BREAST CANCER DETECTION IN THE PHILIPPINES USING MACHINE LEARNING APPROACHES: A PILOT STUDY

by

Maria Maura S. Tinao
University of the Philippines Open University
mariamaura.tinao@upou.edu.ph

Ruth B. Rodriguez
University of the Philippines Open University
ruth.rodriguez@upou.edu.ph

Eunelfa Regie F. Calibara
University of the Philippines Open University
eunelfaregie.calibara@upou.edu.ph

ABSTRACT

This study presents a comprehensive exploration of breast cancer prediction using advanced machine learning (ML) techniques, namely K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. The primary objective is to identify those who are at high risk of having breast cancer and to enable early detection prior to starting treatment. By harnessing the potential of these ML algorithms, this research aims to enhance healthcare outcomes and decision-making in breast cancer management, focusing on predictive accuracy and variable relationships.

Using 10 significant breast cancer characteristics derived from Rabiel et al.'s (2022) study, this study applies a variety of ML methods to aid in early detection, prevention, and treatment. The chosen metrics—accuracy, Jaccard index, F1-score, and log loss—evaluate model performance, while Pearson Correlation and linear regression analyses elucidate associations and directions of influence among variables.

Utilizing a sample of 112 randomly selected women from Olongapo City and Zambales Province, the study employs anonymized data to maintain confidentiality and ethical considerations. Python programming, along with the Jupyter Notebook application in Anaconda, is used for model development and testing.

The analysis of breast cancer features reveals valuable insights. The prevalence of life event stress, family problems, and marital status emerges, while a minority report a personal history of breast cancer. Key findings include an average personal other cancer history score of 1.70 and a substantial representation of individuals with a family history of breast cancer. Smoking behavior, breast density, and other variables are also considered. Correlation analyses indicate weak positive relationships between several variables and breast density. Linear regression demonstrates a

limited explanatory power of predictor variables, suggesting the need for refinement and additional predictors to enhance model performance.

The study underscores the significance of performance evaluation metrics in ML model application. Notably, the KNN model attains the highest accuracy rate of 0.8696, correctly classifying 86.96% of breast cancer cases. Contrastingly, the Decision Tree model achieves a lower accuracy rate, while SVM and Logistic Regression present distinct trade-offs between precision and recall.

Keywords: Cancer, Breast Cancer, Breast Cancer Detection, Healthcare, Machine Learning Techniques, Python Programming.

INTRODUCTION / BACKGROUND OF THE STUDY

Breast cancer (BC) remains a devastating illness that disproportionately affects women, not only in Asia but worldwide, making it a significant global health concern. According to a study by Lim, et al. (n.d.) 45.4% of breast cancer patients are of Asian ethnicity, and the Philippines ranks fifth in the highest incidence of this disease in East Asia and the Pacific region. Unbelievably, there are 27,000 new cases of breast cancer reported each year, with 9,000 of those cases tragically resulting in mortality. Alarming statistics from the Department of Health (DOH) indicate that 3 out of every 100 women in the nation will receive a breast cancer diagnosis during their lifetime. Dr. Norman San Agustin, a respected health expert, has described the prevalence of breast cancer in the Philippines as "staggering". Compared to COVID-19, breast cancer rates continue to rise, while screening rates for breast and cervical cancer remain abysmally low, with just 1% of the 54 million women in the country currently receiving cancer screenings.

In low- and middle-income countries like the Philippines, many women struggle to manage this disease due to various barriers, including limited access to breast cancer screening, low health literacy, and prohibitive healthcare costs. Early diagnosis is vital for improving survival rates, but unfortunately, it remains a challenge.

Despite the grim situation, there is confidence in the technological advancements that have been transforming the healthcare industry. Arthur Samuel, an American pioneer in computer

gaming and artificial intelligence, defined Machine Learning (ML) as a subfield of computer science that enables "computers to learn without being explicitly programmed," allowing them to learn patterns from data. The application of machine learning in the health sector aims to improve patient outcomes and provide medical insights that were previously unavailable. It offers a way to validate physicians' reasoning and decisions through predictive models. Consequently, ML simplifies the process of training machines and constructing predictive models, benefiting both patients and healthcare providers.

Machine learning (ML) is particularly valuable for analyzing patient data to aid in the early diagnosis, prevention, and treatment of various illnesses, including breast cancer. In countries with low survival rates, like the Philippines, ML can accurately identify patterns and predict diseases in similar cases, addressing the delay in diagnosis and improving outcomes.

Given its unique ability to identify critical features from complex breast cancer datasets, machine learning (ML) is widely regarded as the preferred approach for BC pattern classification and forecasting. In simple terms, machine learning has the potential to uncover hidden patterns in data, leading to improved cancer prediction and management.

RESEARCH OBJECTIVES

The main objective of this study is to predict breast cancer using several and recognized machine learning (ML) approaches such as K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. The goal of analyzing the findings of these algorithms is to identify individuals who are at high risk of breast cancer and to assist early detection even before the patient receives decisive treatment. In addition, this study aims to describe in detail the pertinent breast cancer characteristics found in the sample and explore potential relationships between these features. By achieving these research goals, it could lead to better breast cancer prediction and improve healthcare outcomes for affected individuals.

CONCEPTUAL/THEORETICAL FRAMEWORK

Machine learning (ML) has emerged as a significant driving force in the healthcare industry, especially in the domain of artificial intelligence. This technology enables systems to autonomously learn from data and identify patterns, reducing the need for extensive human intervention. Unlike traditional programming, ML algorithms are presented with data, allowing researchers to draw their own conclusions and make informed decisions. In healthcare, ML is invaluable in deciphering the vast amount of data generated daily within electronic health records, revealing patterns and insights that would be difficult to discern manually. As ML gains broader adoption in healthcare, there is an opportunity for healthcare providers to adopt a more predictive approach, leading to advancements in precision medicine.

Elsadig, et al. (2023) underline the essential impact of early detection in enhancing breast cancer patient survival rates in their study. They illustrate the effectiveness of artificial intelligence, specifically machine learning, in improving breast cancer diagnosis. Multi-layer perception (MLP), support vector machine (SVM), and stack classifiers stand out among the classifiers investigated, with SVM attaining the greatest accuracy rate of 97.7% and remarkably low classification errors (0.029 false negatives and 0.019 false positives).

El Massari, et al. (2023) propose an ontological model based on machine learning techniques to distinguish between patients with aggressive and benign breast cancer. Their study, "Effectiveness of Applying Machine Learning Techniques and Ontologies in Breast Cancer Detection," presents rules implemented in the ontological reasoner using the Semantic Web Rule Language (SWRL). The results demonstrate a remarkable prediction accuracy of 97.1%.

Reza Rabiel, et al. (2022) introduce an ML technique using the Random Forest (RF) algorithm to predict breast cancer. With an accuracy rate of 80%, sensitivity of 95%, specificity of 80%, and an area under the curve (AUC) of 0.56, this approach shows promise in improving breast cancer diagnosis.

In contrast, Fannizi, et al. (2020), propose a machine learning approach to multiscale texture analysis for breast microcalcification diagnosis, acting as a support tool for radiologists.

The study utilizes the Speeded Up Robust Feature (SURF) and Minimum Eigenvalue Algorithm (MinEigenAlg) to train a Random Forest binary classifier on selected features using filter and embedded methods. The embedded method yields the best-performing results, with median AUC values of 98.16% and 92.08% and accuracies of 97.31% and 88.46% for normal/abnormal and benign/malignant classifications, respectively.

In a study conducted by high school students then Shreya Nag and Jaydeep Nag in May 2021, ML techniques like Random Forest and Logistic Regression were analyzed for their accuracy and performance in breast cancer prediction. The Random Forest algorithm demonstrated superior performance in malignant prediction, achieving an accuracy rate of 98%, while the Logistic Regression algorithm excelled in benign prediction with an accuracy rate of 99% using the Wisconsin Breast Cancer dataset from the UC Irvine repository. Their findings highlight the potential of ML techniques, even in the hands of young researchers.

