Chapter 61
Prediction of Breast Cancer Recurrence With Machine Learning

Mohammad Mehdi Owrang O.
American University, USA

Ginger Schwarz
American University, USA

Fariba Jafari Horestani
American University, USA

ABSTRACT

Medical prognostication is the science of estimating the complication and recurrence of a disease. A Breast cancer recurrence (BCR) event is characterized by the cancer "coming back" after at least a year of remission after the treatment. Many factors, including tumor grade, tumor size, and lymph node status may influence or correlate with prognosis for breast cancer patients. Early detection of recurrence events (i.e., while still asymptomatic) is more likely to be curable than after the cancer symptoms are seen again. Machine learning techniques can help to provide some necessary information and knowledge required by physicians for accurate predictions of BCR and better decision-making. The aim of this chapter is to use machine learning classifiers to examine the factors that are most predictive of the BCR. Several attributes/features selection schemes have been used to find the most significant features contributing to BCR. Five different machine learning algorithms were tested and compared for the prediction of BCR. The decision tree was found to be the best model for the dataset.

INTRODUCTION

Breast cancer is the most common female cancer in the U.S., the second most common cause of cancer death in women ("American Cancer Society", 2021), and the main cause of death in women ages 40 to 59 (Siegel et al., 2012). In 2023, an estimated 297,790 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 55,720 new cases of non-invasive (in situ) breast cancer ("Cancer.net", 2023). In 2023, about 43,170 women in the U.S. were expected to die from breast cancer.

DOI: 10.4018/978-1-6684-7366-5.ch061

There is a one in six probability that a woman will develop breast cancer in their lifetime (one in eight for invasive disease) ("American Cancer Society", 2021; Siegel et al., 2012). Breast cancer incidence is increasing while mortality is declining in many high-income countries (Torre et al., 2016). In recent years, great progress has been made in the understanding of breast cancer, with new classification techniques that have significant prognostic value and provide guides to treatment options (Ahmed et al., 2013; Lafourcade et al., 2018).

Medical prognostication is an essential part of medicine that encompasses the science of estimating the complication and recurrence of disease and predictive survival of patients (Ohno, 2001). Medical prognosis plays an increasing role in health care outcomes. Many factors, including tumor grade, tumor size, and lymph node status may influence or correlate with prognosis for breast cancer patients (Bradley, 2007; Lafourcade et al., 2018).

Early detection of breast cancer allows doctors to treat the disease aggressively giving women a better chance of survival. After the diagnosis and treatment of the primary incidence of breast cancer, the cancer may come back (recurrence event) after a year in the same organ (local recurrence), or in a close-by-organ (reginal recurrence) or in another part of the body (distant recurrence) (Fan et al., 2010; Ahmad et al., 2013; Guo et al., 2017; Alva, 2018; Lafourcade et al., 2018). However, local, or distant recurrence can occur at any time. In cancer, relapse-free survival is the length of time after primary treatment for a cancer that the patient survives without any signs or symptoms of that cancer. It is one way to measure how well a new treatment works.

Because of the high prevalence and mortality of breast cancer among women, it is important and necessary to explore and develop new techniques to help predict a patient's chance for recurrence and develop a treatment process to prevent it. A significant topic in breast cancer research is understanding and improving the effectiveness of breast cancer treatments thus increasing relapse-free survival time. In addition, the practical application of data mining can help to provide some necessary information and knowledge required by physicians for accurate prediction of Breast Cancer Recurrence (BCR) and better decision-making.

Current studies reported in the literature have utilized machine learning/data mining techniques to predict breast cancer recurrence, and to identify asymptomatic women at risk of cancer recurrence (Mannell, 2017; Mulatu & Gangardle, 2017; Mosayebi et al., 2020; Lou et al., 2020; Massafra et al., 2021). However, it is still an open task which requires the gathering of suitable and quality datasets, using proper features selection schemes, and defining more accurate prediction models.

To predict the probability of recurrence, it is important to know which features/factors of a patient have the highest risk factors (Chang & Lin, 2008; Mannell, 2017; Darst et al., 2018; Lafourcade et al., 2018). This chapter presents the preliminary results of a feature important analysis collected by patient records aimed to develop an automated system to predict BCR. The goal of this study is to use machine learning classifiers to examine the factors that are most predictive of the BCR. The main objective is to compare different data mining/machine learning algorithms to select the most accurate model for predicting BCR.

BREAST CANCER RECURRENCE OVERVIEW

Medical prognosis is a field in medicine that includes estimating the complication and recurrence of disease and predicting survival of a patient (Bradley, 2007; Ohno-Machado, 2001). Survival analysis is

a field in medical prognosis that deals with the application of various methods to estimate the survival of a patient suffering from a disease.

Breast Cancer Recurrence Risk Factors

Breast cancer recurrence (or recurrent breast cancer) describes breast cancer that comes back after treatment (mastectomy, chemotherapy, or other treatments) (Fan, 2010; Ahmad, 2013; Alva, 2018; Lafourcade et al., 2018). The goal of cancer treatments is to kill cancer cells. Treatments can reduce tumors so much that tests don't detect their presence. These weakened cells can remain in the body after treatment. Over time, they start to grow and multiply again. Recurrent breast cancer may occur months or years after initial treatment (Lou et al., 2020).

For breast cancer patients, factors that increase the risk of a recurrence include (Lafourcade et al., 2018; Mannell, 2017; "Mayoclinic.org", n.d.):

- Lymph node involvement: Finding cancer in nearby lymph nodes at the time of the original diagnosis.
- Larger tumor size.
- Positive or close tumor margins: "During breast cancer surgery, the surgeon tries to remove the cancer along with a small amount of the normal tissue that surrounds it. A pathologist examines the edges of the tissue to look for cancer cells. If the borders are free of cancer when examined under a microscope, that's considered a negative margin. If any part of the border has cancer cells (positive margin), or the margin between the tumor and normal tissue is close, the risk of breast cancer recurrence is increased", (Mayoclinic.org, n.d.).
- Lack of radiation treatment following a lumpectomy: For treatment, most patients choose lumpectomy and radiation therapy to reduce the risk of recurrence. The risk of local BCR is higher without radiation therapy.
- Younger age: Particularly those under age 35 at the time of their original breast cancer diagnosis.
- Inflammatory breast cancer: Patients with inflammatory breast cancer have a higher risk of local recurrence.
- Lack of endocrine therapy for hormone receptor-positive breast cancer: Patients who are not receiving endocrine therapy have a high risk for BCR.
- Cancer cells with certain characteristics: Triple Negative Breast Cancer (TNBC) patients may have a higher risk if the breast cancer does not respond to hormone therapy or treatments directed at the HER2 gene.
- Obesity: Patients with higher body mass index.

Initially, BCR was thought to happen within the first five years after treatment (Guo, 2017). However, recent studies show that the risk of recurrence can last more than thirty years (Lou, 2020). For late BCR, patients who had primary tumors larger than 20 mm, lymph node-positive disease, and estrogen receptor-positive tumors are at higher risk for late recurrence. Recurrent breast cancer is most likely to appear in patients who had an original tumor with more than three positive lymph nodes.

BREAST CANCER RECURRENCE RELATED WORKS

Many studies have been reported in the literature suggesting the use of machine learning and data mining techniques in predicting breast cancer recurrence (Ahmad et al., 2013; Pritom et al., 2016; Mannell, 2017; Lafourcade et al., 2018; Roberto et al., 2020; Lou et al., 2020; Yang et al., 2021; 2021; Massafra et al., 2021). These research studies had different goals for the BCR predictions including identifying risk factors for BCR, modeling BCR using small sample size in the Wisconsin breast cancer dataset, comparing machine learning algorithms for BCR prediction, and prediction for early and late BCR. These studies are reviewed next.

There have been several studies to identify risk factors for BCR. In (Mannell, 2017), authors examined pathogenic pathways leading to Breast Cancer (BC) and suggested modifications to lifestyle, surgical procedures and treatment plans which could reduce the recurrence of BC. Patient's factors associated with increased risk included the extremes of age, ethnicity, genetic inheritance, obesity, and alcohol ingestion. Although the incidence of BC in young women is small, age at diagnosis had a significant impact on outcome. Young age is a risk factor for LRR (Locoregional Recurrence (Shikama et al., 2011)) and death from BC. In another study, authors tried to identify risk factors by characterizing the presence of breast cancers' receptors, including ER, PR, HER2 (Human Epidermal Growth Factor Receptor 2) and TNBC, each subtype would have a higher risk of recurrence than others (Chacon & Costanzo, 2010; Roberto et al., 2020).

The research in (Mosayebi et al., 2020) compared different data mining algorithms to select the most accurate model for predicting breast cancer recurrence. A total of 5,471 independent records were available on the dataset. Authors found some features such as the LN (Lymph Node) involvement rate, HER2 value, Tumor size, and free or closed tumor margin to be the most important features in the dataset to predict breast cancer recurrence. Results show that the C5.0 algorithm possibly could be a helpful tool for the prediction of breast cancer recurrence at the stage of distant recurrence and nonrecurrence, especially in the first to third years.

There have been several studies to predict BCR using data mining and machine learning algorithms using small sample size in Wisconsin breast cancer dataset (Pritom et al., 2016; Alva, 2018; Lachman, 2019; Massafra et al., 2021), containing 286 records (and 10 attributes) of patients. The set of data has 201 records of patients who did not show breast cancer recurrence and 85 records in which there is evidence of recurrence. In (Pritom et al., 2016), authors applied Naive Bayes, C4.5 Decision Tree and Support Vector Machine (SVM) classification algorithms and calculated their prediction accuracy. Feature selection algorithms have been used for each model by reducing some lower ranked attributes, resulting in a much-improved accuracy rate for all three algorithms. The outcome of these SVM, Naïve Bayes and C4.5 has 75.75%, (67.17%) and (73.73%) respectively. Authors in (Alva, 2018) applied the SMOTE technique, which involved synthetically oversampling the minority class no-recurrence. They built a classification model to classify patient data into 2 classes: Recurrence Events, Non-Recurrence Events, using the following classifiers Decision Tree Classifier: 64.27%, Logistic Regression: 73.17%, k Nearest Neighbor: 67.54%%, Support Vector Machine Model (SVC): 70%, Gaussian Naive Bayes Algorithm: 73.17%, and Random Forest Classifier: 70.21%. On testing the model accuracy of each algorithm, they found the Regression Classifier and the Gaussian Naïve Bayes Classifiers had the highest accuracy at 73%.

Most of the studies reported on BCR have been focused on comparing machine learning algorithms for BCR (Ahmad et al., 2013; Toloie Eshlaghy, 2013; Roberto et al., 2020; Yang et al., 2021; Mosayebi et

al., 2020). In (Roberto et al., 2020), authors compared Decision Trees (DT), Naïve Bayes (NB) and Support Vector Machines (SVM), and then integrated the optimum results with Simple K-Means algorithm to generate significant improvements in precision. The study in (Toloie Eshlaghy, 2013) implemented artificial neural networks (ANN), DT, and Support Vector Machine (SVM) for BCR prediction. The dataset consists of 1,189 records with 22 predictor variables and a single outcome variable. The SVM outperforms other techniques and scores the highest accuracy and minimum error rate.

