Breast Cancer Detection Applying Machine Learning

Trish Girmus
Omaha, NE, USA
tkgirmus@gmail.com

Abstract
Breast cancer disease affects 27% of women worldwide. It affects women physically as well as emotionally. Early detection is crucial for saving and prolonging lives. This topic has been on the radar of researchers more recently to address further research and prediction. With the availability of data, doctors could determine a quicker diagnosis with more precision and less human error. Not only does this save time, it also gives medical professionals another set of eyes to evaluate this information. Machine learning can perform many types of classification modeling with algorithms which learn from this data and calculate a prediction. This paper will highlight several different types of algorithms with high-performance accuracy rates and the outcomes of its results.

Author Keywords
Breast Cancer; Detection; Machine Learning; Classification; Algorithms; Artificial Intelligence

ACM Classification Keywords
•Computing methodologies•Computing methodologies~Machine learning•Computing methodologies~Machine learning~Machine learning approaches

Introduction

Breast cancer is the most common diagnosed cancer in women, affecting one out of eight women [Figure 1]. A definition of breast cancer is a group of diseases in which cells in breast tissue change and divide uncontrolled, typically resulting in a lump or mass. [22] It is also the second cause of cancer death with lung cancer being the first. [4] Men are also susceptible to this disease, but the odds are fewer with one in one thousand developing it. [21] An estimated 41,760 women will die this year from breast cancer as well as 500 men. Various, but not all reasons for developing breast cancer include lack of exercise, alcohol consumption, hormones, family history, and genetics. See Figure 2 for a summary of increased risk factors for breast cancers.

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.

Figure 1 – Probability of Breast Cancer Diagnosis or Death for US Women Based on Age

When breast cancer is detected early, the five year-survival rate percentages for the following are: [22]

- A localized stage is 99%
- A regional stage is 86%
- Diagnosis of metastatic disease is 27%

According to the American Cancer Society, death rates from breast cancer have declined 40% through 2017. Early detection and treatment improvements are attributed to this. [22] It is also important to note that not all breast lumps are cancerous; however, it can increase a women’s risk of developing breast cancer. [9] These non-cancerous (benign) tumors are not life-threatening and do not spread outside of the breast. At the early stage of breast cancer, a tumor small in size is the easiest to treat and has the best outcome. If no symptoms are present, it difficult to diagnose without screening. Early signs of breast cancer include architectural distortion, mass, microcalcification and breast asymmetries. [2]

To detect breast cancer, there are two strategies involved. These are early diagnosis and screening. [4] Screening detection techniques used for medical imaging include mammography, magnetic resonance imaging (MRI), computed tomography (CT) scans, biopsy (costly but effective), fine needle aspiration cytology, ultrasound, nuclear imaging, etc. [2,3,5] Images are produced which are analyzed to detect a benign or malignant pattern. Mammogram screening is the primary technology for diagnosis. However, if breast tissue is dense, tumors may hide and not be visible which can result in a high false negative/positive rate. [2] Another disadvantage of using mammography

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.

Figure 2 – Increased Risk Factors for Invasive Breast Cancer in Women

is x-ray radiation, which can also cause cancer. Other human error can result including inter-operation and intra-operation difference, fatigue, and experience of the doctor. [3] Computer aided diagnosis (CAD) with machine learning units can help radiologists predict and determine the prognosis, accurately, efficiently, and effectively. This also gives a second opinion when diagnosing this disease.

Methods

This section discusses several machine learning algorithms that favored high accuracy with detecting and predicting breast cancer.

Machine learning (ML) is gaining momentum in the medical profession as it can help with predicting cancer susceptibility, survival rates and treatments. [4] It creates and trains models which can be used for classification or prediction purposes. Machine learning could help with identifying and predicting cancer. In figure 3, the supervised classification model can determine whether a tumor is malignant or benign.

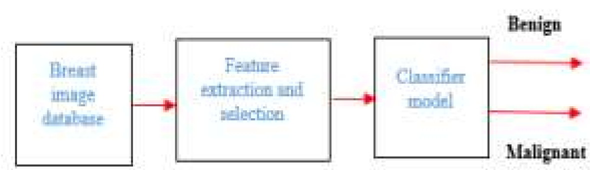

Figure 3 – Breast Cancer Image Classification Model [4]

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. [4] Using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset (obtained from the University of California, Irvine), the study concluded the MLP classifier had the highest accuracy to classify a tumor as benign or malignant.