Overall, the research presented in these studies showcases the significant impact of machine learning in breast cancer prediction, detection, and diagnosis. The application of ML algorithms in the healthcare domain holds promise for improving patient outcomes and revolutionizing the field of oncology.

The conceptual framework employed in this study involves the use of machine learning classification algorithms to predict breast cancer.

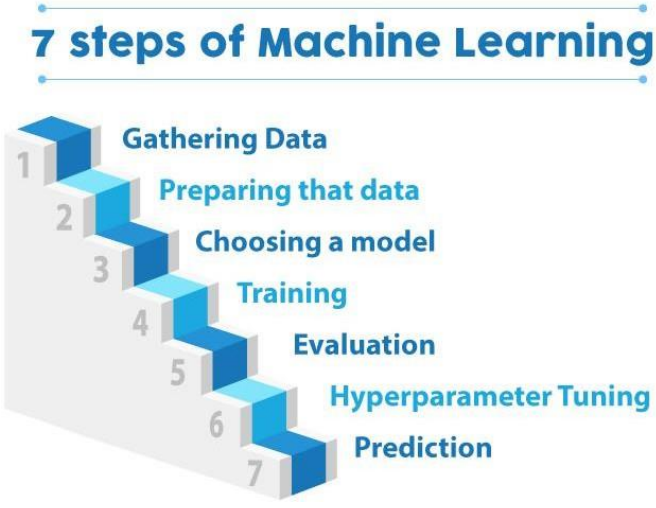

Figure 1. Machine Learning Classification Algorithm

The initial step in the process is collecting data, during which the problem is defined and the project's aim for machine learning is specified. Then, pertinent data is gathered from a variety of resources, including databases, APIs, web scraping, or pre-existing datasets. In the next stage, called data preprocessing, the data is cleaned up by dealing with missing values, outliers, and inconsistent formatting. To normalize features, data transformations like normalization or scaling are used. Categorical variables are encoded using methods like one- hot encoding or label encoding. Following that, the data is divided into training, validation, and testing sets.

The data is then visualized and examined via Exploratory Data Analysis (EDA) in order to gain knowledge about its distribution, correlations, and trends. Finding probable connections between features and the target variable is another step in the process. To better capture underlying data patterns, new features are created and existing ones are modified in the field of feature engineering. The emphasis is on choosing relevant attributes that improve the model's capacity for prediction.

Model selection is the process of selecting an acceptable machine learning method for a given problem. Versatile libraries like Scikit-Learn, TensorFlow, or PyTorch may be used. Model Training, which involves training the chosen model on the training dataset, is the next step in the process. The hyperparameters are then fine-tuned using methods like grid search or random search. Using appropriate metrics (accuracy, precision, recall, F1-score, etc.) on the validation dataset, the model's performance is evaluated during model evaluation. To increase the dependability of the results, procedures like cross-validation are frequently used.

The next step is model tuning, which entails parameter changes based on validation results to improve performance while being careful to prevent overfitting using strategies like regularization or dropout. Final Model Selection, when the model demonstrating the highest performance based on validation outcomes is chosen, is the culmination of these efforts. The performance of the resulting model in the actual world is then evaluated using the test dataset. The well-performing model is then deployed to make predictions on fresh, unforeseen data once confidence in its performance is established. This completes the process.

The four different machine learning classification algorithms used in this study are as follows: K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. Each algorithm processes the data differently, identifying patterns and creating decision boundaries to classify breast cancer cases into different categories.

The ultimate goal of this conceptual framework is to identify high-risk individuals and predict breast cancer cases accurately. By utilizing machine learning algorithms and evaluating their performance on test data, this study aims to contribute to the early detection and improved management of breast cancer.

RESEARCH METHOD

In this study, the ten (10) relevant breast cancer features from Rabiel, et al.'s study (2022) were utilized. These are: (1) age at the latest mammographic test; (2) Life event stress; (3)Marital status; (4) Smoker; (5) Biopsy; (6) Breast density; (7) Personal breast cancer history; (8) Personal history of other cancer; (9) Family breast cancer history; and (10) Family history of other cancer.

This quantitative research design utilizes machine learning (ML) techniques to analyze respondent data for early detection, prevention, and treatment of breast cancer. The study aims to measure the effectiveness of the ML algorithms in achieving accurate predictions and utilizes metrics such as accuracy, Jaccard, F1, and Logloss rates to evaluate the model's performance.

This research employs several ML algorithms, including K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. KNN is a classification method that identifies neighbors based on similarity to make predictions on the test dataset. The number of neighbors is determined by the parameter k, and the algorithm calculates the distance between test and training data to find similarities.

Decision trees are known for their efficiency and easy interpretation, presenting a flowchart-like structure with rules for data splitting. The algorithm selects the best predictive attribute to create a training model that predicts the target variable's class or value.

SVM, a powerful ML algorithm for classification and regression tasks, finds the optimal decision boundary that separates different groups of data on a graph.

Logistic Regression is employed for categorical variable classification, predicting the probability of class labels based on independent variables. The model calculates the logistics of the result by summing the input features.

To demonstrate the degree and direction of relationships between variables, Pearson Correlation and linear regression tests were conducted. Correlation-based forecasts are more reliable and realistic. Linear regression predicts the relationship between two variables, assuming a linear relationship, and determines the best-fitting line.

The research design's utilization of these ML algorithms and statistical tests allows for a comprehensive analysis of breast cancer prediction and relationships between variables. The findings can potentially contribute to improving healthcare outcomes and decision-making in breast cancer management.

Sample of the study

Finding an active and up-to-date cancer registry in the Philippines proved to be a challenging task, especially when seeking specific information on breast cancer. Based on the findings of the Rabiel, et al. (2022) investigation, ten (10) important breast cancer features were chosen. These are: (1) age at the latest mammographic test; (2) Life event stress; (3) Marital status; (4) Smoker; (5) Biopsy; (6) Breast density; (7) Personal breast cancer history; (8) Personal history of other cancer; (9) Family breast cancer history; and (10) Family history of other cancer. Additionally, due to the sensitive nature of this study, finding an appropriate research sample posed further difficulties. To address these challenges, the study initially gathered information from a loose and informal group of breast cancer suspects and patients. Subsequently, the sample size expanded to include 112 randomly selected women from Olongapo City and Zambales Province. Online surveys and interviews were conducted with women who had undergone a mammogram. The use of anonymized data ensured that no written permission was required from the participants, ensuring confidentiality and ethical considerations.

Models' Implementation Tools

For the implementation of the machine learning models, the study utilized the Python programming language and the Jupyter Notebook application in Anaconda. Jupyter Notebook is a web-based, interactive computing environment that enables the creation of human-readable documentation while describing the data analysis process. With this powerful tool, the ML models were developed and tested, facilitating the evaluation and interpretation of the results. Statistical Package for Social Science or SPSS was utilized to compute linear regression

Overview and simplified steps of Machine Learning Models

1. K-Nearest Neighbors (KNN): KNN is a simple and intuitive classification algorithm based on the similarity of data points.

1.1. Choose the value of K (number of neighbors).
1.2. Compute the distance between the input data point and all other data points in the training set.
1.3. Select the K-nearest neighbors based on the computed distances.
1.4. Assign a class label to the input data point based on majority voting among the K-nearest neighbors.
1.5. Make predictions for new data points by repeating steps 1.2-1.4.

2. Decision Trees are hierarchical structures that make decisions based on feature values.

2.1. Select a feature from the dataset as the root node of the tree.
2.2. Split the data based on the selected feature, creating child notes.
2.3. Choose the best feature and split again at each child node.
2.4. Continue recursively until a stopping criterion is met (e.g., maximum depth, minimum samples per leaf).
2.5. Assign a class label to the leaf nodes based on majority voting of training samples in that node.
2.6. To make predictions, traverse the tree from the root to a leaf node based on feature values.

3. Logistic Regression is a binary classification algorithm that models the probability of a data point belonging to a certain class.

3.1. Initialize the weights and bias.

3.2. Compute the weighted sum of input features and apply the sigmoid function to get the predicted probability.

3.3. Define a loss function (e.g., cross-entropy) that qualifies the difference between Predicted probabilities and actual labels.