There have been few studies on the prediction of early and/or late breast cancer recurrence.

In the study (Lou et al., 2020), the authors attempted to identify the elements significantly associated with recurrent breast cancer and employ the ANN model to detect the recurrence within ten years after breast cancer surgery. A total of 1,140 patients' data is involved in this study. The model scores an accuracy of 0.988 and a sensitivity of 0.954. In another report (Massafra et al., 2021), authors provided the preliminary results of a prediction model of the Breast Cancer Recurrence (BCR) within five and ten years after diagnosis. Using a small dataset (256 patients), they implemented several feature selection techniques (Neighborhood Component Analysis (NCA), Random Forest (RF) and Support Vector Machine Recursive Feature Elimination (SVM-RFE)) and then evaluated the prediction performances of BCR within 5 and 10 years after the first diagnosis by means of different classifiers. By using a small number of features, the models reached highly performing results both with reference to the BCR within 5 years and within 10 years with an accuracy of 77.50% and 80.39% and a sensitivity of 92.31% and 95.83% respectively, in the hold-out sample test.

DATA PREPARATION AND VISUALIZATION

Data Description

The dataset that has been used for this study was the Breast Cancer METABRIC dataset ("Kaggle", n.d.). This dataset consists of 2510 entries of patient information related to breast cancer. The complete list of dataset features is shown in Table 1. It is important to understand these features in their entirety to determine which attributes are important for our model of predicting breast cancer recurrence.

Data Visualization

Data visualizations are particularly important in understanding how the features are related to one another and how the data can be represented. By visualizing the data, we can also better understand the distribution of the patient's data as well as any anomalies and outliers that could cause problems down the road. As seen through these visualizations, there are slightly more cases of patients having no recurrence than those who had breast cancer return (Figure 1).

In Figure 2, it can also be seen that when it comes to recurrence and chemotherapy, there are more patients that did not have recurring breast cancer, specifically those who did not do chemotherapy. This might be alarming at first, because one would hope that chemotherapy keeps cancer from returning, however there is an issue of NA or empty values in the dataset. There are currently 529 missing entries for chemotherapy which could have a greater impact on the visualization of that data. This will be fixed with some proper data preprocessing before there is any important modeling performed.

Table 1. List of the attributes/features in METABRIC dataset ("Kaggle", n.d.)

Age at Diagnosis
Type of Breast Surgery
Cancer Type
Cancer Type Detailed
Cellularity
Chemotherapy
Pam50 + Claudin-low subtype
Cohort
ER status measured by IHC
ER Status
Neoplasm Histologic Grade
HER2 status measured by SNP6
HER2 Status
Tumor Other Histologic Subtype
Hormone Therapy
Inferred Menopausal State
Integrative Cluster
Primary Tumor Laterality
Lymph nodes examined positive
Mutation Count
Nottingham prognostic index
Oncotree Code
Overall Survival (Months)
Overall Survival Status
PR Status
Radio Therapy
Relapse Free Status (Months)
Relapse Free Status
Sex
3-Gene classifier subtype
Tumor Size
Tumor Stage
Patient's Vital Status

Because this study is focusing specifically on breast cancer recurrence, there are several visualizations that highlight the data of relapse. These visualizations will look at the distribution of recurrence and non-recurrence based on age, NPI (Nottingham Prognostic Index) score (NPI, n.d.), Hormone Therapy, and Radio Therapy as shown in Figure 3-6, respectively. Similar histograms can be created for positive lymph nodes, PR status, ER status, and Her2 status.

Figure 1. A histogram depicting the percentage of the patients who had breast cancer recurrence

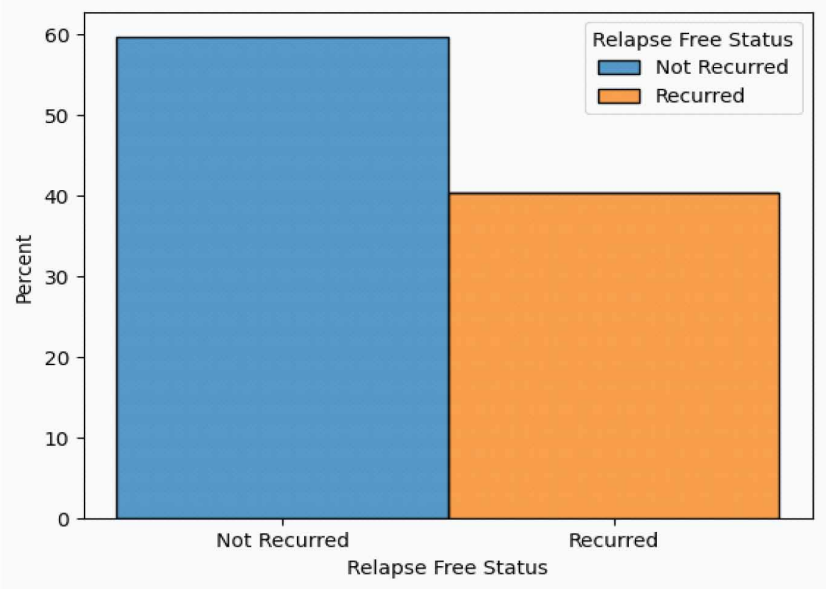


Figure 2. A histogram depicting the Chemotherapy distribution of those patients who had chemotherapy treatment and breast cancer recurrence or non-recurrence

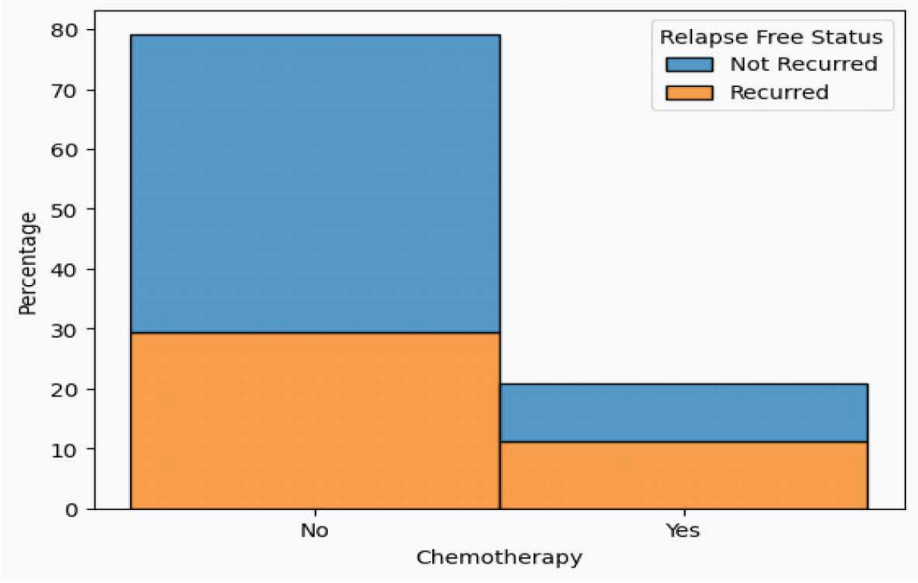


Table 2 shows the list of the features and the number of missing entries contained in the dataset.

Some of these numbers are concerning because there are only 2510 patient entries, meaning some of these features have a good portion of their data missing. To perform proper modeling and feature selection, the number of missing values needed to be reduced (Makaba & Dogo, 2019). To start, the

Figure 3. A histogram depicting the age distribution of those who had breast cancer recurrence and non-recurrence

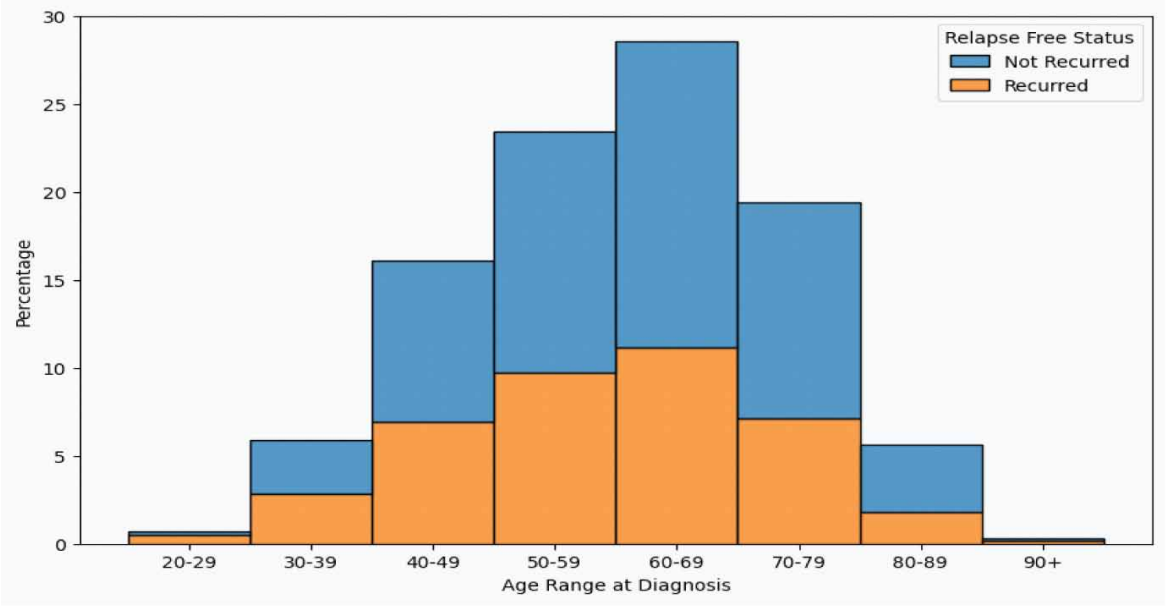


Figure 4. A histogram depicting the NPI score distribution of those patients who had breast cancer recurrence and non-recurrence

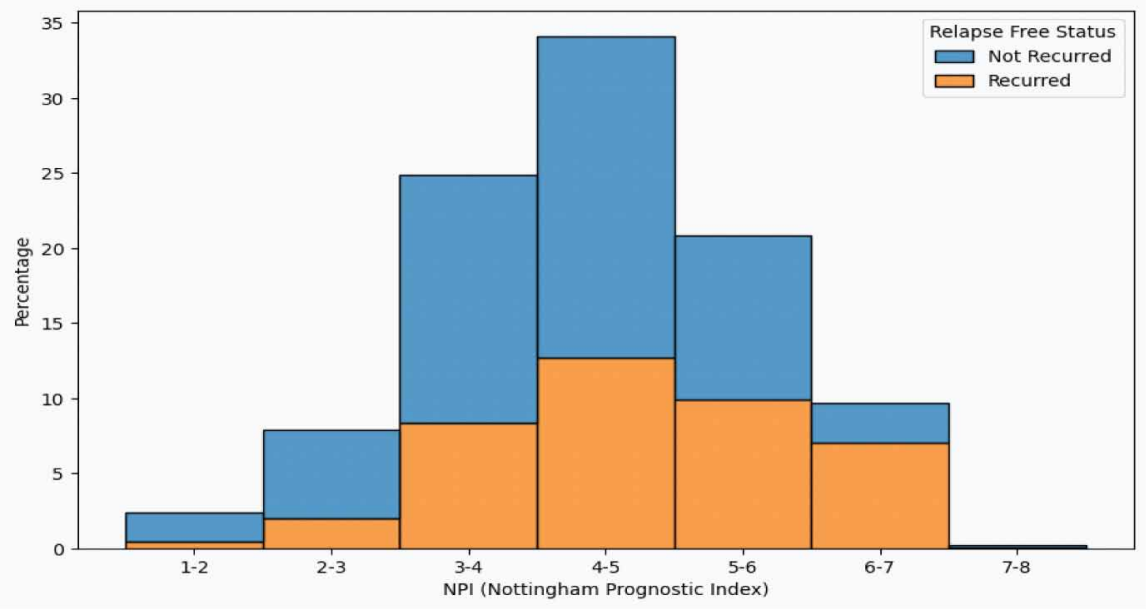


Figure 5. A histogram depicting the Hormone therapy distribution of those patients who had hormone therapy treatment and breast cancer recurrence or non-recurrence

