HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

# Training an Artificial Intelligence Model to Predict Breast Cancer Recurrence

Avi Shah

Viewpoint School, USA

## ABSTRACT

Predicting cancer recurrence in patients with breast cancer is challenging. This study aimed to train and use an Artificial Intelligence (AI) model to predict breast cancer recurrence. The model successfully predicted recurrence versus no recurrence in 92.94% of patients. The three traits at presentation that correlated most to recurrence were positive ovarian status, negative human epidermal growth factor 2 receptor status, and negative estrogen receptor status. AI models can predict cancer recurrence and may become a useful tool in the management of cancer.

## Introduction

In the United States, breast cancer is diagnosed in 1 in every 8 women and is the second leading cause of cancerous death in women[1]. Recurrence of breast cancer is associated with higher mortality. In one study, women with early recurrence of breast cancer had a ten-year cumulative mortality rate of 72%[2]. Identifying patients at higher risk of recurrence and modifying their treatment may help reduce mortality rates.

AI systems have been developed to improve cancer detection in mammography screenings and MRI screenings[3]. Many FDA-approved software tools utilize AI to help improve cancer detection in cancer screening imaging[4]. However, the use of machine learning to predict future cancer recurrences is still in the early stages of development.

The goal of this study was to develop an artificial intelligence model that can analyze breast cancer patients and accurately predict recurrence. The Duke Breast Cancer MRI dataset published in 2018 was used to train and test an AI algorithm created with Weka software to predict breast cancer recurrence.

## Methods

### Weka

Weka is a data mining system that uses artificial intelligence algorithms to complete statistical tasks. This program was developed by the University of Waikato in New Zealand. This program can be applied directly to a dataset to perform several functions including classification, regression, and association[5].

### Database

The database that was used for this study was published by Duke in 2018 and consists of breast cancer patients with clinical data and MRI scans[6]. Specific patient attributes were used for the algorithm (Table 2). Because Weka software requires every data point to be complete, patients with missing data for these attributes were excluded.

### Algorithm, Pruning, and Procedure

A pruned dataset with randomly selected patients with and without recurrence was used for this study based on Weka software requirements. The dataset consisted of 86 patients without recurrence and 84 patients who experienced a recurrence of their cancer. The specific algorithm selected was a random forest classifier. It was trained and tested with the data by means of the standard 10-fold cross validation. Weka was also used to rank attributes based on correlation to recurrence. The program assigns a merit score to each attribute based on its predictive ability and its degree of redundancy.

Random Tree and J48 algorithms were also tested, but these resulted in reduced accuracy. For cross validation, 8-fold and 12-fold were tested, but these also had lower accuracy. In addition, the dataset prior to pruning was tested, but this also resulted in poor accuracy because of overfitting.

## Results

The model was able to predict recurrence or no recurrence with an overall accuracy of 92.94% (Table 1). The sensitivity was 86.9%. The specificity was 98.8%.

For further data analysis, the dataset was divided into two groups: presentation and treatment. This was done to compare the models' prediction of recurrence between only using patient attributes at presentation and only using the treatments received by each patient. The following were the attributes selected for the presentation group: human epidermal growth factor 2 receptor status (HER2), estrogen receptor status (ER), progesterone receptor status (PR), lymph node involvement based on pathologic/clinical findings, skin/nipple involvement in MRI imaging, presence or absence of ovaries, tumor molecular subtype (luminal-like, ER/PR positive and HER2 positive, ER/PR negative and HER2 positive, or triple negative), metastatic disease at presentation, lymphadenopathy or suspicious lymph nodes in MRI imaging, tumor size (T1 is < 2 cm, T2 is between 2 and 5 cm, T3 is > 5 cm, T4 is a tumor growing into the chest wall), multifocal versus multicentric tumor in MRI imaging, contralateral breast involvement in MRI imaging, pectoral/chest involvement in MRI imaging, age, mitotic activity grade on tumor pathology, tubule presence grade on tumor pathology, and nuclear abnormality grade on tumor pathology. The following were the attributes selected for the treatment group: received neoadjuvant (before surgery) therapy, neoadjuvant chemotherapy, neoadjuvant radiation therapy, adjuvant (after surgery) anti-Her2 neu therapy, neoadjuvant anti-Her2 neu therapy, adjuvant radiation therapy, therapeutic or prophylactic oophorectomy, adjuvant chemotherapy, neoadjuvant endocrine therapy medications, and adjuvant endocrine therapy medication.