3.4. Use an optimization algorithm (e.g., gradient descent) to minimize the loss function and update the weights and bias.

3.5. To make predictions, compute the predicted probability and classify based on threshold (e.g., 0.5).

4. Sector Vector Machine (SVM) is a binary classification algorithm that finds a hyperplane to separate data points of different classes. Choose the kernel function (linear, polynomial, radial basis function, etc.)

4.1. Select C, the regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error.

4.2. Formulate and solve the optimization problem to find the optional hyperplane that maximizes the margin between classes.

4.3. To classify new data points, evaluate which side of the hyperplane they fall on.

Validation method

The accuracy metric is a widely used evaluation measure to assess how well a model predicts the correct class labels.

The formula for accuracy is as follows:

$$\text{Accuracy} = \mathbf{(TP + TN) / (TP + TN + FP + FN)}$$

where:

•**TP** (*True Positive*) represents the number of positive instances correctly predicted by the classification model.

•**TN** (*True Negative*) represents the number of negative instances correctly predicted by the classification model.

•**FP** (*False Positive*) represents the number of negative instances incorrectly predicted as positive by the model.

•**FN** (*False Negative*) represents the number of positive instances incorrectly predicted as negative by the model.

Confusion Matrix

The confusion matrix is a table that presents a summary of the model's performance by comparing predicted class labels against the actual class labels. It helps visualize the classification errors and correct predictions made by the model.

The confusion matrix typically looks like this:

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Table 1. Confusion Matrix

The confusion matrix is then used to evaluate the model's performance, showing the distribution of correct and incorrect predictions. The ultimate goal is to minimize the false negatives and false positives to enhance the model's predictive capabilities.

RESULTS AND DISCUSSION

Breast cancer is a significant health concern worldwide, and early detection plays a crucial role in improving patient outcomes. In recent years, machine learning algorithms have shown promise in aiding medical professionals in accurately diagnosing breast cancer based on various features extracted from mammograms, such as texture, shape, and size. This research aims to explore the effectiveness of four popular machine learning models in detecting breast cancer: K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression.

Profile of the Respondents

Variables	Frequency	Percentage
1. Age at latest mammographic test		
18-24	5	4.5
25-34	16	14.3
35-44	5	4.5
45-54	33	29.5
55-64	10	8.9
TOTAL	**112**	**100.0**
2. Life event stress		
Family problem	40	35.7
Death of loved ones	14	12.5
Work related problems	20	17.9
Others	18	16.1
None	20	17.9
TOTAL	**112**	**100.0**
3. Marital status		
Single	41	36.6
Married	49	43.8
Widowed	14	12.5
Divorced	5	4.5
No answer	3	2.7
TOTAL	**112**	**100.0**
4. Personal breast cancer history		
Yes	13	11.6
No	99	88.4
TOTAL	**112**	**100.0**
5. Personal other cancer history		
Yes	34	30.4
No	78	69.6
TOTAL	**112**	**100.0**
6. Family history of breast cancer		
Yes	22	19.6
No	90	80.4
TOTAL	**112**	**100.0**

7. Family history of other cancer		
Yes	52	46.4
No	60	53.6
TOTAL	**112**	**100.0**
8. Smoker		
Yes	9	8.0
No	103	92.6
TOTAL	**112**	**100.0**
9. Breast density		
Fatty tissue	27	24.1
Heterogeneously dense	6	5.4
Extremely dense	11	9.8
Dense	22	19.6
No info	29	25.9
Glandular and fibrous tissue	17	15.2
TOTAL	**112**	**100.0**
10. Biopsy		
No malignancy detected	37	33.0
Malignancy detected	47	42.0
Not applicable	28	25.0
TOTAL	**112**	**100.0**

Table 2. Profile of the Respondents based on the ten (10) Distinct Characteristics of Breast Cancer

Table 2 depicts the response profile based on the ten (10) unique breast cancer features, which are as follows:

1. Age at latest mammographic test

"Age at latest mammographic test" refers to the ages at which respondents have undergone a mammogram, a medical imaging procedure used to detect breast cancer or other breast-related issues. The mean age at the latest mammographic test is 3.37. This value represents the average age of individuals in the dataset who have undergone the test.

The standard deviation of 1.280 indicates the dispersion or spread of ages around the mean. A higher standard deviation suggests greater variability in ages. The highest frequency (33) and percentage (29.5%) are observed in the age range of 45 to 54, indicating that this age group has the most individuals who have undergone mammographic tests. The next highest frequency is in the age range of 25 to 34, with 16 individuals (14.3%). The age ranges of 35 to 44 and 65 to 74 also have noticeable frequencies of 5 and 10, respectively. The age range of 18 to 24 has the lowest frequency of 5 (4.5%).

2. Life event stress

The provided data showcases the distribution of different stressors among respondents. The mean life event stress score of 2.68 indicates the average level of stress reported by the participants. The standard deviation of 1.532 reflects the variability or spread of stress scores around the mean.

Among the identified stressors, family problems (1.0, *see Figure 2*) appear to be the most prevalent, with 35.7% of respondents indicating this as a significant source of stress. The death of a loved one follows, with 12.5% of respondents reporting it as a stressor. Work-related problems account for 17.9% of stress sources, and various other factors contribute to 16.1% of stressors. It's notable that 17.9% of respondents did not specify any particular stressor, suggesting a general sense of stress without a defined cause.

Figure 2. Histogram of Life event stress

3. Marital Status

The data presents information about the marital status distribution within the surveyed group. The mean marital status score of 1.86 gives an average indication of respondents' marital status, while the standard deviation of 0.985 reflects the variability around this average. The majority of respondents are either "Married" or "Single," with "Married" being the most common category.

The "Widowed" category represents individuals who have experienced the loss of a spouse, and a small portion of respondents chose not to answer.

4. Personal breast cancer history

The data offers insights into individuals' personal histories of breast cancer within the surveyed group. The mean personal breast cancer history score of 1.88 gives an average indication of respondents' personal breast cancer experiences, while the relatively low standard deviation of 0.332 suggests a relatively moderate variability around this average. A small proportion of respondents answered "Yes" to having a personal history of breast cancer, indicating that a notable segment of the surveyed group has been diagnosed with breast cancer at some point. Conversely, a larger percentage of respondents answered "No," = yes, = No (2.0, *see Figure 3*), suggesting that most individuals in the group do not have a personal history of breast cancer.

The data offers insights into individuals' personal history of breast cancer within surveyed group. The mean personal breast cancer history score of 1.88 gives an average indication of respondents' personal breast cancer experiences, while the relatively low standard deviation of 0.332 suggests a relatively moderate variability around this average.

A small proportion of respondents answered "Yes"(1.0, *see Figure 3*) to having a personal history of breast cancer, indicating that a notable segment of the surveyed group has been diagnosed with breast cancer at some point. Conversely, a larger percentage of respondents answered "No," (2.0, *see Figure 3*) suggesting that most individuals in the group do not have a personal history of breast cancer.

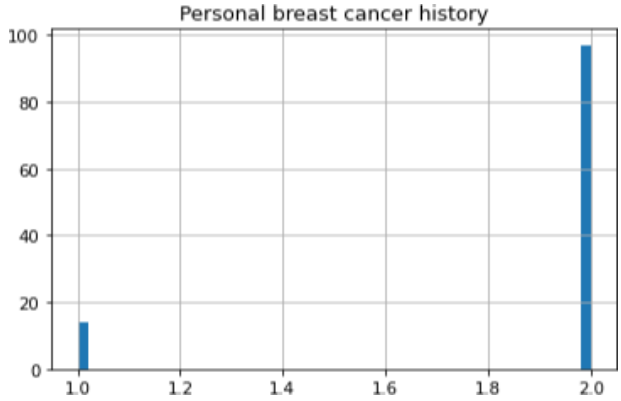

Figure 3. Histogram of Personal Breast Cancer

5. Personal other cancer history

The mean personal other cancer history score of 1.70 gives an average indication of respondents' experiences with other types of cancer. The relatively low standard deviation of 0.462 suggests a relatively moderate variability around this average.