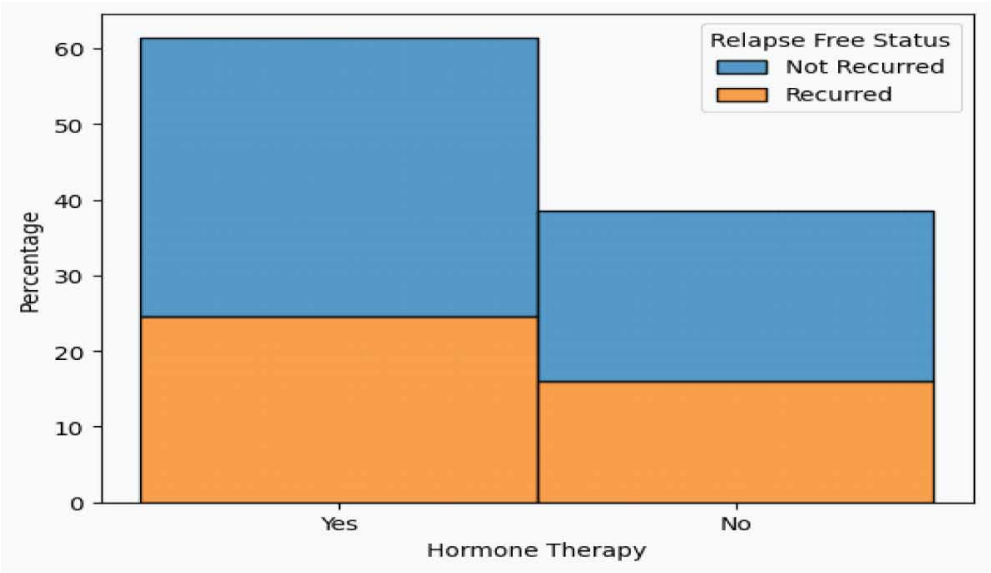


Figure 6. A histogram depicting the Radio therapy distribution of those patients who had radio therapy treatment and breast cancer recurrence or non-recurrence

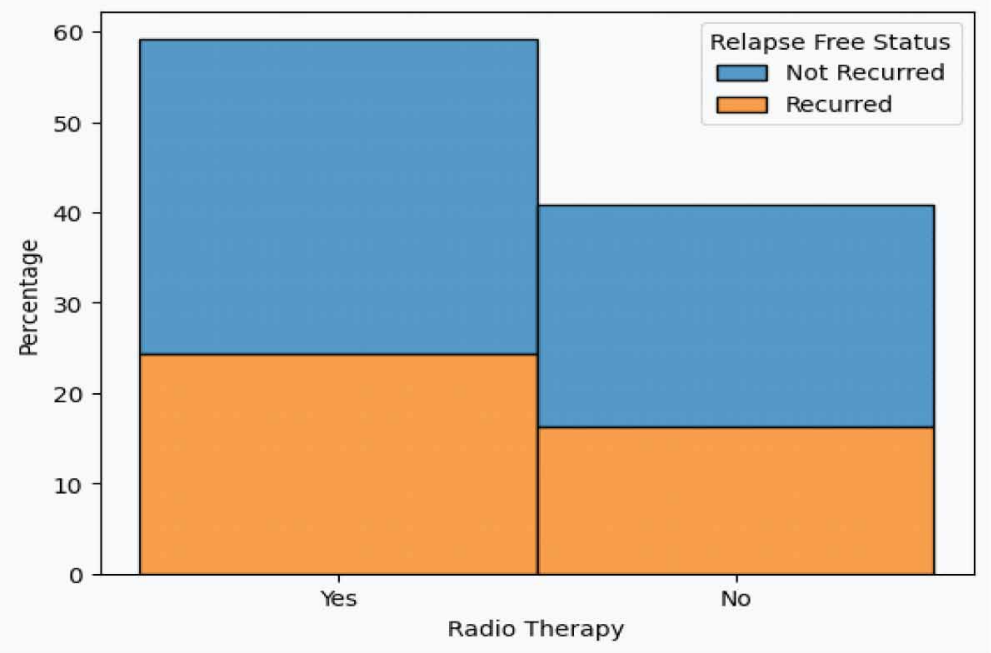


Table 2. List of features with missing entries

Feature	Missing entries
Age at Diagnosis	11
Type of Breast Surgery	554
Cellularity	592
Chemotherapy	529
Pam50 + Claudin-low subtype	529
ER Status	40
Neoplasm Histologic Grade	121
HER2 Status	529
Tumor Other Histologic Subtype	135
Hormone Therapy	529
Inferred Menopausal State	529
Primary Tumor Laterality	639
Lymph nodes examined positive	266
Mutation Count	152
Overall Survival (Months)	528
Overall Survival Status	528
PR Status	529
Radio Therapy	529
Relapse Free Status (Months)	121
Relapse Free Status	21
Tumor Size	149
Tumor Stage	721
Patient's Vital Status	529

features' datatypes have been examined to see if some of the missing values can be filled with a mean, minimum, or maximum value. For the features that are categorical, the empty entries could be filled with the most common value or drop the row all together. It seemed dropping the rows were more important to maintaining the integrity of the dataset and not skewing it in any way that could affect results later.

Before beginning feature selection, looking at the dataset, there were several attributes that were unnecessary for the study of breast cancer recurrence. These features included: Cancer type, Sex, NPI, Patient ID, Cancer Type Detailed, Cohort, ER status measured by IHC, HER2 status measured by SNP6, Integrative Cluster Theory and Oncotree Code, and 3-Gene classifier subtype. Cancer type and sex have been removed because they were the same for every patient in the dataset, and thus had no importance to the research. Patient ID and Cohort were unnecessary to this study as they were methods of classifying and tracking patients and were unimportant for the methods of prediction. It seemed that the features of Cancer type detailed, ER status measured by IHC, HER2 status measured by SNP6, and 3-Gene classifier subtype were repetitive in nature since it was known that the cancer type is Breast Cancer, and there are other features — ER status and HER2 status — that already provided the necessary information. There was no need to know how the status was measured, rather only needed the overall status. Integrative

cluster Theory and Oncotree Code were removed since there was little previous studies and research that proved that these features were important or necessary to predicting breast cancer recurrence.

After filling and dropping the necessary entries, the features to convert the categorical features into numerical values were encoded. Numerical data were needed to perform certain machine learning methods for feature selection and recurrence predictions (i.e., regression algorithms). Before using machine learning prediction models to predict if a patient will have breast cancer recur or not, it was necessary to figure out which features are the most important to recurrence itself. /The next section will discuss the methods for choosing these features, their importance, and their results. These features will then be used in the overall prediction of breast cancer recurrence.

ATTRIBUTE AND FEATURE SELECTION

Techniques and Methods

This section will describe the techniques and methods used to perform feature selection on the Breast Cancer METABRIC dataset (Kaggle, n.d.). A total of 6 methods have been used to retrieve rankings and correlation scores of the attributes as compared to the target feature of "Relapse Free Status". Most of these techniques were performed using the programming language of Python and its related libraries (Raschka & Mirajalili, 2017). Feature selection is the technique where one chooses the features that contribute the most to the target variable (Chang & Lin, 2008). The advantages of feature selection are a reduction in overfitting, a possible improvement in accuracy, and faster training time.

SelectKBest

The first method used was SelectKBest. SelectKBest uses the Scikit-learn API (Bisong, 2019) and provides the class for extracting best features of a given dataset. The SelectKBest method selects the features according to the K highest score. Sklearn provides a universal function SelectKBest which can select K best features based on some metric. The main idea of SelectKBest is to calculate some metrics between the target and each feature, sort them, and then select the K best features (Bisong, 2019). Selecting the best features is an important process when we prepare a large dataset for training.

Results: A ranking of attributes from highest to lowest: Chemotherapy, ER status, Neoplasm Grade, Overall Survival Status (Months), PR Status, Relapse Free Status (Months), and Tumor Size.

LinearSVC

The second method performed was LinearSVC. SVM/SVC (Chang & Lin, 2008) are machine learning models used for classification. The main objective in SVM is to find the optimal hyperplane to correctly classify data points of different classes. Once having fitted the linear SVM to the data, it is possible to access the classifier coefficients on the trained model. These weights figure the orthogonal vector coordinates orthogonal to the hyperplane. Their direction represents the predicted class. Feature importance can, therefore, be determined by comparing the size of these coefficients to each other. By

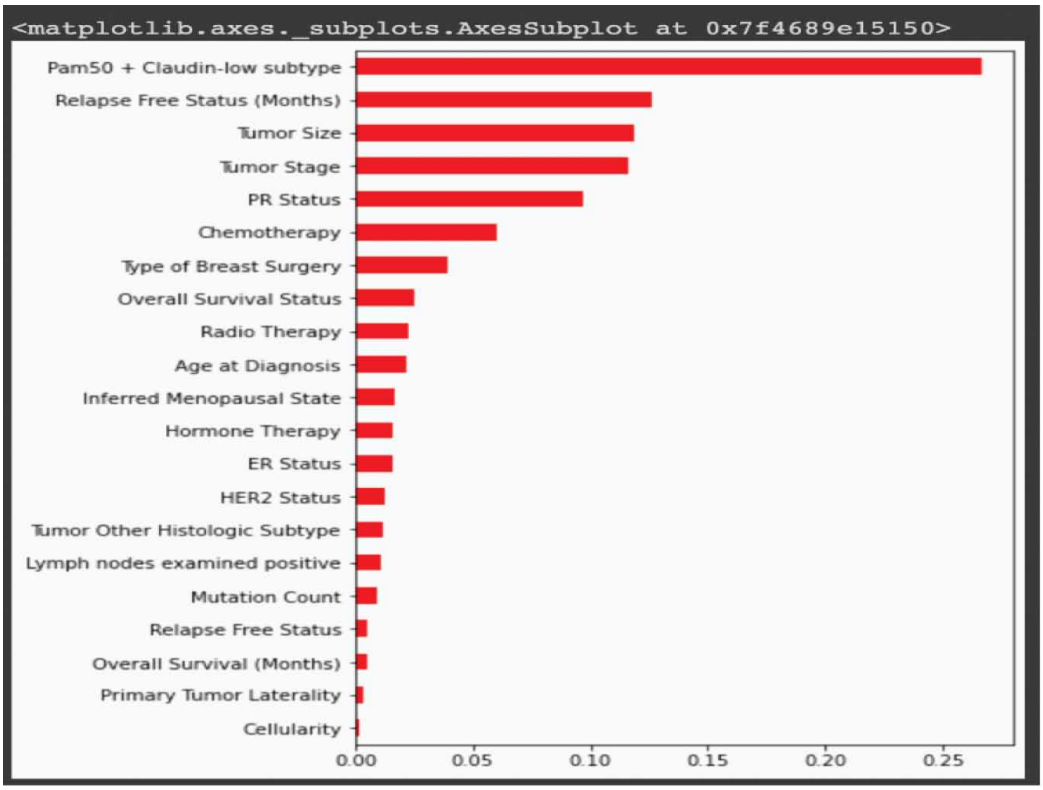
looking at the SVM coefficients, one can identify the main features used in classification, and get rid of the not important ones.

