Breast Cancer Prediction

# **Bhushan Suryawanshi** [**DSC680-T302 Applied Data Science (2215-1)**](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_500232_1)

# **Domain and Background:**

Cancer is a disease that attacks thousands of people annually. Breast cancer is the second leading cause of death from cancer in women. The ability to quickly and accurately diagnose this disease greatly increases one’s ability to overcome and survive this affliction.

Currently, breast cancer is diagnosed by completing one or more the following:

* **Breast ultrasound:** A machine that uses sound waves to make detailed pictures, called sonograms, of areas inside the breast.
* **Diagnostic mammogram:** If you have a problem in your breast, such as lumps, or if an area of the breast looks abnormal on a screening mammogram, doctors may have you get a diagnostic mammogram. This is a more detailed X-ray of the breast.
* **Magnetic resonance imaging (MRI):** A kind of body scan that uses a magnet linked to a computer. The MRI scan will make detailed pictures of areas inside the breast.
* **Biopsy:** This is a test that removes tissue or fluid from the breast to be looked at under a microscope and do more testing. There are different kinds of biopsies (for example, fine-needle aspiration, core biopsy, or open biopsy).

Using machine learning to predict breast cancer can be utilized as an additional tool for early detection and prevention of complexities in cancer treatment.

# **Data Understanding:**

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. n the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is available through -

UCI Machine Learning Repository: [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

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

Attribute Information:

1) ID number  
2) Diagnosis (M = malignant, B = benign)  
3 to 32) 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. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

# Research Question:

Mammogram screenings are an excellent tool in identifying a potential tumorous mass, even before symptoms may appear. While imaging technology has increased resolution and clarity, interpretation of the data and the image as still exposed to potential false positive or false negative findings by breast imaging radiologists. Can predictive analytics be used to hone and improve cancerous tumor diagnosis? Data from the "Diagnostic Wisconsin Breast Cancer Database" can be analyzed to create a predictive model. The goal of this project is to examine the results of multiple features being utilized in providing a predictive value between benign and malignant breast tumors. In doing so, it will be another tool in the arsenal of early detection and minimizing the damaging effects of the disease.

# **Method:**

Data is available through Kaggle and the UCI machine learning repository. Having data from trusted sources ensures quality. However, the exploratory analysis will be performed to find the variability, skewness, or missing values in the dataset. Once the data is cleaned, feature sets will be analyzed for statistical summary and correlation. Feature’s distribution plot can help in identifying outliers. Identifying outliers will help in reducing overfitting issues.   
Based on statistical and correlation methods accompanied with feature selection techniques like PCA (Principe Component Analysis) or LassoCV can give us finalized set of features. These features will be used for model building.   
The data will be split into training and test set. This is a binary classification problem. The models that make more sense to use here are - Random Forest Classifier, Logistic Regression CV, K-Nearest Neighbor Classifier, and Support Vector Machines (SVM) with Support Vector Classifier. The comparative results will be presented as the final outcome of the study.

# Potential Issues:

There are multiple risks involved in getting the data science project completing successfully. In the larger organization setup, you encounter risks such as human resources, technology (hardware/software), poor direction from management, poor requirements, bad project planning, etc. However, for this course work, I already eliminated people risk by working individual. However, the following are some of the risks still expected and respective mitigation plans.

**1. Low Data Quality**  
Since I am using data from an online data source where I don’t have control over its availability, I have kept a copy downloaded. Also to verify data quality I will do some exploratory data analysis. In case this data is found to be lacking quality there is a breast cancer dataset available under sklearn.datasets (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\_breast\_cancer.html).   
**2. Wrong Model and Libraries**  
While working I may end up choosing the wrong model. However, for this project, I will be comparing different model performances for the dataset. This study in itself will eliminate the risk of using a single model.

# Concluding Remarks:

Breast cancer is one of the main reasons behind deaths due to cancer in women. The imaging technology has been evolved and a lot of deep learning models can be utilized to identify Breast cancer very early. The dataset shares information about tumor features, that were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. For each observation, there are ten features, which detail tumor size, density, texture, symmetry, and other characteristics of the cell nuclei present in the image. This rich data can be utilized to find a predictive value between benign and malignant breast cancer. Multiple models will be evaluated with this dataset to compare the performance. This comparison is the final deliverable of the project.

**References -**

1. American Cancer Society. (2020, January). How Common Is Breast Cancer? - <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>.

2. American Cancer Society. Cancer Facts & Figures 2020. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf>.

3. National Cancer Institute. (2018, January). BRCA Mutations: Cancer Risk and Genetic Testing. <https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet>.

4. Ak MF. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare (Basel). 2020;8(2):111. Published 2020 Apr 26. doi:10.3390/healthcare8020111 [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7349542/]

5. Raghupathi W., Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf. Sci. Syst. 2014;2:3. doi: 10.1186/2047-2501-2-3. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/]

6. Petri I., Kubicki S., Rezgui Y., Guerriero A., Li H. Optimizing energy efficiency in operating built environment assets through building information modeling: A case study. Energies. 2017;10:1167. doi: 10.3390/en10081167. [https://www.mdpi.com/1996-1073/10/8/1167#cite]

7. Jothi N., Rashid N.A., Husain W. Data mining in healthcare—A review. Procedia Comput. Sci. 2015;72:306–313. doi: 10.1016/j.procs.2015.12.145. [https://www.sciencedirect.com/science/article/pii/S1877050915036066]

8. Dai H., Cheng Z., Baii J. Breast cancer cell line classification and its relevance with breast tumor subtyping. J. Cancer. 2017;8:3131–3141. doi: 10.7150/jca.18457. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665029/]

9. Rizki A, Weaver VM, Lee SY, Rozenberg GI, Chin K, Myers CA. et al. A Human Breast Cell Model of Preinvasive to Invasive Transition. Cancer research. 2008;68:1378–87. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2662716/]

10. Kurebayashi J, Otsuki T, Kurosumi M, Yamamoto S, Tanaka K, Mochizuki M. et al. Isolation and characterization of a new human breast cancer cell line, KPL-4, expressing the Erb B family receptors and interleukin-6. British Journal of Cancer. 1999;79:707–17. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2362677/]