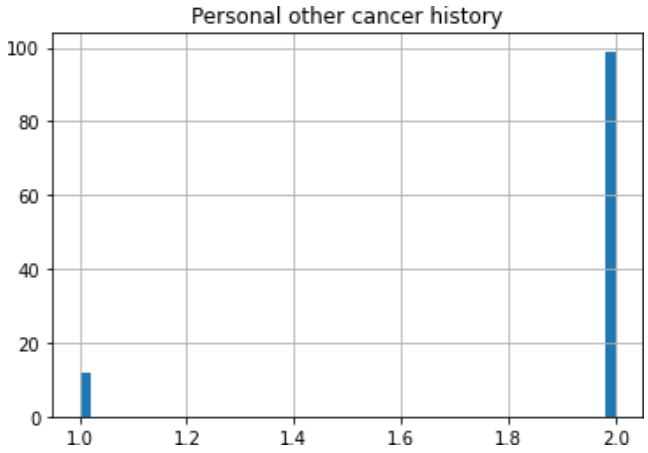

Figure 4. Histogram of Personal Other Cancer

Those who answered "Yes"(*see Figure 4*) to having a personal history of cancers other than breast cancer. The frequency of 34 individuals represents 30.4% of the respondents. This suggests that a significant proportion of the surveyed group has experienced other types of cancer. The "No" category (*see Figure 4*) consists of individuals who do not have a personal history of cancers other than breast cancer. The frequency of 78 respondents represents the majority, making up 69.6% of the sample. This indicates that a larger portion of the surveyed group has not experienced other types of cancer.

6. Family history of breast cancer

The mean family history breast cancer score of 1.80 gives an average indication of respondents' family history experiences with breast cancer. However, the high standard deviation of 19.6 indicates a significant variability around this average, suggesting a wide range of family history scores. The frequency of 22 individuals represents 19.6% of the respondents. This suggests that a notable proportion of the surveyed group has a family history of breast cancer. The "No" (2.0 *see*

Figure 5) category consists of individuals who do not have a family history of breast cancer. The frequency of 90 respondents represents the majority, making up 80.4% of the sample.

This indicates that the larger portion of the surveyed group does not have a family history of breast cancer.

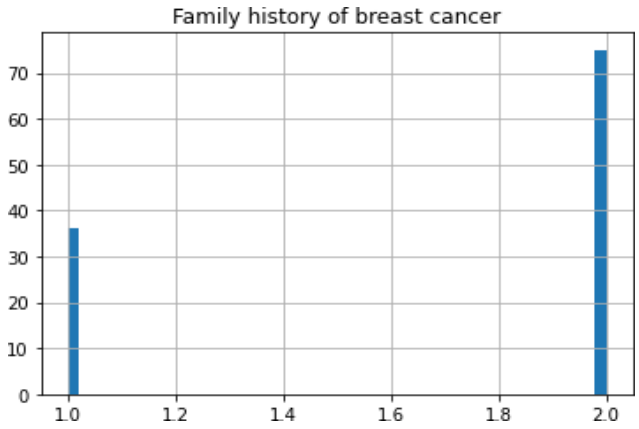

Figure 5. Histogram of Family History of Breast Cancer

7. Family history of other cancer

The mean family history other cancer score of 1.54 gives an average indication of respondents' family history experiences with other types of cancer. The relatively low standard deviation of 0.501 suggests a moderate variability around this average

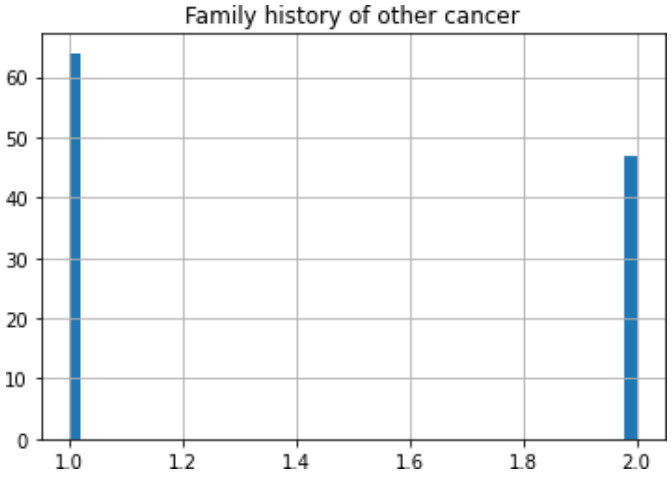

Figure 6. Histogram of Family History of Other Cancer

Respondents under "Yes" category (*see Figure 6*) have a family history of cancers other than breast cancer. The frequency of 52 individuals represents 46.4% of the respondents. This suggests that a significant proportion of the surveyed group has relatives who have been diagnosed with other types of cancer. The opposite or "No" category consists of individuals who do not have a family history of cancers other than breast cancer. The frequency of 60 respondents represents the majority, making up 53.6% of the sample. This indicates that a larger portion of the surveyed group does not have a family history of other cancers.

8. Smoker

The mean score of 1.92 represents the average value of the smoking scores. This numerical value provides insight into the central tendency of the data, indicating the average smoking behavior within the surveyed group. The standard deviation of 0.273 indicates the extent of variability or dispersion of smoking scores around the mean. A lower standard deviation suggests that the data points are closely clustered around the mean.

Respondents include individuals who are smokers. The frequency of 9 individuals represents 8.0% of the respondents. This suggests that a relatively small proportion of the surveyed group are smokers. Non-smokers on the other hand, consist of 103 respondents representing the majority, making up 92.0% of the sample. This indicates that the larger portion of the surveyed group are non-smokers.

9. Breast Density

Breast density refers to the proportion of different types of tissue (fatty tissue, glandular and fibrous tissue) within the breast. The mean score of 3.63 represents the average value of the breast density scores. The standard deviation of 1.801 indicates the extent of variability or dispersion of breast density scores around the mean. A higher standard deviation suggests that the data points are spread out from the mean.

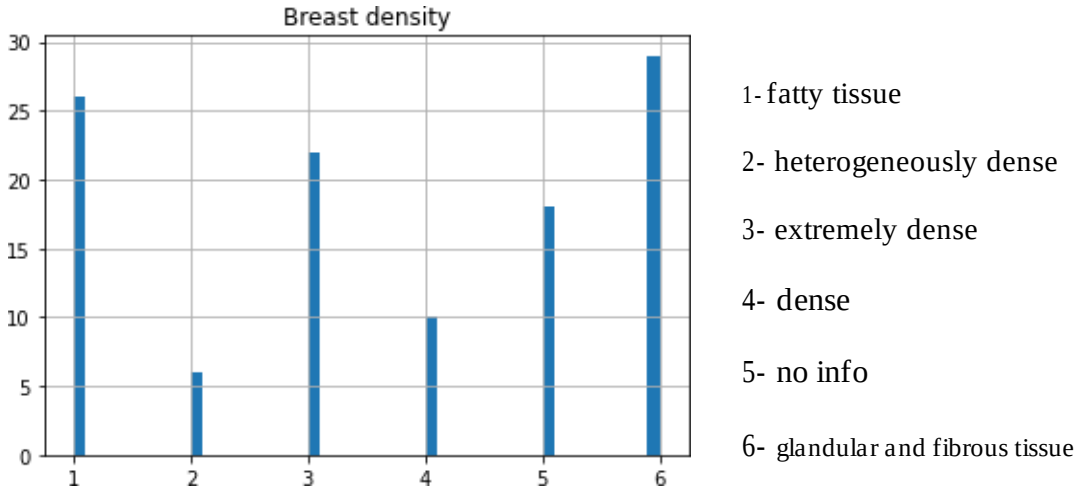

Figure 7. Breast Density

This category includes breasts with a higher proportion of fatty tissue. The frequency of 27 individuals represents 24.1% of the respondents. The frequency of 6 individuals represents 5.4% of the sample for the heterogeneously dense. Individuals in this category have breasts with an extremely dense tissue composition. This category has a frequency of 27 and a percent of 24.1. The "Dense" category includes breasts with a denser tissue composition. The frequency of 22 respondents represents 19.6% of the sample. The accounts for cases of no information has a frequency of 29 respondents represents 25.9% of the sample. For the Glandular and fibrous tissue, this category indicates a higher proportion of glandular and fibrous tissue within the breasts. The frequency of 17 respondents represents 15.2% of the sample.