### Model Accuracy

| | Overall Accuracy (%) | # of Correctly Predicted Recurrence Cases | # of Correctly Predicted Non-Recurrence Cases | False Negatives | False Positives |
|---|---|---|---|---|---|
| All Attributes | 92.94% | 73/84 | 85/86 | 11 | 1 |
| Presentation Attributes | 70.00% | 49/84 | 70/86 | 35 | 16 |
| Treatment Attributes | 77.06% | 66/84 | 65/86 | 18 | 21 |

Using a correlation evaluation ranker in Weka, it was determined that these attributes were most correlated to recurrence: received neoadjuvant therapy, neoadjuvant chemotherapy, therapeutic or prophylactic oophorectomy, adjuvant anti-Her2 neu therapy, and adjuvant chemotherapy (Table 2).

Overall Attribute Correlation

| Correlation Merit Score | Attribute |
|---|---|
| 0.5704 | Received Neoadjuvant Therapy |
| 0.5221 | Neoadjuvant Chemotherapy |
| 0.4416 | Therapeutic or Prophylactic Oophorectomy |
| 0.4205 | Adjuvant Anti-Her2 Neu Therapy |
| 0.389 | Adjuvant Chemotherapy |
| 0.3469 | Positive HER2 Status |
| 0.3444 | Tumor Size (T stage) |
| 0.3207 | Positive Ovarian Status |
| 0.2975 | High Lymph Node Involvement |
| 0.2722 | High Lymphadenopathy or Suspicious Nodes on MRI |
| 0.2598 | PR Negative |
| 0.2588 | High Skin/Nipple Involvement on MRI |
| 0.2363 | High Tubule Pathology Grade |
| 0.2056 | Adjuvant Endocrine Therapy Medications |
| 0.1935 | Neoadjuvant Anti-Her2 Neu Therapy |
| 0.1824 | High Mitotic Activity Pathology Grade |
| 0.1746 | ER Negative |
| 0.1705 | High Nuclear Pathology Grade |
| 0.1356 | Metastatic at Presentation |
| 0.1356 | Pectoral/Chest Involvement on MRI |
| 0.1147 | Triple Negative Molecular Subtype |
| 0.1142 | Age |
| 0.1104 | Neoadjuvant Endocrine Therapy Medications |

| | |
|---|---|
| 0.0898 | Adjuvant Radiation Therapy |
| 0.0547 | Contralateral Breast Involvement |
| 0.0422 | Metastasis on MRI |
| 0.0347 | Multicentral/Multifocal Tumor on MRI |
| 0 | Neoadjuvant Radiation Therapy |

With training and testing on presentation attributes, the model predicted an overall accuracy of 70.00% (Table 1). The sensitivity was 58.3%. The specificity was 75.6%. The following attributes had the strongest correlation: positive ovarian status, positive human epidermal growth factor 2 receptor status, and negative estrogen receptor status (Table 3).

## Presentation Attribute Correlation

| Correlation Merit Score | Attribute |
|---|---|
| 0.4102 | Positive Ovarian Status |
| 0.3261 | Positive HER2 Status |
| 0.166 | ER Negative |
| 0.1653 | High Skin/Nipple Involvement on MRI |
| 0.164 | High Lymph Node Involvement |
| 0.1578 | Triple Negative Tumor Molecular Subtype |
| 0.1356 | Metastatic at Presentation |
| 0.0978 | Metastasis on MRI |
| 0.0936 | High Lymphadenopathy or Suspicious Nodes on MRI |
| 0.0698 | Multicentral/Multifocal Tumor on MRI |
| 0.0597 | Tumor Size (T stage) |
| 0.0577 | PR Negative |
| 0.0377 | Contralateral Breast Involvement |
| 0.0369 | Pectoral/Chest Involvement on MRI |
| 0.0363 | Age |

| | |
|---|---|
| 0.0362 | High Mitotic Activity Pathology Grade |
| 0.0246 | High Tubule Pathology Grade |
| 0.0181 | High Nuclear Pathology Grade |

With training and testing on administered treatment attributes, the model predicted an overall accuracy of 77.06% (Table 1). The sensitivity was 78.6%. The specificity was 81.4%. The following attributes had the highest correlation: received neoadjuvant therapy, neoadjuvant chemotherapy, and therapeutic or prophylactic oophorectomy (Table 4).