Results: LinearSVC can produce a feature ranking of: Pam50 + Claudin-low subtype, ER Status, Overall Survival Status (Months), Overall Survival Status, PR Status, and Radiotherapy.

Decision Tree- Feature Importance

Feature Importance technique within the Decision Tree model (Myles et al., 2004) has been used as the third method of feature selection. Decision trees learn how to best split the data into smaller and smaller subsets to predict the target value. The condition, or test, is represented as the "leaf" and the possible outcomes as "branches". This splitting process continues until no further gain can be made or a preset rule is met, meaning the maximum depth of the tree is reached. Feature importance is calculated as the decrease in node impurity weighed by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature, as shown in Figure 7. Given this graph, one can see that the features with the higher values are more important.

Figure 7. A bar graph illustrating the features and their score rankings based on how well they predict a target variable



Results: A ranking of attributes from highest to lowest: Pam50 + Claudin-low subtype, Relapse Free Status (Months), Tumor Size, Tumor Stage, PR status, Chemotherapy, Type of Breast Surgery

Recursive Feature Elimination

The fourth feature selection method was Recursive Feature Elimination using Linear Regression (Darst et al., 2018). Recursive Feature Elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model's coefficient or feature importance's attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model. RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features.

Results: The most important features selected because of RFE are: Type of Breast Surgery, Cellularity, Pam50 + Claudin-low subtype, ER Status, Tumor Other Histologic Subtype, Hormone Therapy, Inferred Menopausal State, Overall Survival Status, Radio Therapy, Patients Vital Status.

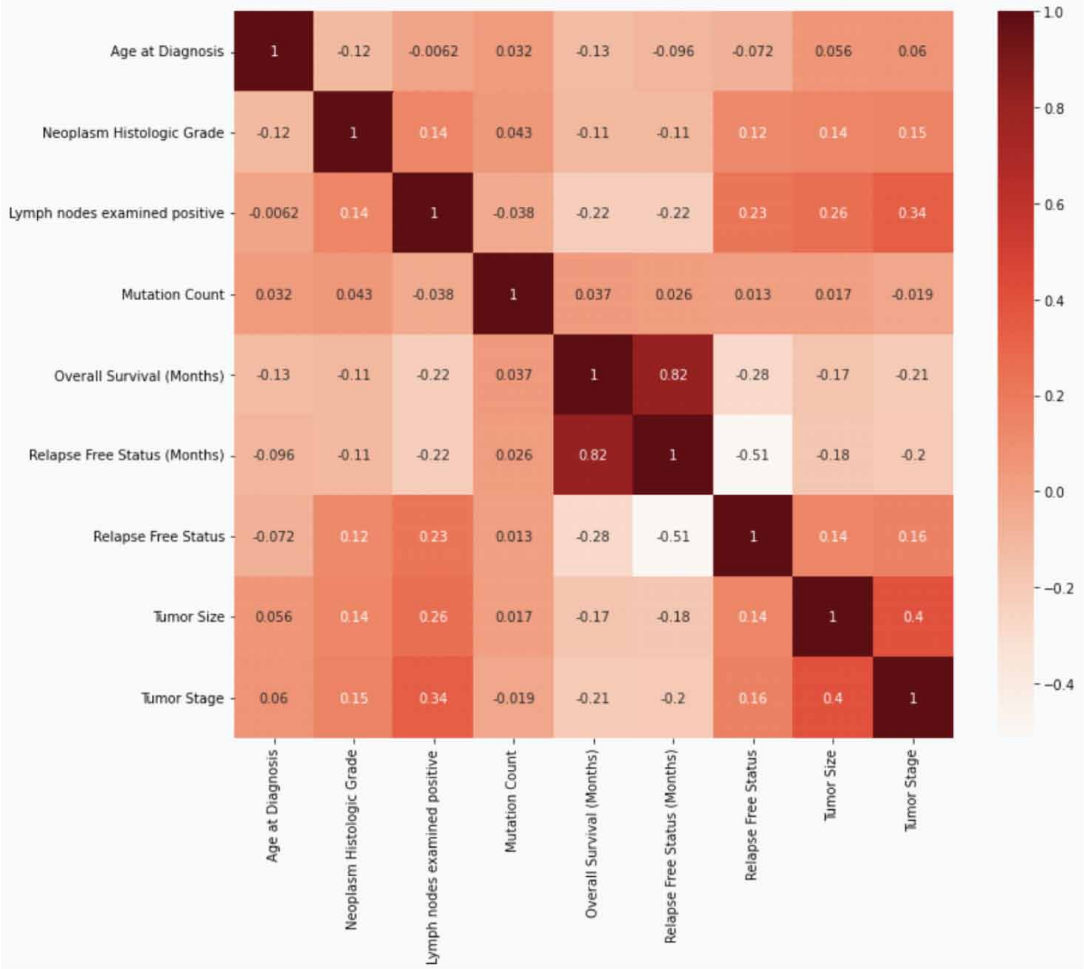
Pearson Correlation

The fifth method of feature selection was Pearson Correlation (Benesty et al., 2009). The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation. If two variables/features are correlated, we can predict one from the other. Therefore, the model only needs one, as the second does not add additional information. We need to set an absolute value, say 0.5, as the threshold for selecting the variables.

Correlation heatmaps (Heatmap, n.d.) are a type of plot that visualizes the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. A correlation plot typically contains several numerical variables, with each variable represented by a column. The rows represent the relationship between each pair of variables. The values in the cells indicate the strength of the relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship. Correlation heatmaps can be used to find potential relationships between variables and to understand the strength of these relationships. In addition, correlation plots can be used to identify outliers and to detect linear and nonlinear relationships. The color-coding of the cells makes it easy to identify relationships between variables briefly. Correlation heatmaps can be used to find both linear and nonlinear relationships between variables. A heat map of the correlations is shown in Figure 8.

Results: By looking at this heat map, it is clear to see which features have the highest positive correlation to the target variable "Relapse Free Status". These features are ranked as being: Lymph nodes examined positive, Tumor Stage, Tumor Size, Neoplasm Histologic Grade, Mutation Count

Figure 8. A heat map showing the linear correlation between attributes



Chi-Square Tests

The last and sixth method of feature selection for the Breast Cancer METABRIC dataset was Chi-Square Tests (Tallarida & Murray, 1987). Equation (1) shows the Formula for Chi Square test. Chi-square test in hypothesis testing is used to test the hypothesis about the distribution of observations/frequencies in different categories. A chi-square test is used in statistics to test the independence of two events. As seen in (1), given the data of two variables, the observed count O and expected count E can be obtained. Chi-Square measures how expected count E and observed count O deviate from each other.

Formula for Chi-Square:

$$X_c^2 = \sum \frac{\left(O_i - E_i\right)^2}{E_i} \tag{1}$$

where:

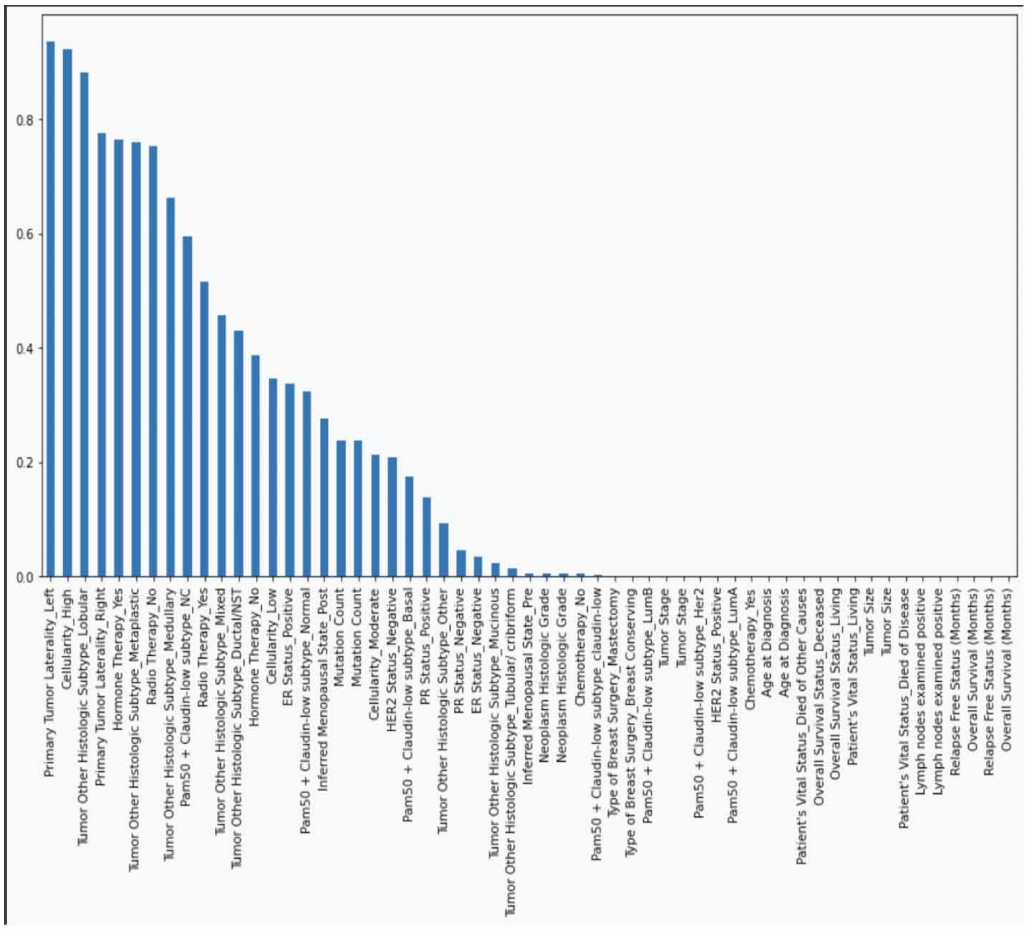
C= degrees of freedom

O=observed value (S)

E= expected value(S)

The primary use of the chi-square test is to examine whether two variables are independent or not. In feature selection, the aim is to select the features which are highly dependent on the response. When two features are independent, the observed count is close to the expected count, thus a smaller Chi-Square value is produced. So high Chi-Square value indicates that the hypothesis of independence is incorrect. The higher the Chi-Square value the feature is more dependent on the response, and it can be selected for model training (Figure 9). In our use of chi-square tests, we are looking to see if there is a relationship between breast cancer recurrence and the other features individually.