10. Biopsy

A biopsy is a medical procedure that involves the removal of a sample of tissue for examination to determine the presence of abnormalities, such as malignancy (cancer). The mean biopsy score of 1.92 provides an average indication of biopsy outcomes, while the standard deviation of 0.761 suggests a moderate spread of biopsy scores.

The category of “No malignancy” indicates cases where no malignancy (cancer) was detected in the biopsy sample. The frequency of 37 individuals represents 33.0% of the respondents. “Malignancy detected” respondents in this category had malignancy detected in their biopsy sample, indicating the presence of cancer. The frequency of 47 individuals represents

42.0% of the sample. The category of "Not applicable" where a biopsy has not been performed has frequency of 28 respondents represents 25.0% of the sample.

Correlation

Breast density is a crucial factor that can influence the likelihood of getting breast cancer. It improves risk assessment, diagnosis, and treatment planning's precision, accuracy, and thoroughness. The Pearson correlation coefficient is a statistical measure used to quantify the strength and direction of a linear relationship between two variables. Specifically, the hypothesis under investigation is whether there is a significant correlation between various factors: age at the latest mammographic test, life event stress, marital status, personal breast cancer history, personal other cancer history, family history of breast cancer, family history of other cancer, smoker, biopsy, and breast density.

The results are:

1. **Age at the Latest Mammographic Test and Breast Density**: The positive correlation coefficient of 0.153 suggests a weak positive relationship between age at the Latest Mammographic Test and breast density. This implies that as age increases, there is a slight tendency for breast density to increase. However, the p-value of 0.108 indicates that this relationship is not statistically significant.

2. **Life Event Stress and Breast Density**: The negative correlation coefficient of -0.099 indicates a weak negative correlation between life event stress and breast density. This suggests that higher levels of life-event stress might be associated with slightly lower breast density. However, the p-value of 0.297 suggests that this correlation is not statistically significant.

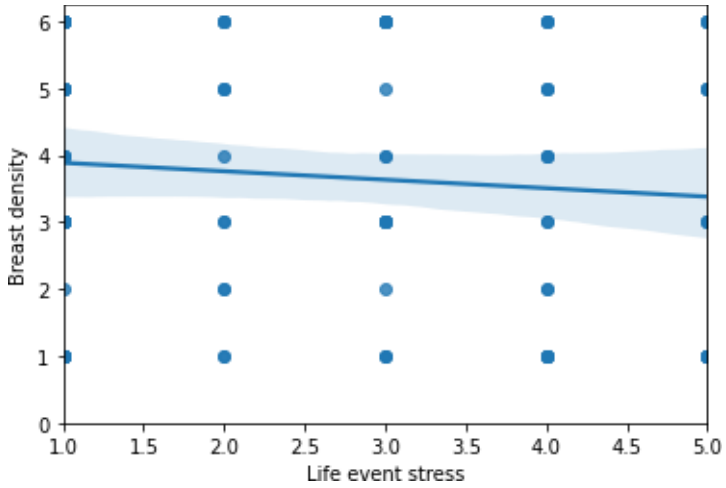

Figure 8. Correlation of Life Event Stress and Breast Density

3. **Personal Breast Cancer History and Breast Density**: The negative correlation coefficient of -0.250 suggests a moderately negative correlation between personal breast cancer history and breast density. This implies that individuals with a personal history of breast cancer may have lower breast density. The low p-value of 0.008 indicates that this correlation is statistically significant.

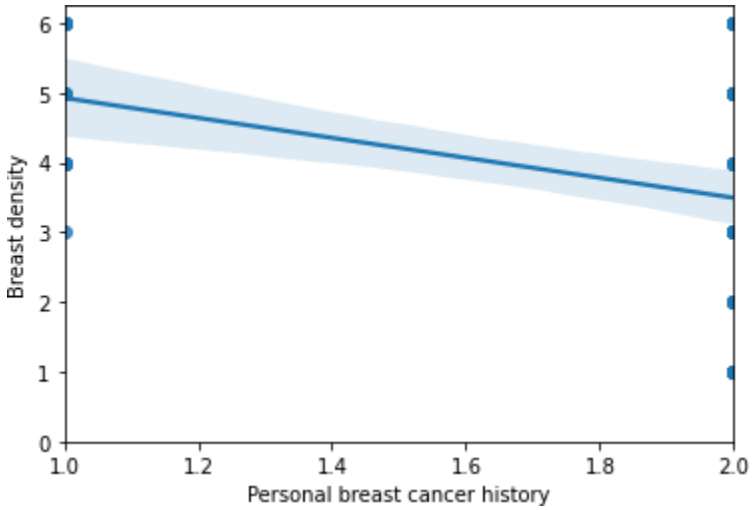

Figure 9. Correlation of Personal Breast Cancer History and Breast Density

4. **Personal Other Cancer History and Breast Density**: The negative correlation coefficient of -0.181 suggests a weak negative correlation between personal history of other cancers and breast density. The p-value of 0.057 indicates that while the correlation is not strongly significant, it approaches significance.

5. **Family History of Breast Cancer and Breast Density**: The negative correlation coefficient of -0.219 indicates a moderate negative correlation between family history of breast cancer and breast density. This suggests that individuals with a family history of breast cancer may have lower breast density. The p-value of 0.021 suggests statistical significance.

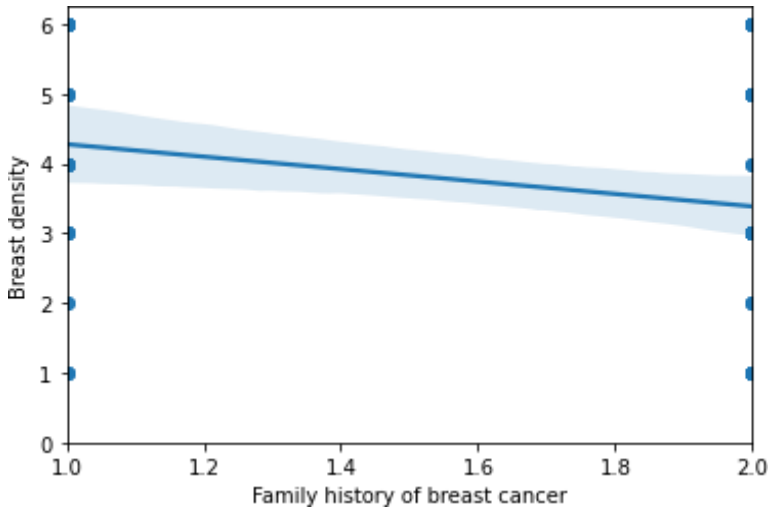

Figure 10. Correlation of Family History of Breast Cancer and Breast Density

6. **Family History of Other Cancer and Breast Density**: The positive correlation coefficient of 0.290 suggests a moderate positive correlation between family history of other cancers and breast density. This implies that individuals with a family history of other cancers might have slightly higher breast density. The low p-value of 0.002 indicates statistical significance

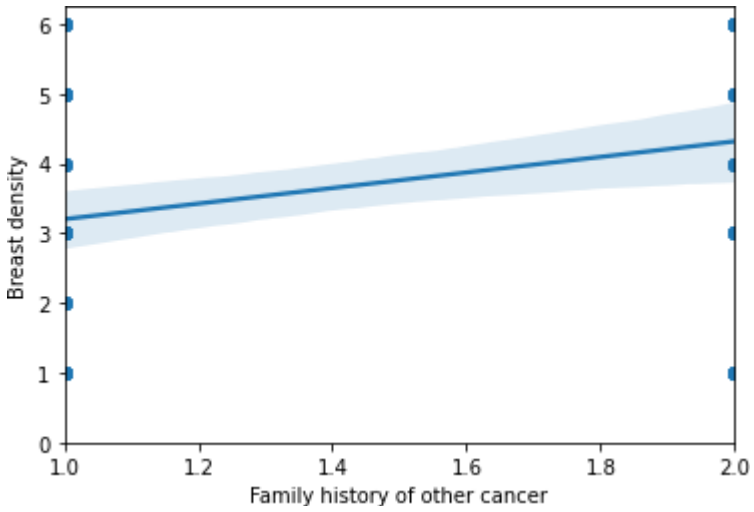

Figure 11. Correlation of Family History of Other Cancer and Breast Density

7. The **Pearson Correlation Coefficient of Marital Status and Breast density** is 0.057 with a p-value = 0.553. This suggests a very weak positive correlation between marital status and breast density. The p-value indicates that this correlation is not statistically significant. Marital status may not have a substantial impact on breast density, and this correlation might be due to random chance.