**Table 4.** Treatment Attribute Correlation

| Correlation Merit Score | Attribute |
|---|---|
| 0.5888 | Received Neoadjuvant Therapy |
| 0.4483 | Neoadjuvant Chemotherapy |
| 0.3043 | Therapeutic or Prophylactic Oophorectomy |
| 0.2722 | Adjuvant Chemotherapy |
| 0.1958 | Neoadjuvant Endocrine Therapy Medications |
| 0.1301 | Adjuvant Endocrine Therapy Medications |
| 0.1286 | Adjuvant Radiation Therapy |
| 0.1265 | Neoadjuvant Anti-Her2 Neu Therapy |
| 0.011 | Adjuvant Anti-Her2 Neu Therapy |
| 0 | Neoadjuvant Radiation Therapy |

## Discussion

This AI model executed using the Weka software performed well at predicting recurrence versus no recurrence in this database of patients with breast cancer. The specificity (98.8%) was excellent, while the sensitivity (86.9%) was not quite as high. While this algorithm performed best using all patient attributes (92.94% accuracy), the accuracy using only patient attributes at presentation was 70.00%. This is an important part of the model to improve on because clinically, AI models would be most useful if they could use presentation attributes to predict cancer recurrence risk. Other studies that have used AI modeling to try to predict breast cancer recurrence have published a large range of accuracy results: from 65% - 97%[7]. As AI models improve and as more extensive cancer databases become available, AI may become an important resource for optimizing and individualizing cancer treatment.

Interestingly, the characteristics with the highest merit score related to tumor recurrence were all treatment-based. This could be because the patients' oncologists recognized patients with more aggressive tumor characteristics and prescribed more aggressive treatments, such as neoadjuvant chemotherapy and prophylactic oophorectomy.

It will be interesting to see the results of improved software in the future. The database used in our study also contained MRI images of each patient. The Weka software used for this AI model does not have the capability to utilize MRI images. It would be interesting to see an algorithm that could incorporate patient attributes as well as their MRI scans.

Another limitation of this study was the size of the database. As seen in this study, when more attributes are available to the algorithm, results are better. When more complex databases with more data points become available, AI may become more useful.

## Conclusion

AI models can predict cancer recurrence and may become a useful tool in the management of cancer.

## Acknowledgments

This project was inspired by my mother, who was recently diagnosed with breast cancer. Her strength during her fight with cancer was inspiring and I am so grateful that she was able to overcome it.

## References

1. R. L. Siegel, A. N. Giaquinto, A. Jemal. (2024). Cancer statistics, 2024. CA Cancer J Clin. 74, 12-49. https://doi.org/10.3322/caac.21820
2. R. N. Pedersen, L. Mellemkjær, B. Ejlertsen, M. Nørgaard, D. P. Cronin-Fenton, 2022. Mortality after late breast cancer recurrence in Denmark. J Clin Oncol. 40, 1450-1464. https://doi.org/10.1200/JCO.21.02062
3. M. Cè, E. Caloro, M. E. Pellegrino, M. Basile, A. Sorce, D. Fazzini, G. Oliva, M. Cellina, 2022. Artificial intelligence in breast cancer imaging: risk stratification, lesion detection and classification, treatment planning and prognosis-a narrative review. Explor Target Antitumor Ther. 3, 795-816. https://doi.org/10.37349/etat.2022.00113
4. K. C. Potnis, J. S. Ross, S. Aneja, C. P. Gross, I. B. Richman, 2022. Artificial intelligence in breast cancer screening: evaluation of FDA device regulation and future recommendations. JAMA Intern Med. 182, 1306-1312. https://doi.org/10.1001/jamainternmed.2022.4969
5. E. Frank, M. A. Hall, I. H. Witten, 2016. The WEKA workbench online appendix for data mining: practical machine learning tools and techniques (4th edition), Morgan Kaufmann.
6. A. Saha, M. R. Harowicz, L. J. Grimm, C. E. Kim, S. V. Ghate, R. Walsh, M. A. Mazurowski, 2018. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. British journal of cancer. 119, 508-516. https://doi.org/10.1038/s41416-018-0185-8
7. C. Mazo, C. Aura, A. Rahman, W. M. Gallagher, C. Mooney, 2022. Application of artificial intelligence techniques to predict risk of recurrence of breast cancer: a systematic review. J Pers Med. 12, 1496-1517. https://doi.org/10.3390/jpm12091496