# Breast cancer detection applying machine learning


Trish Girmus

Bellevue University

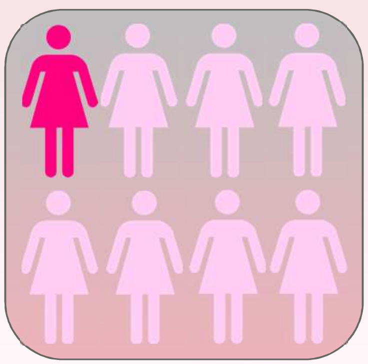
tgirmus@my365.bellevue.edu


## INTRODUCTION

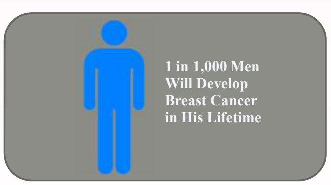
Breast cancer is the most common diagnosed cancer in women. It is also the second cause of cancer death with lung cancer being the first. Men are also susceptible to this disease, but the odds are fewer. An estimated 41,760 women will die this year from breast cancer as well as 500 men. When breast cancer is detected early at a localized stage, the five-year survival rate is 99%. Early detection and an accurate diagnosis are vital to increase survival rates. Mammogram screening is the primary technology for diagnosis and human error can result as doctors make these observations/diagnoses themselves. Machine learning classifiers can help doctors make accurate diagnoses quickly and affordably, thus an option to help prolong lives. The goal of this project is to identify how machine learning techniques can be applied to detect breast cancer and highlight several algorithms that might be best suited based on research.

## FIGURE 1 – Comparison of Women vs. Men in Development of Breast Cancer

**One in Eight Women Will Develop Breast Cancer in Her Lifetime**



**Men Can Also Be Diagnosed with Breast Cancer But Less Likely Than Women**




https://www.nationalbreastcancer.org/breast-cancer-facts

## FIGURE 2 – Probability of Breast Cancer Diagnosis or Death for US Women Based on Age

| Current age | Diagnosed with invasive breast cancer | Dying from breast cancer |
|---|---|---|
| 20 | 0.1% (1 in 1,479) | <0.1% (1 in 18,503) |
| 30 | 0.5% (1 in 209) | <0.1% (1 in 2,016) |
| 40 | 1.5% (1 in 65) | 0.2% (1 in 645) |
| 50 | 2.4% (1 in 42) | 0.3% (1 in 310) |
| 60 | 3.5% (1 in 28) | 0.5% (1 in 193) |
| 70 | 4.1% (1 in 25) | 0.8% (1 in 132) |
| 80 | 3.0% (1 in 33) | 1.0% (1 in 101) |
| **Lifetime risk** | **12.8% (1 in 8)** | **2.6% (1 in 39)** |

Note: Probability is among those who have not been previously diagnosed with cancer. Percentages and "1 in" numbers may not be numerically equivalent due to rounding.



https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

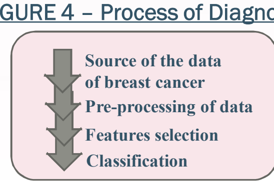
## FIGURE 3 – Increased Risk Ractors for Invasive Breast Cancer in Women

| Relative risk | Factor |
|---|---|
| >4.0 | Age (65+ versus <65 years, although risk increases across all ages until age 80) |
| | Atypical hyperplasia |
| | Lobular carcinoma in situ |
| | Pathogenic genetic variations (e.g. *BRCA1, BRCA2, PALB2, TP53*) |
| 2.1-4.0 | Ductal carcinoma in situ |
| | High endogenous hormone levels (postmenopausal) |
| | High-dose radiation to chest (e.g. Hodgkin lymphoma treatment) |
| | Mammographically dense breasts |
| | Two or more first-degree relatives with breast cancer |
| 1.1-2.0 | Alcohol consumption |
| | Early menarche (<11 years) |
| | Excess body weight |
| | High endogenous estrogen or testosterone levels (premenopausal) |
| | Late age at first full-term pregnancy (>30 years) |
| | Late menopause (≥55 years) |
| | Never breastfed a child |
| | No full-term pregnancies |
| | One first-degree relative with breast cancer |
| | Obesity (postmenopausal) |
| | Personal history of ovarian or endometrial cancer |
| | Physical inactivity |
| | Proliferative breast disease without atypia (usual ductal hyperplasia, fibroadenoma) |
| | Recent and long-term use of menopausal hormone therapy containing estrogen and progestin |
| | Recent hormonal contraceptive use |
| | Weight gain in adulthood |
| | Tall height |