8. The **Pearson Correlation Coefficient of Smoker and Breast dens**ity is 0.086 with a p-value = 0.336. This indicates a weak positive correlation between smoking habits and breast density. However, the p-value suggests that this correlation is not statistically significant. While smoking is a known risk factor for various cancers, including breast cancer, the correlation with breast density might not be strong enough to draw definitive conclusions.

9. The **Pearson Correlation Coefficient of Biopsy and Breast density** is 0.070 with a P-value of P = 0.461. This implies a very weak positive correlation between a history of biopsy and breast density. The p-value again suggests that this correlation is not statistically significant. Biopsies are often performed to assess suspicious breast lesions, and their correlation with breast density might be influenced by a range of other factors.

Linear Regression

This study aims to investigate the relationship between various predictor variables (Age at the latest mammographic test, life event stress, personal breast cancer history, personal other cancer history, family history of breast cancer, family history of other cancer, smoker, marital status, and biopsy) and the dependent variable (breast density).

The linear regression analysis aimed to explore the relationship between breast density (dependent variable) and a set of predictor variables, including age at the latest mammographic test, life event stress, personal breast cancer history, personal other cancer history, family history of breast cancer, family history of other cancer, smoker, marital status, and biopsy. Various statistical measures were calculated to assess the fit and significance of the model.

The R = 0.284 in the table below is the correlation coefficient (Pearson's correlation) between the dependent variable (breast density) and the combination of all predictor variables. It indicates a weak positive linear relationship between the variables. It indicates a weak positive linear relationship between the variables. On the other hand, R Square = 0.081 is the coefficient of determination, indicating that approximately 8.1% of the variance in breast density can be explained by the predictor variables collectively. The Adjusted R-squared =.009 The adjusted R-squared accounts for the number of predictor variables and adjusts the R-squared value accordingly. In this case, the adjusted R-squared is very close to the original R-squared, suggesting that the inclusion of predictor variables has not substantially improved the model fit.

MODEL SUMMARY

Model	R	R Square	Adjusted R square	Std. Error of the Estimate
1	.284a	.081	.009	1.793

a. Predictors: (Constant), Life event stress, Personal breast cancer history, Smoker, Marital status, Age at latest mammographic test, Family history other cancer, Family history of breast cancer, Personal other cancer history

ANOVAα

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29.001	8	3.625	1.128	.351b
	Residual	330.990	103	3.213		
	TOTAL	359.991	111			

a. Dependent Variable: Breast Density
b. Predictors: (Constant), Life event stress, Personal breast cancer history, Smoker, Marital status, Age at latest mammographic test, Family history other cancer, Family history of breast cancer, Personal other cancer history

Table 3. Analysis of Variance Between Various Predictor Variables

The Standard Error of the Estimate = 1.793: is a measure of the accuracy of predictions made by the regression model. It represents the average distance between the actual values and the predicted values. It represents the average distance between the actual values and the predicted values. The Sum of Squares (SS) = 29.001, df = 8, Mean Square (MS) = 3.625, F = 1.128, Sig = 0.351. These values are related to the ANOVA (analysis of variance) for the regression model. The F-statistic tests the overall significance of the model, comparing the variation explained by the model to the variation not explained. The associated p-value (Sig) tests whether the model as a whole is statistically significant.

The analysis results suggest that the overall model, which includes all the predictor variables, does not have a statistically significant effect in explaining the variance in breast density. This conclusion is based on the F-statistic (F = 1.128) and the associated p-value (Sig = 0.351), which exceeds the typical threshold of significance (e.g., $\alpha = 0.05$).

Furthermore, the low values of R-squared (0.081) and Adjusted R-squared (0.009) suggest that only a small portion of the variability in breast density is explained by the predictor variables included in the model.

Based on the provided linear regression analysis, it can be concluded that the model's ability to predict breast density using the specified predictor variables is weak and statistically non-significant. Further refinement of the model, inclusion of additional predictors, and careful consideration of data quality and feature selection may help improve the model's performance and provide more meaningful insights into the relationship between breast density and the chosen variables.

Machine Learning Models

Breast cancer represents a substantial healthcare challenge, where timely detection plays a pivotal role in enhancing patient prognoses. Leveraging the potential of machine learning, this research also aims to assess and compare the capabilities of four distinct algorithms - k-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression - in the domain of breast cancer detection. The table below presents the evaluation metrics of the Jaccard Index. It measures the proportion of correctly predicted positive instances (true positives) relative to the total number of actual positive instances and the number of false negatives. Simply,

it measures, the overlap between predicted and actual sets. A high Jaccard Index (close to 1) indicates that the model's predictions align very well with the true class labels. A low Jaccard Index (closer to 0) suggests that the model's predictions are not well-aligned with the true outcomes. The F1-Score is a popular metric for assessing the performance of a classification model, especially when there is an imbalance between the classes. It is the score of both precision and recall combined into a single score. A high F1-Score (close to 1) indicates that the model is making accurate positive predictions (high precision) while also capturing a large proportion of actual positive instances (high recall). A low F1-score suggest that the model is either producing low precision or low recall. The Logarithmic Loss, often referred to as Log Loss, is a measure used to evaluate how well a classification model predicts the probability of different classes. It tells how closely the predicted probabilities match the actual results. A low Log Loss suggests that the model is making accurate and confident predictions, an indication that the model is performing well and has a good understanding of the data. A High Log Loss values can be a sign that the model needs improvement, adjustments, or further tuning.

	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.3	0.43	NA
1	Decision Tree	0.19	0.3	NA
2	SVM	0.26	0.38	NA
3	LogisticRegression	0.26	0.38	1.08

Table 4. Report Summary of 4 Machine Learning Models

1. K-Nearest Neighbors (KNN):

KNN is a non-parametric classification algorithm that makes predictions based on the majority class of its k-nearest neighbors. In this study, KNN achieved an accuracy rate of 0.8696 with k=1, indicating that 86.96% of the breast cancer cases were correctly classified. The Jaccard index of 0.3 suggests a moderate level of similarity between predicted and actual instances, while the F1-score of 0.43 signifies a reasonable balance between precision and recall.

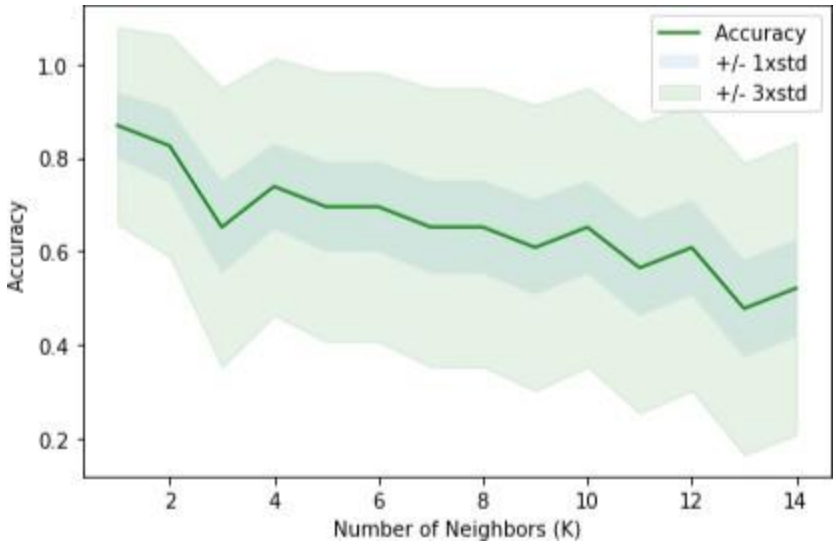

Figure 12. K-Nearest Neighbor (KNN) model

2. Decision Tree:

The Decision Tree algorithm constructs a tree-like model for classification by recursively partitioning the data based on attribute values. The obtained accuracy rate of 0.7059 indicates that 70.59% of the breast cancer cases were accurately classified. The Jaccard index of 0.19 suggests a relatively lower level of similarity compared to KNN, while the F1-score of 0.3 indicates a trade-off between precision and recall, similar to KNN.