Figure 9. A bar graph with a ranking of Chi-Square test scores



Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Counts of experimental data are called observed frequencies. Expected values are a type of frequency that may be calculated using probability theory.

Results: Below is a ranking of the features based on their test scores. These scores include encoded values, so we must look at the overall feature. The selected features are: Primary Tumor Laterality, Cellularity, Tumor Other Histologic Subtype, Hormone Therapy, Pam50 + Claudin-low subtype, Radio Therapy, ER Status, Inferred Menopausal State, Mutation Count, HER2 Status, PR Status.

Final Features Selected From Attribute/Feature Selection Schemes

Given all these methods of feature selection and extraction, one must now compare all the results and pick the top recurring attributes. By getting the union of all these features, it is possible to determine the attributes that contribute the most to BCR. The following six attributes have been picked which were most common among all these methods:

- ER Status
- Pam50 + Claudin-low subtype
- Relapse Free Status (Months)
- Tumor Size
- PR Status
- Radio Therapy
- These will be the features used in the prediction models

Data Visualization for Selected Features

Based on the following plots (Figure 10) that show the percentage value of each feature, ER and PR status is positive for most of the patients. In addition, the number of patients that did Radio Therapy is more than those who did not. For Pam50 + Claudin-low subtype, most of the patients are in LUM A group and the LUM B is second. For Relapse Free Status (Months), cancer returns after around 27 months for most the patients (around 150 patients). Regarding the Tumor size, around half of the patients had tumors with a 25 mm size.

Considering the plots in Figure 11 that are drawn based on the percentage value of each feature and the target number (Relapse-Free Status), the cancer did not return for most of the positive ER patients. For positive and negative PR status, the number of patients with recurrence and non-recurrence are equal approximately. For the patients that did Radio therapy, the number of recurrences is less than the number of non-recurrences. For patients that had 27 months of Relapse Free Status, cancer returned for around 70% of the patients. For Tumor size, the cancer did not return for most patients who had tumor size of 25 mm. Finally, for the LUM A Pam50 + Claudin-low subtype which has the greatest number of patients, the cancer recurrence is less than the non-recurrence.

In the plots of Figure 12, drawn with Tableau Data Visualization tool (Tableau, n.d.), shows and compares all features to visualize which patient's cancer recurred or not. It should be noted that the Tumor Size (mm) and Relapse Free Status (months) are removed because there are more than 1000 different tumor sizes and different number of months for Relapse Free Status. In this kind of plot, having

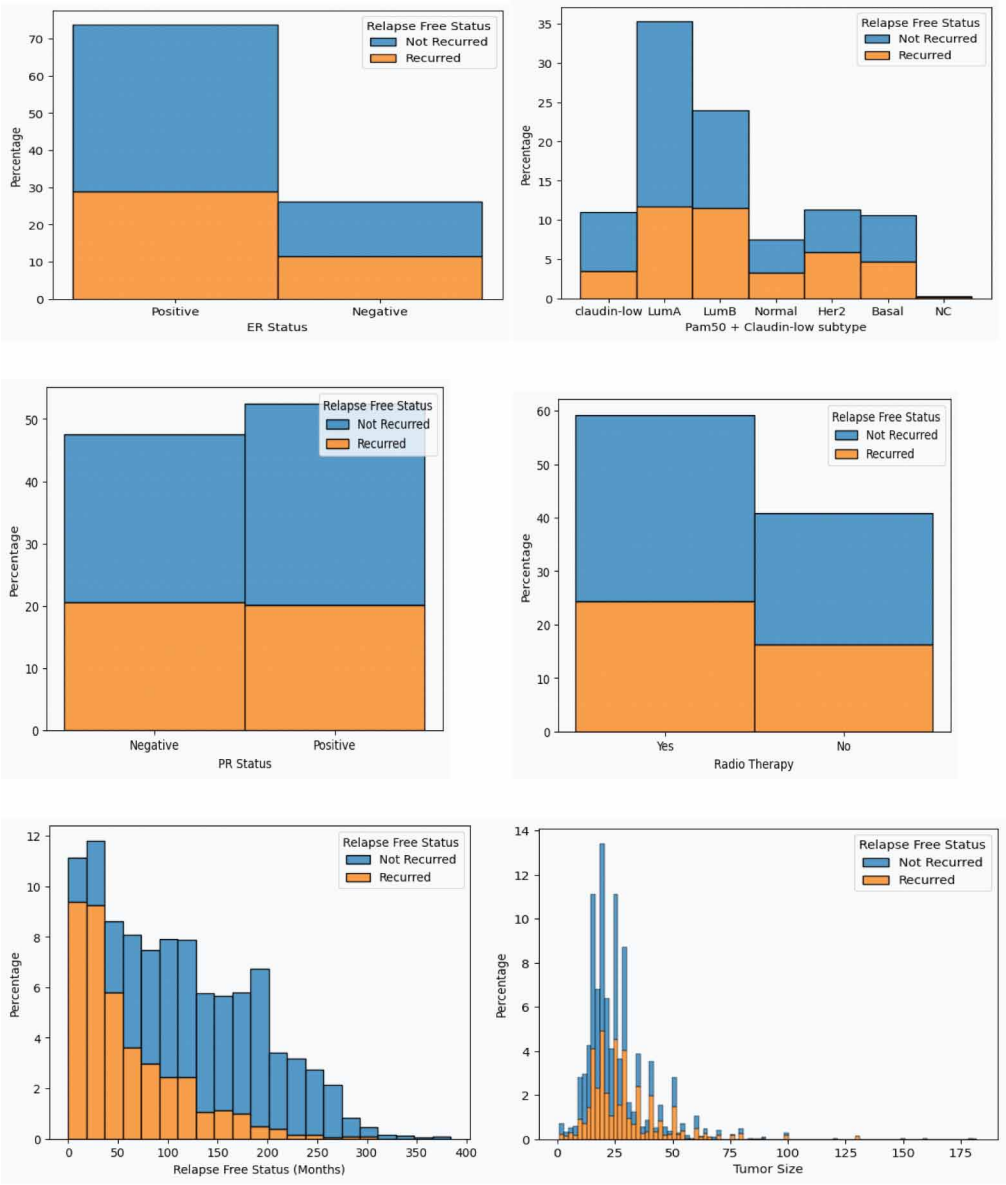
Figure 10. Percentage of the selected features

each feature plot based on other features could lead to a long plot. Therefore, two features are removed to make the plot shorter.

Considering the plot in Figure 12, the following observations can be made:

1. When ER is negative, the number of cancer recurrences is less.
2. When the ER is negative, and the PR is negative, and the patient has not had Radio Therapy, the number of patients in terms of Pam characteristics is more in the Claudin-low subgroup, and the number of patients whose cancer has not returned is more.

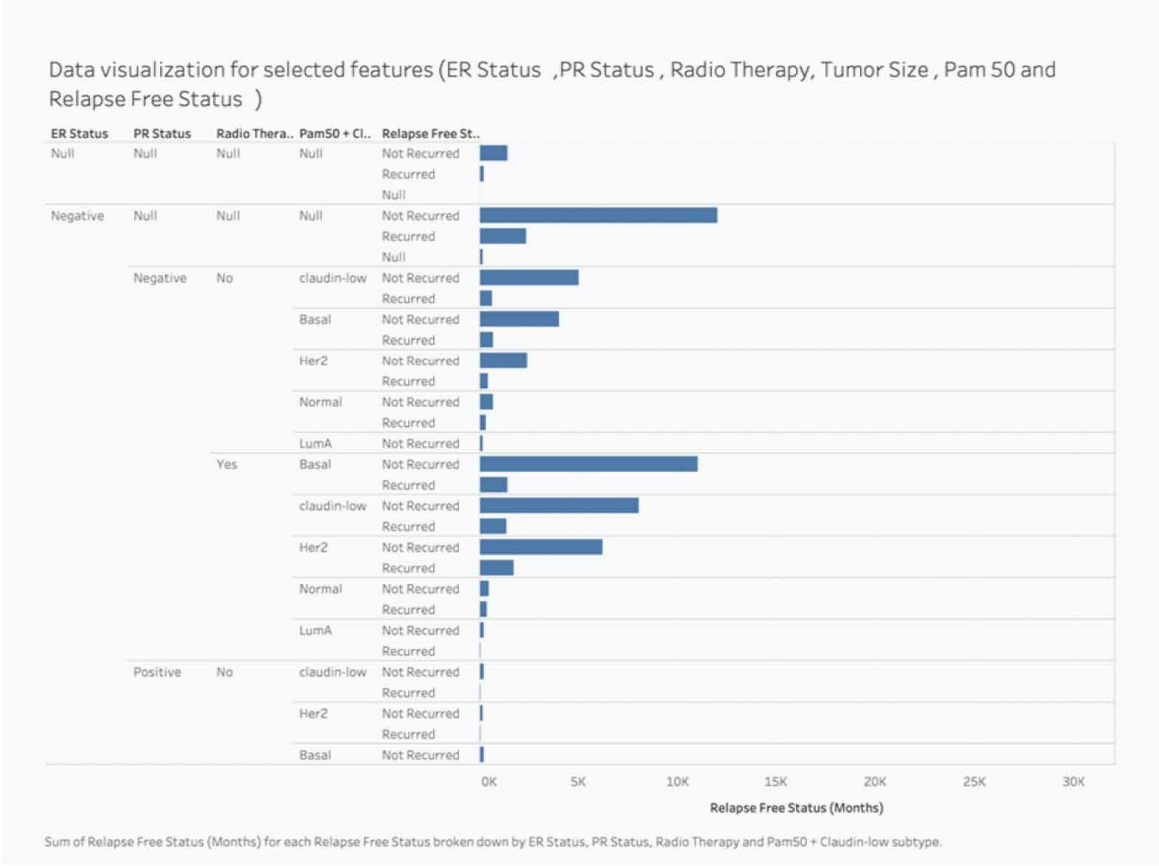
Figure 11. Percentage value of each selected feature and the target number (relapse free status)



3. If the Radio Therapy is the only one changed, i.e., the patients who have done radiotherapy and their PR and ER status is negative, then most of the non-recurrence patients are in the Basal category.
4. When ER is negative, but the PR becomes positive and Radio Therapy has not been performed, there is no significant change in the cancer recurrence or non-recurrence patients. However, the number of recurrences is still lower than non-recurrence.

In summary, from the plots in Figure 12 with some other plots for other attributes interactions, it can be understood that the patient whom cancer returns are the patients who are ER positive and have a

Figure 12. Tableau data visualization sample plot demonstrating associations among features



tumor size of around 25 mm. They did not have Radio Therapy, and the Pam50 + Claudin-low subtype is in Lum A or Lum B and +/- PR status.