Note: Relative risks for some factors vary by breast cancer molecular subtype.

https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

## FIGURE 4 – Process of Diagnostic

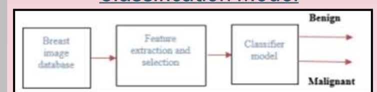



https://www.semanticscholar.org/paper/Proposed-approach-for-breast-cancer-diagnosis-using-Saoud-Ghadi/6fec944321a94b13f6671e6cc23d6fabc7ba58c3/figure/3

## DELIVERABLE

Machine learning (ML) is gaining momentum in the medical profession as it can help with predicting cancer susceptibility, survival rates and treatments. It creates and trains models which can be used for classification or prediction purposes. Machine learning could help with identifying and predicting cancer. The supervised classification model can determine whether a tumor is malignant or benign.

## FIGURE 5 – Breast Cancer Image Classification Model




http://www.ijmlc.org/vol9/794-L0201.pdf

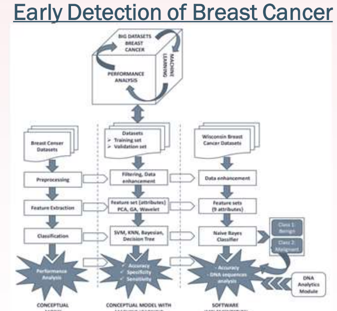
In the study using this classification model five non-linear machine learning algorithms were used: Multilayer perceptron (MLP), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Naïve Bayes (NB) and Support Vector Machines (SVM). The steps for this model were to measure the classification test accuracy, recall and precision. Using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, the study concluded the MLP classifier had the highest accuracy to classify a tumor as benign or malignant.

## FIGURE 6 – Accuracy Scores Using Five Classifiers

| | MLP | KNN | CART | NB | SVM |
|---|---|---|---|---|---|
| **Benign** | 97% | 88% | 90% | 90% | 95% |
| **Malignant** | 100% | 100% | 96% | 97% | 100% |
| **Average** | 99% | 96% | 94% | 95% | 98% |

http://www.ijmlc.org/vol9/794-L0201.pdf

## FIGURE 7 - Conceptual Model for Early Detection of Breast Cancer



https://aip.scitation.org/doi/pdf/10.1063/1.5014010

Steps for the conceptual model, steps include pre-processing, feature extraction, classification, and performance analysis. As in Figure 5, this model the classification step is important as it classifies whether the data is benign or malignant. Classifiers such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, and Decision Tree are examples of algorithms for this type of study.

## FIGURE 8 – NAÏVE BAYES Classifier Performance Results

| *Number of Attributes* | *Features sets* | *Accuracy* |
|---|---|---|
| 9 | {Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses} | 97,86 % |

https://aip.scitation.org/doi/pdf/10.1063/1.5014010

The Naïve Bayes classifier was used for this study and as Figure 8 shows, the accuracy was 97.86%

## WHY IS THIS DATA SCIENCE?

Machine learning and data science are providing efficiency, accuracy and giving professionals in the medical field a second opinion on the diagnosis of breast cancer disease. The use of machine learning algorithms can help identify, classify, and predict outcomes effectively. With the increase in breast cancer over the years, more data has been compiled which can be used for partnering together data science and medicine for further study and research.

## CONCLUSIONS

Breast cancer is a horrible disease and claimed the lives of many women as well as men. It is commonly detected by a doctor through a medical examination. Unfortunately, the human eye and other factors can cause a misdiagnosis. Early detection is vital to save lives as well as to increase survival rates of those who have been diagnosed with breast cancer. By utilizing machine learning and applying classification algorithms, doctors and other medical professionals can accurately diagnose this disease. This poster highlighted several different methods and common supervised learning models, showing the reliability and accuracy with prediction techniques.


## ACKNOWLEDGMENTS

I'd like to take a moment to acknowledge anyone who is currently diagnosed with cancer or had a loved one who lost the battle to this disease. My thoughts are with you. We will win this fight.

A sincere appreciation and thank you to my friends and family for their encouragement and support. I would also like to thank Professor Singh and the wonderful faculty at Bellevue University for the opportunity to make a difference in this world.