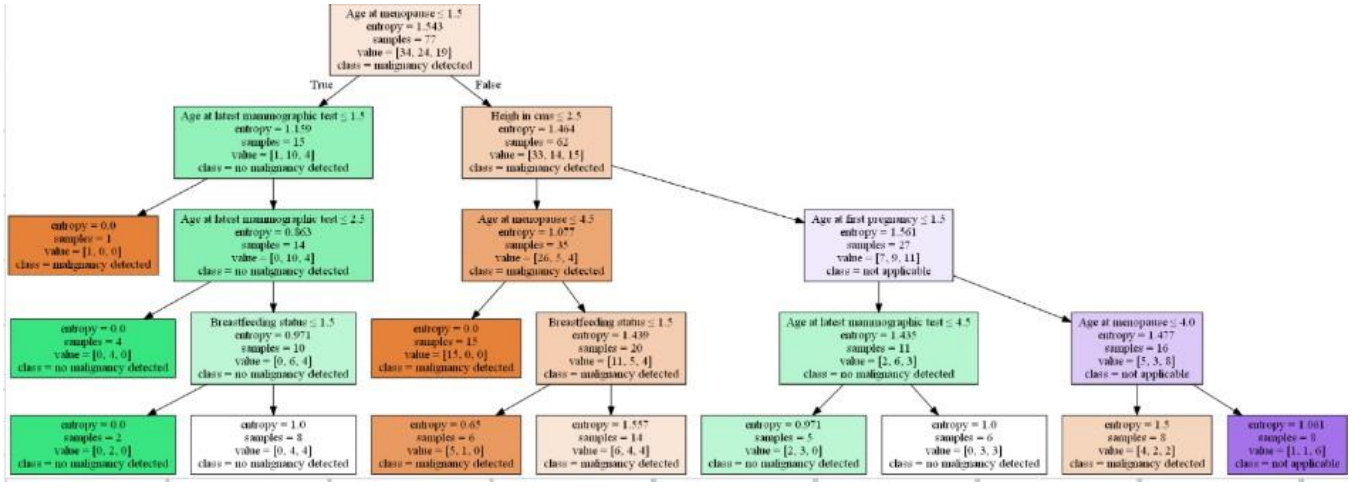

Figure 13. Decision Tree Model

3. Support Vector Machine (SVM):

SVM is a powerful algorithm that aims to find the optimal hyperplane for separating classes in a high-dimensional space. The Jaccard index of 0.26 indicates a moderate level of similarity in predictions, while the F1-score of 0.38 implies a balanced trade-off between precision and recall.

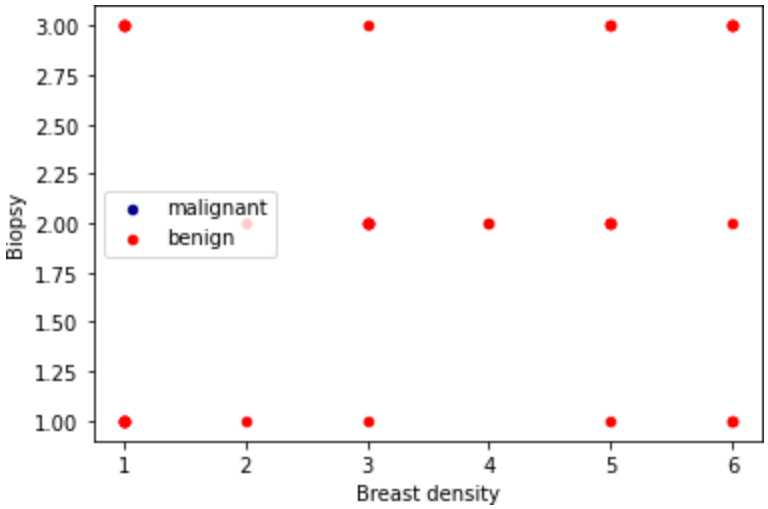

Figure 14. Support Vector Machine model

4. Logistic Regression:

Logistic Regression is a linear classification algorithm that models the probability of the binary outcome. The Jaccard index of 0.26 suggests a moderate similarity between predicted and actual instances, and the F1-score of 0.38 indicates a balanced precision-recall trade-off. The Logloss value of 1.08 quantifies the cross-entropy loss between predicted probabilities and actual class labels, with lower values indicating better model calibration.

Confusion Matrix

A confusion matrix is typically a square matrix that compares the predicted classifications made by a model with the actual classes of the data. It is divided into four regions, based on the two classes being predicted and the two classes present in the actual data.

The four regions are as follows:

1. **True Positive** (TP): This represents the number of instances where the model correctly predicted a positive class (Class A) when the actual class was indeed positive.

2. **True Negative** (TN): This represents the number of instances where the model correctly predicted a negative class (Class B) when the actual class was indeed negative.

3. **False Positive** (FP): Also known as a Type I error, this is the number of instances where the model incorrectly predicted a positive class when the actual class was negative. In other words, it's a "false alarm."

4. **False Negative** (FN): Also known as a Type II error, this is the number of instances where the model incorrectly predicted a negative class when the actual class was positive. In other words, it's a failure to detect the positive class.

It is an important tool for understanding, analyzing, and improving the performance of machine learning models, to make informed decisions and optimizations based on real-world performance metrics rather than relying solely on accuracy.

For this study, the model predicted the positive class 7 times, and all of these predictions were correct. It did not predict the negative class) at all. Additionally, there were no instances where the model made incorrect predictions for either class.

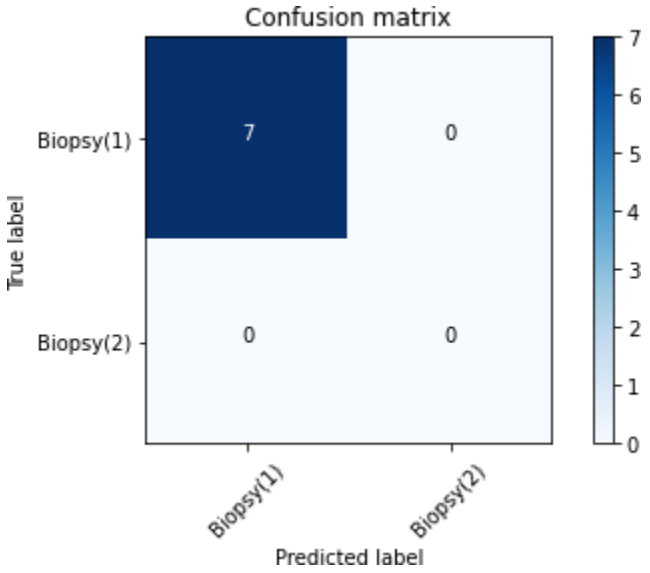

Figure 15. Confusion Matrix

Overall, this research demonstrated the application of machine learning models for breast cancer detection and highlights the importance of performance evaluation metrics. The KNN model exhibits the highest accuracy, Jaccard index, and F1-score among the tested models. However, the Decision Tree model shows a lower performance in comparison. Both SVM and Logistic Regression models provide competitive results in terms of Jaccard index and F1-score. The Logistic Regression model shows a relatively higher log loss, suggesting potential room for improvement.

The confusion matrix suggests that the model performed perfectly for the positive class but did not make any predictions for the negative class While a perfect score for the positive class is desirable, it's important to consider in probability estimation. The overall context and other metrics when evaluating a model's performance, as this confusion matrix might indicate some limitations or issues with the model.