BREAST CANCER PREDICTION MODELS

The purpose of this research is to be able to predict if a patient has a high chance of breast cancer recurrence, given certain attributes. As established in the feature selection procedure, there were six features determined to be contributing factors of recurrence. These factors are ER Status, Pam50 + Claudin-low subtype, Relapse Free Status (Months), Tumor Size, PR Status, and Radio Therapy. These factors will be used to predict if a patient's breast cancer would relapse. To do this, various machine learning models including Linear Regression, Decision Tree (DC), and Random Forest, Support Vector Machine (SVM), and Neural Network (Ahmad et al., 2013; Abreu et al., 2016; Mohammed et al., 2016; Alva, 2018; Aishwarja et al., 2021) have been used. After explaining these models and what they do, the results of each model will be established and determined which is the best at prediction for the data and the selected features. To better show how feature selection approaches affect the

outcome and prediction, the data set after feature selection and the original dataset with 33 features have been used to train DC algorithm.

The Linear Regression and Random Forest models are evaluated using different statistical measures of Mean Squared Error, R-squared error, and Mean Absolute Error (Hodson, 2022). Following the modeling, the results were calculated using error and loss functions (Hodson, 2022). Those calculations are for mean squared error, mean average error, and r-squared error. They are described as the following. The Mean absolute error (MAE) represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset. Unlike the mean squared error (MSE), the MAE calculates the error on the same scale as the data. The MAE doesn't square the differences and is less susceptible to outliers. Both values are negatively oriented which means that while both range from 0 to infinity, lower values are better. The Mean squared error (MSE) represents the error of the estimator or predictive model created based on the given set of observations in the sample. Intuitively, the MSE is used to represent the penalty of the model for each of the predictions. In other words, it can be used to represent the cost associated with the prediction. There is no correct value for MSE. However, the lower the value the better and 0 means the model is perfect. R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score, the value of R-square will be less than one. If the R-square value is 100%, the two variables are perfectly correlated, i.e., with no variance at all. A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases. These measures are useful for this research as it is looking to predict if patients will have breast cancer recurrence or not. The dataset contains the actual value of 1 if the patient's breast cancer recurred.

Linear Regression

Regression is a method of modeling a target value based on independent predictors. This method is mostly used for predicting the cause-and-effect relationship between variables. Simple linear regression (Maulud & Abdulazeez, 2020) is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent and dependent variables. After performing linear regression, the mean average error, mean squared error, and R-squared error were calculated.

Results:
MAE: 0.375
MSE: 0.181
RSE: 0.425

These numbers show the errors of our predicted values compared to the actual values. The mean absolute error, mean squared error, and r-squared error are different error metrics to see how well our model performed. While the model is not perfect, i.e., the mean absolute error and mean square error are not zero, the values are low enough to determine that the predicted values did not vary that far from the actual values. This means that the model was close to predicting if a patient would have breast cancer recurrence or not. The R-squared value is higher because if there is perfect correlation between the features and breast cancer recurrence it would be 100%. Because there isn't a perfect correlation,

the R-Squared value is less than 100%. The reason there could not be perfect correlation could be that the dataset doesn't provide other important factors like race, ethnicity, or other preexisting conditions.

Random Forest

Random forest (Liu et al., 2012) is a type of supervised learning algorithm that uses ensemble methods (bagging) to solve both regression and classification problems. The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees. In the case of a regression problem, the final output is the mean of all the outputs. After performing random forest regression, the mean average error, mean squared error, and R-squared error were calculated.

Results:
MAE: 0.378
MSE: 0.378
RSE: 0.615

This Decision Tree and Random Forest model, in comparison with the Linear Regression model, performed less accurately. The mean absolute and mean squared error are higher than the Linear Regression model. This means that the predicted values didn't match the actual values quite as well. The ability for these models to predict breast cancer recurrence based on the selected features was not as accurate. The R-squared error is higher than the Linear Regression model which means that there is a higher correlation between these features and breast cancer recurrence according to these models.

It should be noted that a confusion matrix is designed for classification tasks (Decision Tree, Support Vector Machine, Neural Network), assessing how well models categorize data into classes. It doesn't apply to linear regression, which predicts continuous numerical values, not classes. As explained earlier, in regression, metrics like MAE, MSE, RMSE, and R-squared quantify differences between predicted and actual values, differing from classification evaluations.

Decision Tree

Decision Tree (DT) (Charbuty & Abduazeez, 2021) is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set, and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. A regression tree is basically a decision tree that is used for the task of regression which can be used to predict continuous valued outputs instead of discrete outputs (Charbuty & Abdulazees, 2021). GradinetBooster is one of the DT models that was chosen because it is more accurate than regular DT. Gradient boosting works by using simpler prediction models sequentially where each model tries to predict the error left over by the previous model. In gradient boosting decision trees, many weak learners are combined to come up with one strong learner. All the trees are connected in series and each tree tries to minimize the error of the previous tree. Due to this sequential connection, the gradient boosting algorithm is usually slow to learn, but also highly accurate.

After this model was trained using the entire data set, the Recall, F1 Score, Precision, and Accuracy were calculated.

Results:
Recall: 0.378
F1 Score: 0.352
Precision: 0.371
Accuracy: 0.412

After running the model with the target variable of recurrence, it gave the following.

Results:
Recall: 0.782
F1 Score: 0.720
Precision: 0.745
Accuracy: 0.831

Given the results, there is an improvement from the first model. In fact, it shows how feature selection techniques can help to have a better prediction.

- **Optimizing Decision Tree:** For optimization, this model was tested with different depths of (2, 3, 5, 7), different learning rates (0.001, 0.01, 0.1), and different nodes (200, 250, 300).
- **Depth (2, 3, 5, 7):** The maximum depth of the individual regression estimators

The maximum depth limits the number of nodes in the tree. You must tune this parameter for best performance; the best value depends on the interaction of the input variables. The values must be in the range (200, 250, 300). This is the number of trees you want to build before taking the maximum voting or averages of predictions. The higher number of trees gives you better performance. In other words, it is the number of boosting stages to perform.

- **Learning rate (0.001, 0.01, 0.1):** Learning rate, denoted as α, simply means how fast the model learns. Each tree added modifies the overall model. The magnitude of the modification is controlled by the learning rate. The lower the learning rate, the slower the model learns.

The best situation for this model was {'learning rate': 0.01, 'max depth': 3, 'n estimators': 200} that means if our model has 200 nodes and depth of 3 and the learning rate be 0.01.

After optimization the results were:

Recall: 0.790
F1 Score: 0.729
Precision: 0.752
Accuracy: 0.859

As a result, Accuracy increased to 0.859.

Summary of the model:

GridSearchCV(estimator=GradientBoostingClassifier(),

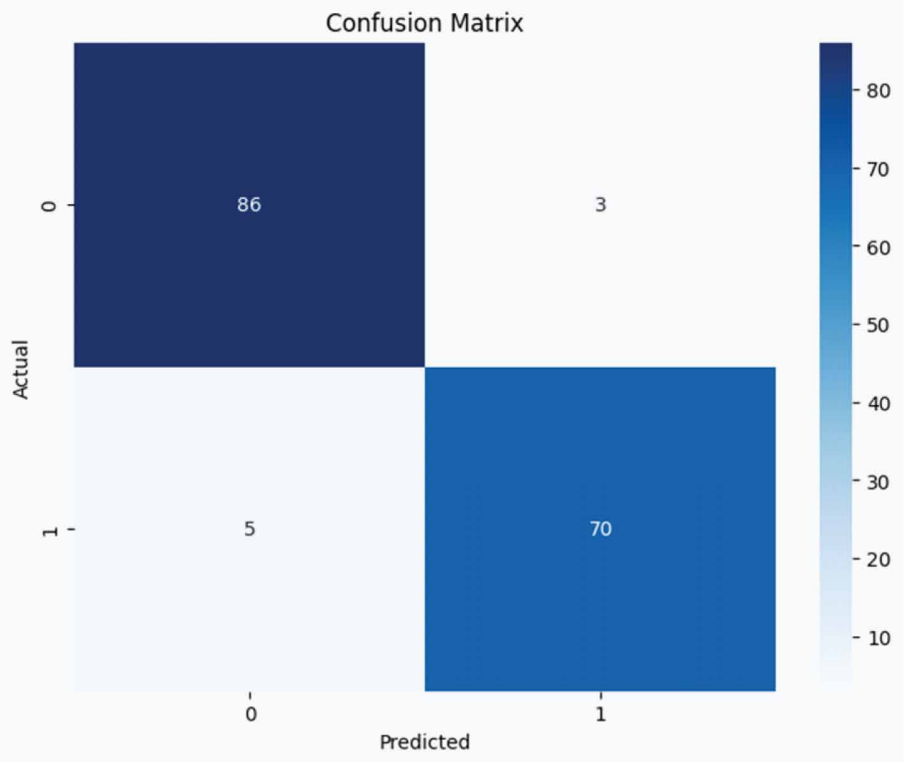
param_grid={'learning_rate': (0.001, 0.01, 0.1),

'max_depth': (2, 3, 5, 7),

'n_estimators': (200, 250, 300)})

Figure 13 represents the confusion matrix for Decision Tree model.

Figure 13. Confusion matrix for Decision Tree model



The confusion matrix illustrates the model's performance in classifying instances into two classes. The matrix is divided into four categories:

- **True Positives (TP):** 70 instances were correctly classified as positive.
- **True Negatives (TN):** 86 instances were correctly classified as negative.
- **False Positives (FP):** 3 instances were falsely classified as positive.
- **False Negatives (FN):** 5 instances were falsely classified as negative.

Metrics Analysis

The classification report provides several metrics for assessing the model's performance:

- **Precision:** The precision for class 0.0 is 0.752, indicating that among the instances predicted as class 0.0, 75% were actually class 0.0. For class 1.0, the precision is 0.76.
- **Recall:** The recall for class 0.0 is 0.79, which means the model correctly identified 79% of all actual class 0.0 instances. For class 1.0, the recall is 0.80.
- **F1-Score:** The F1-score balances precision and recall, providing an overall measure of the model's accuracy. The F1-score for class 0.0 is 0.790, and for class 1.0, it's 0.80.
- **Accuracy:** The overall accuracy of the model is 0.859, indicating that it correctly classified 86% of instances.

Conclusion

The high precision, recall, and F1-score values suggest that the model exhibits strong performance in both classes. With an accuracy of 0.86, it demonstrates consistent and reliable results. These metrics collectively highlight the model's effectiveness in correctly classifying instances, underscoring its potential for practical applications.

Support Vector Machine

A support vector machine (SVM) (SVM, n.d.; Chang & Lin, 2008) is a supervised Machine Learning model (Mohammed et al., 2016) that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. After performing SVM with selected feature, the Precision, Recall, and Accuracy were calculated.