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

Figure 4 – Accuracy Scores Using Five Classifiers

Figure 5 is an example of a conceptual model that was used in another study for early detection of breast cancer.

Steps for the conceptual model, steps include pre-processing, feature extraction, classification, and performance analysis. As in Figure 3, 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.

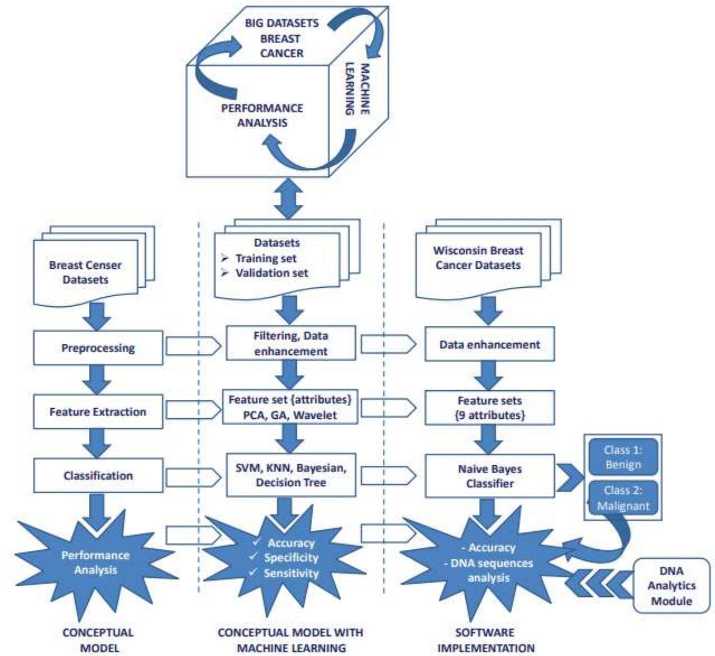

Figure 5 -Conceptual Model for Early Detection of Breast Cancer based on Machine Learning [9]

The Naïve Bayes classifier was used for this study and as Figure 6 shows, the accuracy was 97.86%

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 %

Figure 6 – Naïve Bayes Classifier Performance Results [9]

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.

Conclusion

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 paper explored several different methods and common supervised learning models, showing the reliability and accuracy with prediction techniques.

Acknowledgements

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.

References

[1] Abien Fred M. Agarap. 2018. On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. *ICMLSC '18: Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, (February 2018), 5-9. DOI: https://doi.org/10.1145/3184066.3184080

[2] Fayez AlFayez, Mohamed W. Abo El-Soud, Tarek Gaber. 2020. Thermogram breast cancer detection: A comparative study of two machine learning techniques. *Applied Sciences*, Vol. 10, 2, 551 (January 2020), 20 pages. DOI: https://doi.org/10.3390/app10020551

[3] Kushangi Atrey, Yogesh Sharma, Narendra K.Bodhey, & Bikesh Kumar Singh. 2019. Breast cancer prediction using dominance-based feature filtering approach: A comparative investigation in machine learning archetype. *Brazilian Archives of Biology and Technology*, (Epub November 25, 2019), *62*, e19180486. https://doi.org/10.1590/1678-4324-2019180486

[4] Ali Al Bataineh. 2019. A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. *International Journal of Machine Learning and Computing*, Vol. 9, No. 3 (June 2019), 248-254. DOI: https://doi: 10.18178/ijmlc.2019.9.3.794

[5] Sumaiya Dabeer, Maha Mohammed Khan, Saiful Islam. 2019. Cancer diagnosis in histopathological image: CNN based approach. *Informatics in Medicine Unlocked,* Vol. 16 (2019). DOI: https://doi.org/10.1016/j.imu.2019.100231

[6] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled & Aly Fahmy. 2019. Dataset of breast ultrasound images. *Data in brief*, (November 21, 2019), *28*, 104863. https://doi.org/10.1016/j.dib.2019.104863

