurement features in the dataset, whether an observation identified a benign or a malignant tumor. The **xgboost** model resulted in some impressive results.







Deep Learning

As previously mentioned, my focus was to train a model with the least rate of false negatives; at the very least, to do better than the human rate of 20%. By utilizing the fastai library, I was able to train a model with a false negative rate of about 9% (1 – Negative Predicted Value (NPV)).

Sensitivity: 0.7816111416839521
Specificity: 0.9318509575074622
PPV: 0.8196022727272727
NPV: 0.9150488234616081

Sensitivity & Specificity

Sensitivity or True Positive Rate is where the model classifies a patient has the disease given the patient actually does have the disease. Sensitivity quantifies the avoidance of false negatives
Example: A new test was tested on 10,000 patients, if the new test has a sensitivity of 90% the test will c orrectly detect 9,000 (True Positive) patients but will miss 1000 (False Negative) patients that have the c ondition but were tested as not having the condition

Specificity or True Negative Rate is where the model classifies a patient as not having the disease given the patient actually does not have the disease. Specificity quantifies the avoidance of false positives
Understanding and using sensitivity, specificity and predictive values is a great paper if you are intereste d in learning more about understanding sensitivity, specificity and predictive values.

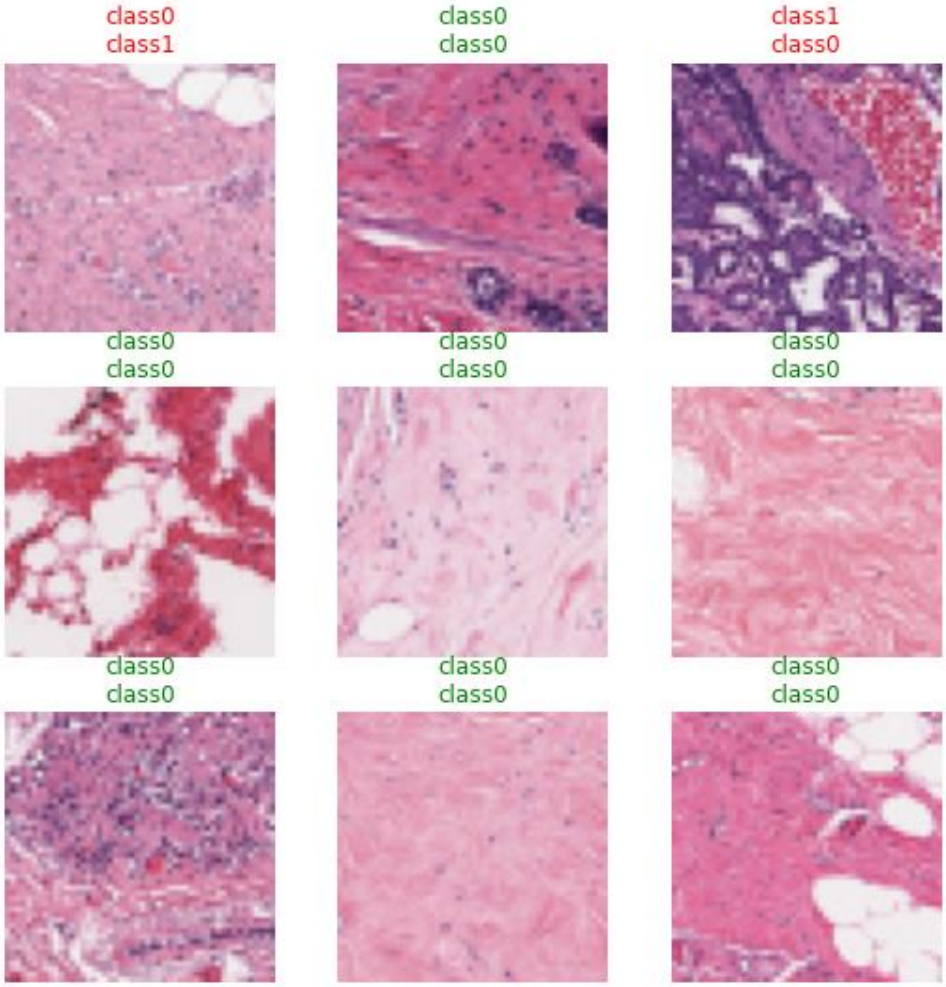
PPV and NPV
Most medical testing is evaluated via PPV (Positive Predictive Value) or NPV (Negative Predictive Value).

PPV - if the model predicts a patient has a condition what is the probability that the patient actually has the condition

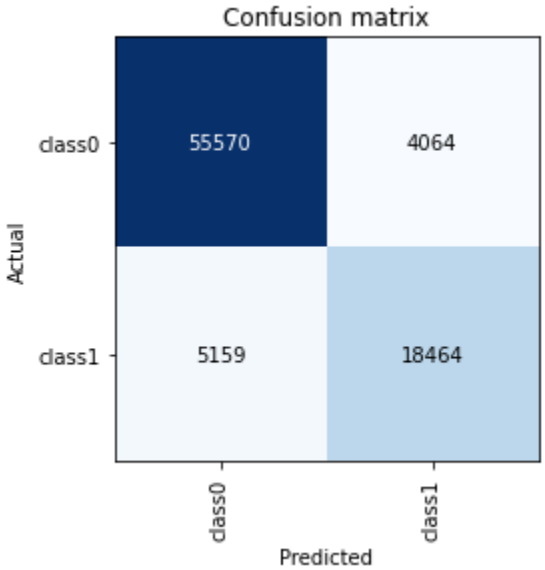
NPV - if the model predicts a patient does not have a condition what is the probability that the patient a ctually does not have the condition

The ideal value of the PPV, with a perfect test, is 1 (100%), and the worst possible value would be zero
The ideal value of the NPV, with a perfect test, is 1 (100%), and the worst possible value would be zero

Source: https://docs.fast.ai/tutorial.medical_imaging.html



In a random sampling of predictions, only one resulted in a false negative.



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?

Conclusions

TBD

References

[1] Sammari S. (2021) Breast Cancer Image Classification

https://www.kaggle.com/midouazerty/breast-cancer-images-classification

[2] National Cancer Institute (n.d.) Breast Cancer – Mammograms

https://www.cancer.gov/types/breast/mammograms-fact-sheet

[3] UCI Machine Learning (2016) Breast Cancer Wisconsin (Diagnostic) Data Set

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

[4] Mooney, P. (2019) Breast Histopathology Images

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