Results:
Recall: 0.734
F1 Score: 0.732
Precision: 0.765
Accuracy: 0.801
Optimizing SVM:

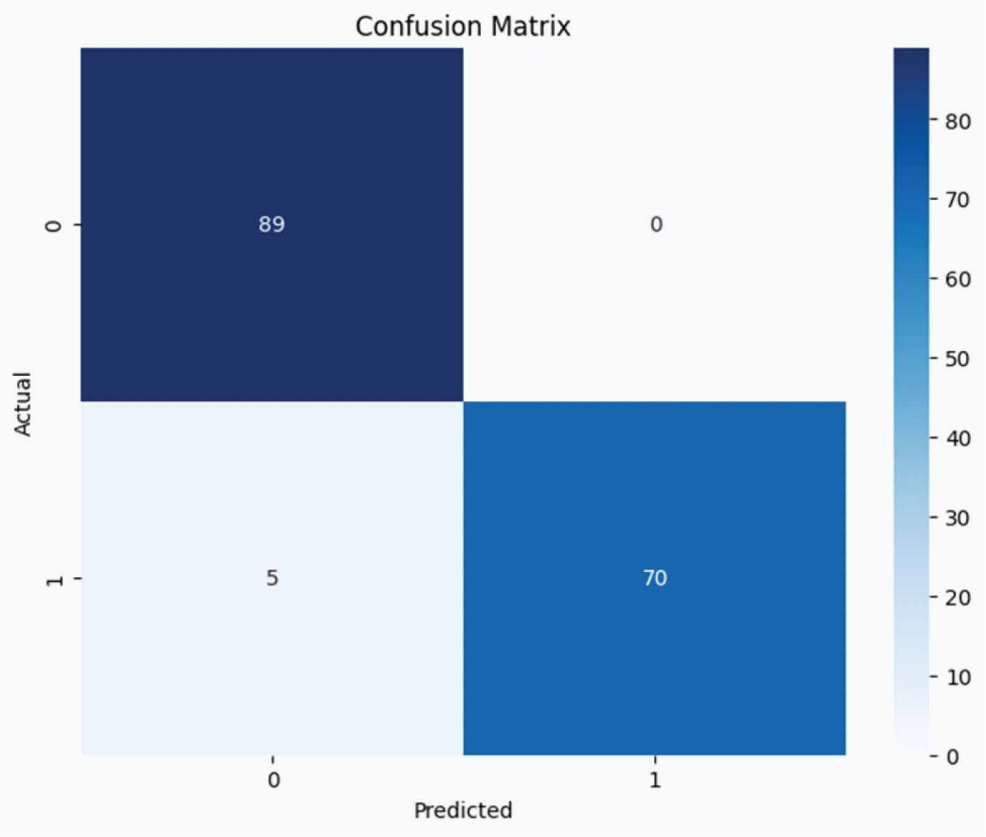
For optimizing this model, C: (0.1, 1, 10, 20) is considered. The C parameter in SVM is the Penalty parameter of the error term. It can be considered as the degree of correct classification that the algorithm must meet or the degree of optimization the SVM must meet. For greater values of C, there is no way that an SVM optimizer can misclassify any single point. For the kernel all possible kernels were considered: kernel: ('rbf', 'poly', 'sigmoid'). Kernel parameters are used as a tuning parameter to improve the classification accuracy. There are mainly four different types of kernels (Linear, Polynomial, RBF, and Sigmoid) that are popular in SVM classifiers. The best parameter for this model was {'C': 1, 'kernel': 'rbf'}.

After optimization the results were:

Results:
Recall: 0.785
F1 Score: 0.741
Precision: 0.762
Accuracy: 0.812

As a result, the Accuracy increased to 0.812.
Figure 14 represents the confusion matrix for SVM model.

Figure 14. Confusion matrix for SVM model



Neural Network (Multilayer Perceptron)

A network model (Zhou, 2019) is a simplified model of the way the human brain processes information. It works by simulating many interconnected processing units that resemble abstract versions of neurons. The processing units are arranged in layers. After performing this model with selected features, the Recall, F1 Score, Precision, and Accuracy were calculated.

Results:
Recall: 0.728
F1 Score: 0.720
Precision: 0.732
Accuracy: 0.758

This means that if a cancer patient with selected features uses our model, 76 percent of patients get the correct prediction of breast cancer recurrence.

Optimizing

For optimizing this model, several kind of layers were used in a 2D faze:

(20,)
(50,)
(100,)
(50, 50)
(50, 100)
(100, 50)
(100, 100))

A layer in a deep learning model is a structure or network topology in the architecture of the model, which takes information from the previous layers and then passes information to the next layer. And the best accuracy achieved in {'hidden_layer_sizes': (100,)}.

After optimization the results were:

Recall: 0.770
F1 Score: 0.741
Precision: 0.756
Accuracy: 0.801

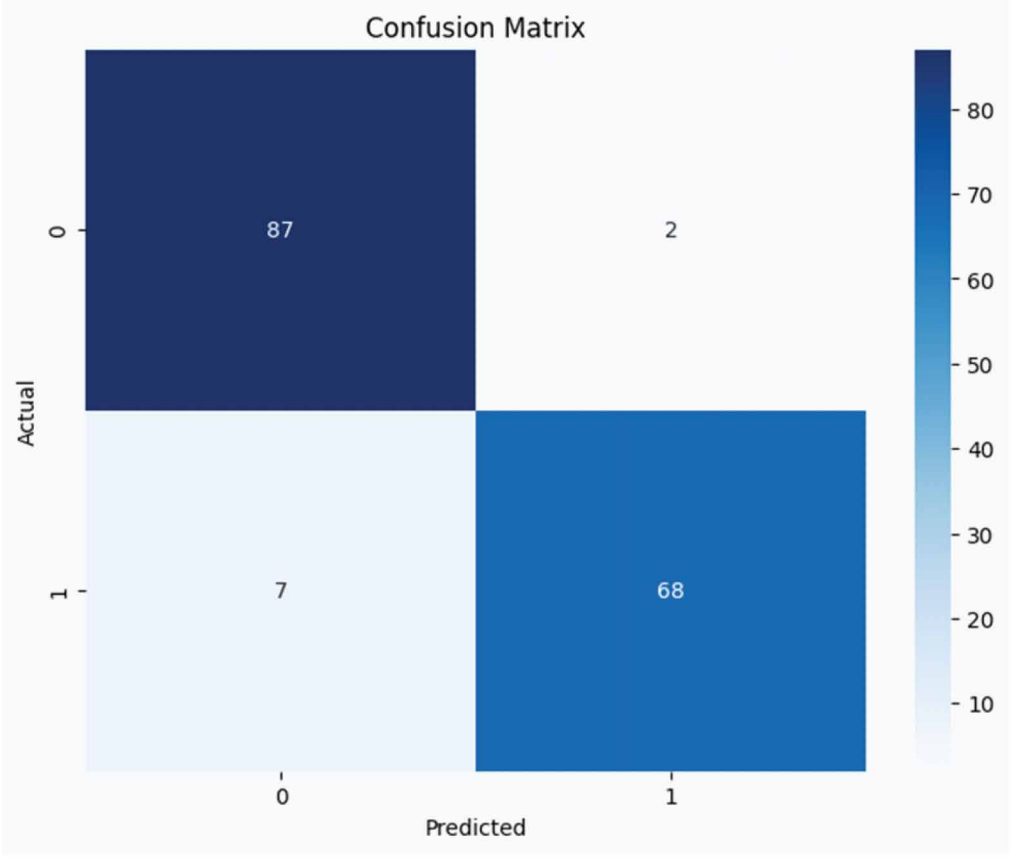
This model had the most increase in compared to its score before optimizing with other two models.

Figure 15 represents the confusion matrix for Neural Network model.

ANALYSIS OF RESULTS AND EVALUATING MODEL

To summarize the procedures within this study, a supervised model decision tree was used to predict recurrence based of all 32 features and the prediction was: 0.378. This means that if one were to take a patient with all the same 32 features as the ones used in the METABRIC dataset, one could predict the breast cancer would recur with 37% accuracy. For a medical project that involves human lives this prediction score is very low.

Figure 15. Confusion matrix for Neural Network model



Feature selections have been used to make the model more accurate, and then proceeded to apply several supervised models on those selected features. Table 3 shows the full summary of the MAE, MSE, and RSE results of the prediction models.

Table 3. A full summary of the MAE, MSE, and RSE results of the prediction models

Model	Mean Absolute Error	Mean Squared Error	R Score	F1 Score	Precision	Accuracy	RSE
Linear Regression	0.375	0.181	***	***	***	***	0.425
Decision Tree	***	***	0.790	0.595	0.745	0.831	***
Random Forest	0.378	0.378	***	***	***	***	0.615
Support Vector Machine	***	***	0.785	0.741	0.762	0.812	***
Neural Network	***	***	0.770	0.741	0.756	0.801	***

The reason that these models produce similar errors is because the algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees. In the case of a regression problem, the final output is the mean of all the outputs. The success of a random forest highly depends on using uncorrelated decision trees. If we use the same or very similar trees, the overall result will not be much different than the result of a single decision tree.

These error calculations assess the difference between the observed and predicted values. When a model has no error, the error equals zero. As model error increases, its value increases. Because we are trying to predict if a patient will have breast cancer recurrence or not, it is important to look at error calculations. Most math-based classification models will not predict 0 or 1, rather they'll predict a value between 0.0 and 1.0. Two common ways to determine the accuracy of a prediction model are to compute the mean squared error (where smaller values are better and 0.0 means perfect prediction) and to compute the predictive accuracy (the percentage of correct predictions). RSE is a good measure of accuracy, but only to compare forecasting errors of different models or model configurations for a particular variable and not between variables, as it is scale-dependent. Because different models are being used and maintain the same features throughout modeling, RSE is a good measure for the model's accuracy.

Several models have been used based using the newly selected features to make the models more accurate and received the following results. All those five models are very powerful models in machine learning, and we got the best accuracy with a Decision Tree model that can show based on our dataset type, this model works better. For analysis models' performance after optimizing, we optimize three of our models that had better accuracy and the accuracy of all of three models increased. For the Decision Tree this score increased more than 0.01 for the Support Vector Machine this increased less than 0.01. But for neural networks we had the highest increase, that was 0.07. That means in this model as the number of layers and nodes increase, the loss decreases and finally the accuracy increases. But because the amount of accuracy (the correct prediction) is the most important factor for us to decide which model works better. The Decision Tree would be the best model for our dataset.

CONCLUSION

Breast cancer is the most common female cancer in the US, the second most common cause of cancer death in women, and the main cause of death in women ages 40 to 59. Because of the high prevalence and mortality of breast cancer among women, it is important and necessary to explore and develop new techniques to aid doctors and breast cancer specialists to discover cancer early on and if there are high chances for recurrence. Subsequently, machine learning based prediction models are being investigated to help predict a patient's chance for recurrence and develop a treatment process to prevent it.

To predict the probability for recurrence, there is a need to find features/factors of a patient that have the highest risk factors before modeling the BCR prediction. Most existing works on breast cancer recurrence prediction use one of the feature selection techniques to find such features/factors. Some of the major problems and shortcomings of models for recurrence prediction have been that forecasting models have low accuracy when compared to conventional models. In others, sample datasets were too small (less than 300 instances) to be able to do a significant prediction. Others compared prediction models to find significant features of recurrence within 10 years of breast cancer surgery.

In this study, 6 different feature selection schemes including SelectkBest, LinearSVC, Decision Tree-Feature Importance, Recursive Feature Elimination, Pearson Correlation, and Chi Square Tests have

been used to find the highest risk factors for breast cancer recurrence. These schemes together give a better idea about the features that might have significance on BCR, compared to a single feature selection scheme used by most existing works. By getting the union of all the features produced by selection schemes, it was determined that the attributes that contributed the most to BCR include ER Status, Pam50 + Claudin-low subtype, Relapse Free Status (Months), Tumor Size, PR Status, and Radio Therapy. The new dataset was created by using the highest risk factors common to different feature selection schemes to build the breast cancer prediction model.