[7] Syed Jamal Safdar Gardezi, Ahmed Elazab, Baiying Lei & Tianfu Wang. 2019. Breast cancer detection and diagnosis using mammographic data: Systematic review. *Journal of medical Internet research*, (July 26, 2019), *21*(7), e14464. https://doi.org/10.2196/14464

[8] Min-Wei Huang, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, & Chih-Fong Tsai. 2017. SVM and SVM ensembles in breast cancer prediction. *PloS one*, *12*(1), e0161501 (January 6, 2017). https://doi.org/10.1371/journal.pone.0161501

[9] Desislava Ivanova. 2017. Big data analytics for early detection of breast cancer based on machine learning. *AIP Conference Proceedings*, Vol. 1910, Issue 1, (2017), 1-8. DOI: https://doi.org/10.1063/1.5014010

[10] Seyed Reza Kamel, Reyhaneh YaghoubZadeh& Maryam Kheirabadi. 2019. Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer. Journal of Big Data. 2019;6(1):1-15. https://doi:10.1186/s40537-019-0247-7

[11] Neha Kumari, Khushboo Verma. 2019. A survey on various machine learning approaches used for breast cancer detection. *International Journal of Advanced Research in Computer Science; Udaipur*, Vol. 10, No. 3 (May 2019), 76-79. DOI: http://dx.doi.org/10.26483/ijarcs.v10i3.6438

[12] Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O. Chappuis, Ivo D. Dinov & Maria C. Katapodi. 2019. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. *Breast cancer research: BCR*, (June 20, 2019), *21*(1), 75. https://doi.org/10.1186/s13058-019-1158-4

[13] Bibhuprasad Sahu, Sachi Nandan Mohanty, Saroj Kumar Rout. 2019. A hybrid approach for breast cancer classification and diagnosis. *EAI Endorsed Transactions on Scalable Information Systems*, (January 2019). DOI: https://doi:10.4108/eai.19-12-2018.156086

[14] Bibhuprasad Sahu, Amrutanshu Panigrahi. 2020. Efficient role of machine learning classifiers in the prediction and detection of breast cancer. *5th International Conference on Next Generation Computing Technologies (NGCT-2019)*, February 27, 2020. http://dx.doi.org/10.2139/ssrn.3545096

[15] Zakia Salod, Yashik Singh. 2019. Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. *Journal of Public Health Research*, Vol. 8, 3: 1677 (Dec. 2019). DOI: https://doi:10.4081/jphr.2019.1677

[16] Hajar Saoud, Abderrahim Ghadi, & Mohamed Ghailani. 2019. Proposed approach for breast cancer diagnosis using machine learning. In Proceedings of the 4th International Conference on Smart City Applications (SCA '19). Association for Computing Machinery, New York, NY, USA, Article 102, 1–5. https://doi.org/10.1145/3368756.3369089

[17] Jenni A M Sidey-Gibbons & Chris J. Sidey-Gibbons. 2019. Machine learning in medicine: a practical introduction. *BMC medical research methodology*. (March 9, 2019), *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4

[18] Gigi F. Stark, Gregory R. Hart, Bradley J. Nartowt & Jun Deng. 2019. Predicting breast cancer risk using personal health data and machine learning models, *PloS one*, (December 27, 2019), *14*(12), e0226765. https://doi.org/10.1371/journal.pone.0226765

[19] Seid Hassen Yesuf. 2019. Breast cancer detection using machine learning techniques. *International Journal of Advanced Research in Computer Science; Udaipur*, Vol. 10, 5 (September 2019), 27-33. DOI: https://doi.org/:10.26483/ijarcs.v10i5.6464

[20] Li Yixuan, Zixuan Chen. 2018. Performance evaluation of machine learning methods for breast cancer prediction, *Applied and Computational Mathematics*. Vol. 7, (November 4, 2018), pp. 212-216. https://doi: 10.11648/j.acm.20180704.15

[21] National Breast Cancer Foundation, Inc. 2019. Retrieved from https://www.nationalbreastcancer.org/

[22] American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. 2019. Retrieved from 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