The five powerful models in machine learning algorithms that were used include Linear Regression, Decision tree, Random Forest, Support Vector Machine, and Neural Network. The best accuracy has been achieved with the Decision Tree model. The Decision Tree would be the best model for the dataset that is used since the amount of accuracy (the correct prediction) is the most important factor to decide which model works better.

FUTURE RESEARCH DIRECTIONS

There are other factors including ethnicity/race, obesity, and genetic-family history among other factors that have significance on breast cancer recurrence. These factors, however, were not available in the dataset that was used in this study. In the future, a larger and more detailed data set can be developed which covers additional attributes related to a patient's medical history. Unsupervised machine learning can be applied on such datasets to identify clusters of patient records and the similarity between them.

Finally, one major problem in defining a universal BCR prediction model is the lack of standard mechanism for defining the features/attributes of the breast cancer datasets. Each breast cancer dataset has its own set of features defined for the dataset. If the same set of attributes/features have been used in provided datasets, then it would be possible to define a universal BCR prediction model.

REFERENCES

Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., & Silva, D. C. (2016). Predicting breast cancer recurrence using machine learning techniques: A systematic review. *ACM Computing Surveys*, *49*(3), 52.

Ahmad, A. (2013). Pathways to breast cancer recurrence. *ISRN Oncology*, *290568*, 1–16. Advance online publication. doi:10.1155/2013/290568 PMID:23533807

Ahmad, L.G., Abbas, Eshlaghy, T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A.R. (2013). Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Health & Medical Journal of Informatics, 4*(2).

Alva, N. (2018). *Using machine learning techniques to predict the recurrence of breast cancer.* Retrieved from https://www.linkedin.com/pulse/using-machine-learning-techniques-predict-recurrence-breast-alva/

American Cancer Society. (2021). *Breast Cancer Facts & Figures*. Retrieved from https://www.cancer.org/

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4). Springer.

Bisong, E. (2019). More Supervised Machine Learning Techniques with Scikit-learn. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress. doi:10.1007/978-1-4842-4470-8_24

Bradley, K. T. (2007). *Prognostic and Predictive Factors in Breast Cancer.* Retrieved from http://www.cap.org

Breast Cancer Q & A/Facts and Statistics. (2012). Retrieved from https://www.komen.org/bei/bhealth/QA/q-and-a.asp

Breastcancer.org. (n.d.). *U.S. Breast Cancer Statistics*. Retrieved from http://www.breastcancer.org / symptoms/ understand_bc/statistics

Cancer.net. (2023). *Breast Cancer: Statistics, 2/2023*. https://www.cancer.net/cancer-types/breast-cancer/statistics

Chacón, R. D., & Costanzo, M. V. (2010). Triple-negative breast cancer. *Breast Cancer Research*, *12*(S2, S3), S3. Advance online publication. doi:10.1186/bcr2574 PMID:21050424

Chang, Y. W., & Lin, C. J. (2008). Feature ranking using linear SVM. In *Causation and prediction challenge* (pp. 53–64). PMLR.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. doi:10.38094/jastt20165

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*, *19*(S1), 65. doi:10.118612863-018-0633-8 PMID:30255764

Fan, Q., Zhu, C., & Yin, L. (2010). Predicting Breast Cancer Recurrence Using Data Mining Techniques. Academic Press.

Guo, J., Fung, B.C.M, Iqbal, F.P., Kuppen, J.K., Tollenaar, R.A.E.M., Mesker, W.E., & Lebrun, J.J. (2017). Revealing determinant factors for early breast cancer recurrence by decision tree. *Inf. Syst. Front*.

Heatmap. (n.d.). *jmp-Statistical Discovery*, https://www.jmp.com/en_us/statistics-knowledge-portal/exploratory-data-analysis/heatmap.html

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. doi:10.5194/gmd-15-5481-2022

Kaggle. (n.d.). *Breast Cancer (METABRIC).* Retrieved from https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric

Lachman, M. (2019). *Wisconsin Breast cancer dataset*. UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA. Retrieved from http://archive.ics.uci.edu/ml/datasets/breast+cancer

Lafourcade, A., His, M., Baglietto, L., Boutron-Ruault, M.-C., Dossus, L., & Rondeau, V. (2018). Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: The French E3N cohort. *BMC Cancer*, *18*(1), 171. doi:10.118612885-018-4076-4 PMID:29426294

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random Forest. In *International Conference on Information Computing and Applications, September*, (pp. 246-252). Springer. 10.1007/978-3-642-34062-8_32

Lou, S.J., Hou, M.F, Chang, H.T., Chiu, C.C, Lee, H.H, Yeh, S.C.J.& Shi, H.Y. (2020). Machine Learning Algorithms to Predict Recurrence within 10 Years after Breast Cancer Surgery: A Prospective Cohort Study. *Cancer, An Open Access Journal, 17*(12).

Makaba, T., & Dogo, E. (2019). A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms, *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 1-7. 10.1109/IMITEC45504.2019.9015889

Mannell, A. (2017). An overview of risk factors for recurrent breast cancer. *South African Journal of Surgery. Suid-Afrikaanse Tydskrif vir Chirurgie*, *55*(1), 29–34. PMID:28876555

Massafra, R., Latorre, A., Fanizzi, A., Bellotti, R., Didonna, V., Giotta, F., Forgia, D. L., Nardone, A., Pastena, M., Ressa, C. M., Rinaldi, L., Russo, A. O. M., Tamborra, P., Tangaro, S., Zito, A., & Lorusso, V. (2021). A Clinical Decision Support System for Predicting Invasive Breast Cancer Recurrence: Preliminary Results. *Frontiers in Oncology*, *11*(March), 576007. doi:10.3389/fonc.2021.576007 PMID:33777733

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. doi:10.38094/jastt1457

Mayoclinic.org. (n.d.). *Recurrent breast cancer*. Retrieved from https://www.mayoclinic.org/diseases-conditions/recurrent-breast-cancer/symptoms-causes/syc-20377135

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine Learning: Algorithms and Applications* (1st ed.). CRC Press. doi:10.1201/9781315371658

Mosayebi, A., Mojaradi, B., Ali Bonyadi, A., Naeini, S. H., & Hosseini, K. (2020, October 15). Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PLoS One*, *15*(10), e0237658. Advance online publication. doi:10.1371/journal.pone.0237658 PMID:33057328

Mulatu, D., & Gangarde, R. R. (2017). Survey of Data Mining Techniques for Prediction of Breast Cancer Recurrence. *International Journal of Computer Science and Information Technologies*, *8*(6), 599–601.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society, 18*(6), 275-285.

NPI. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Nottingham_Prognostic_Index

Ohno-Machado, L. (2001). Modeling medical prognosis: survival analysis techniques. *J Biomed Inform, 34*, 428-39.

Pritom, A. I., Sabab, S. A., Munshi, A. R., & Shihab, S. (2016). Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique. In *19th International Conference on Computer and Information Technology* (pp. 310-314). North South University. 10.1109/ICCITECHN.2016.7860215

Raschka, S., & Mirajalili, V. (2017). Python machine learning machine learning and deep learning with Python, scikit-learn, and TensorFlow. Academic Press.

Roberto Cesar, M. O., German, L. B., Paola Patricia, A. C., Eugenia, A. R., Elisa Clementina, O. M., Jose, C. O., Marlon Alberto, P. M., Fabio Enrique, M. P., & Margarita, R. V. (2020). Method Based on Data Mining Techniques for Breast Cancer Recurrence Analysis. *Advances in Swarm Intelligence: 11th International Conference, ICSI 2020, Belgrade, Serbia, July 14–20, 2020 Proceedings*, *12145*, 584–596. doi:10.1007/978-3-030-53956-6_54

Shikama, N., Sekiguchi, K., & Nakamura, N. (2011). Management of locoregional recurrence of breast cancer. *Breast Cancer, 18*(4), 252-8.

Siegel, R., Naishadham, D., & Jamal, A. (2012). Cancer Statistics. *CA: a Cancer Journal for Clinicians*, *62*(10).

SVM. (n.d.). *Introduction to Support Vector Machines (SVM).* Retrieved from **h**ttps://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/

Tableau. (n.d.). *Tableau-An Introduction.* Retrieved from https://cedar.princeton.edu/sites/g/files/toruqf1076/files/media/introduction_to_tableau_training_0.pdf

Tallarida, R. J., & Murray, R. B. (1987). Chi-square test. In *Manual of pharmacologic calculations* (pp. 140–142). Springer. doi:10.1007/978-1-4612-4974-0_43

Toloie Eshlaghy, A. (2013). Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics*, *2013*(4), 2. doi:10.4172/2157-7420.1000124

Torre, L., Siegel, R. L., Ward, E. M., & Jemal, A. (2016). Global Cancer Incidence and Mortality Rates and Trends—An Update. *Cancer Epidemiology, Biomarkers & Prevention*, *25*(1), 16–27. doi:10.1158/1055-9965.EPI-15-0578 PMID:26667886

Weka 3. (n.d.). *Data Mining with open-source machine learning software.* Retrieved from http://www.cs.waikato. ac.nz/ml/ weka/

Yang, P. T., Wu, C. C., Wu, W. S., Shih, Y. N., Hsieh, J. C. H., & Hsu, J. L. (2021). Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning. *Open Medicine: a Peer-Reviewed, Independent, Open-Access Journal*, *16*(1), 754–768. doi:10.1515/med-2021-0282 PMID:34027105

Zhou, V. (2019). *Machine Learning for Beginners: An Introduction to Neural Networks.* Retrieved from https://towardsdatascience.com/machine-learning-for-beginners-an-introduction-to-neural-networks-d49f22d238f9

ADDITIONAL READING

Breast Cancer Risk Assessment Tool (Gail Model). (n.d.). Retrieved from http://ww5.komen.org/BreastCancer/GailAssessmentModel.html

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Houston, A., Chen, H., Hubbard, S. M., Schatz, B. R., Ng, T. D., Sewell, R. R., & Tolle, K. M. (1999). Medical Data Mining on the Internet: Research on a Cancer Information System. *Artificial Intelligence Review*, *13*(5-6), 437–466. doi:10.1023/A:1006548623067

Quinlan, J. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann.

KEY TERMS AND DEFINITIONS

Attributes/Features Selection Schemes: Attribute/Feature selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

Breast Cancer: A malignant (cancerous) growth that begins in the tissues of the breast in an uncontrolled way.

Breast Cancer Recurrence: Is characterized by the cancer "coming back" after at least a year of remission after the treatment.

Machine Learning Algorithms: A machine learning algorithm is the method by which the AI system conducts its task, generally predicting output values from given input data.

Prognosis: The prediction of the survivability of a disease.

Prognostic Factor: A measurable variable that is used in the prediction of survival of breast cancer.

Prognostic Tool: A software tool that uses risk factors (i.e., age, ethnicity, genetic) to estimate breast cancer survivability rate.

Survival Analysis: A field in medical prognosis that deals with the application of various methods to estimate the survival of a particular patient suffering from a disease.