CONCLUSION

Breast cancer is a serious medical problem that puts enormous strain on the world's healthcare systems. Early breast cancer detection enables more effective and less aggressive treatment options, which improve patient outcomes and prognoses. Early detection can increase the likelihood of successful treatments while improving the quality of life for those who are affected.

Machine learning techniques are becoming more popular in the healthcare industry. Large datasets are used to train machine learning algorithms to find patterns and make predictions. Researchers are investigating how various machine learning algorithms can help healthcare professionals identify breast cancer more precisely and quickly in the setting of breast cancer detection.

This research aims to evaluate and present the performance of four specific machine learning algorithms: k-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. Each of these algorithms has a unique way of analyzing data and making predictions. The effectiveness of the algorithms is measured by metrics including accuracy, precision, recall, Jaccard score, and F1 score. In order to help medical practitioners make better decisions and maybe improve patient outcomes through early and accurate diagnosis, it is important to identify which algorithm(s) provide the highest accuracy and reliability in identifying breast cancer and apply these ML techniques to improve early diagnosis, ultimately leading to better patient outcomes.

Additionally, this study aimed to understand the effectiveness of these models in identifying breast cancer cases and their potential to assist medical professionals in making accurate diagnoses. Presented are the following:

1. The provided data offers insights into various aspects related to age of mammographic testing, life event stress, personal and family cancer history, marital status, smoking, breast density, and biopsy outcomes. Understanding these factors is essential for improving breast cancer awareness, early detection, and overall healthcare strategies. It's crucial for healthcare providers and researchers to consider these findings to tailor interventions and

support the unique needs of individuals based on their demographic and health-related characteristics.

2. While some correlations were statistically significant, others were not. These findings suggest that these factors may play a role in influencing breast density and, potentially, breast cancer risk. However, the relationships are generally weak to moderate. Further research is needed to better understand the underlying mechanisms and potential clinical implications of these correlations.

3. The regression model as a whole does not seem to be a strong fit for explaining the variability in breast density. The low R-square and adjusted R-square suggest that the predictors in the model explain only a small portion of the variability in breast density. Additionally, the non-significant F-statistic and high p-value (Sig) indicate that the model's fit is likely due to chance. This suggests that the current set of predictors might not be adequate for accurately predicting breast density in this context. Further analysis and potentially different predictors may be needed to build a more robust model.

4. The research explores the performance of four machine learning models in breast cancer detection. The KNN model achieved the highest accuracy rate and F1 score, indicating its effectiveness in correctly classifying breast cancer cases while maintaining a balanced precision-recall trade-off. The Decision Tree, SVM, and Logistic Regression models showed lower performance metrics, with varying degrees of precision and recall.

5. Pilot studies are often exploratory in nature. They serve as initial investigations to explore new research questions, hypotheses, or areas of interest. Despite the small sample size, these studies help researchers gain a foundational understanding of breast cancer detection using ML models before conducting more extensive research. The research nonetheless has relevance in offering preliminary insights despite being confined by a small sample size. It serves as a valuable starting point for research endeavors in breast cancer detection using ML approaches that can be done nationwide. It offers preliminary insights, generates hypotheses, identifies challenges, and lays the groundwork for more extensive investigations.

RECOMMENDATIONS

While the provided data offers valuable insights into several important factors related to breast cancer, there are additional variables that could significantly enhance the quality and applicability of this study. Expanding the range of factors such as genetic (genetic information) and hormonal (menstrual history, hormonal replacement therapy, etc.), medical history, socioeconomic factors, lifestyle, environmental exposure, and other psychosocial factors like social support, psychological well-being, and coping mechanisms is considered essential to creating a more comprehensive and accurate understanding of breast cancer risk and detection.

1. The small sample size has an effect on the ML models being used. It's important, therefore, to recognize its limitations and take precautions to avoid bias and inaccuracy. Applying machine learning (ML) models like K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, and Support Vector Machine (SVM) can have distinct consequences and implications due to the small sample size. The efficiency of KNN may have diminished with a smaller dataset. Predictions may be unstable as a result of a tiny dataset's lack of diversity. Despite their propensity to perform well on short datasets, Decision Trees may fit the training data too closely and not generalize well to new, unseen data. Logistic Regression is relatively robust for small datasets, but its ability to capture complex relationships might be limited when dealing with a few instances. SVM can handle small datasets effectively, but a small dataset might make it challenging to find optimal hyperparameters that lead to its best performance.

2. The aforementioned facts suggest that this kind of research should utilize large datasets. As a result, this will provide ML models with additional data, allowing them to develop more accurate representations, more effective generalizations, and more accurate predictions. The advantages for model performance and dependability outweigh any computational difficulties and resource requirements associated with processing a larger dataset.

3. Utilize additional ML models, like Random Forest or Gradient Boosting, or even advanced machine learning techniques, such as deep learning. Combining the benefits of these several models improves the handling of complicated patterns, reduces overfitting, and yields reliable predictions.

4. Conducting a breast cancer detection study on a nationwide scale offers a wealth of benefits, including enhanced representation. By including data from different regions and healthcare facilities across the country, the study's findings can be generalized to a broader population. This ensures that the developed models are more likely to perform well across various generalizations. Regional insights on breast cancer prevalence, risk factors, and healthcare infrastructure can vary between regions. A nationwide study allows for the identification and understanding of these regional variations, potentially leading to the development of tailored models that account for unique characteristics in different areas. Resource utilization means collaborating with medical institutions, researchers, and experts from multiple regions to leverage the expertise and resources available across the country. This collective effort can lead to more comprehensive data collection, thorough analysis, and meaningful insights. Nationwide studies carry greater potential to influence healthcare policies, clinical guidelines, and practices at a national level or simply have a policy impact. Lastly, conducting a nationwide study can set the stage for international comparisons and collaborations, enabling researchers to benchmark their findings against other countries and contribute to a broader understanding of breast cancer detection methods globally.

REFERENCES

1. Arya, N. (2022). Nearest neighbors for classification. *KDnuggets*. https://www.kdnuggets.com/2022/04/nearest-neighbors-classification.html

2. Chauhan, N.S. (2022). Decision tree algorithm explained. *KDnuggets*. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

3. Coursera. (2023). What is machine learning in health care?: Applications and opportunities. https://www.coursera.org/articles/machine-learning-in-health-care

4. El Masari, H., Gherabi, N., Mhammedi, S., Sabouri, Z., Ghandi, H. & Qanouni, F. (2023). Effectiveness of applying machine learning techniques and ontologies in breast cancer detection. *Procedia Computer Science, 218*, 2392-2400.

5. Elsadig, M.A., Altigan,A. & Elshoush, H.T. (2023). Breast cancer detection using machine learning approaches: A comparative study. *International Journal of Electronics and Communications, 13*, 736-745.

6. Fanizzi, A., Basile T.M.A., Losurdo, L., Belloti, R., Bottigli, U., Dentamaro, R., Didonna, V., Fausto, A., Massafra, R., Moschetta, M., Popescu, O., Tamborra, P., Tangaro, S. & La Forgia, D. (2020). A machine learning approach on multiscale texture analysis for breast Microcalcification diagnosis. *BMC Bioinformatics, 21* (1), 98. https://doi.org/10.1186/s12859-020-3358-4

7. Foresee Medical. (2022). Benefits of machine learning in healthcare. https://www.foreseemed.com/blog/machine-learning-in-healthcare

8. Jessica, S. (2022). How does logistic regression work?. *KDnuggets*. https://www.kdnuggets.com/2022/07/logistic-regression-work.html

9. Kasuso – Philippine Foundation for Breast Cancer Care, Inc. (n.d.). About breast cancer. https://www.kasuso.org/about-breast-cancer

10. Lim, Y.X., Lim, Z.L., Ho, P.J. & Li, J. (n.d.). Breast cancer in Asia: Incidence, mortality, early detection, mammography programs, and risk-based screening initiatives.

11. Mali, K. (2021). Data Science Blogathon

12. Mendoza, J.E. (2023, February 22). Breast cancer cases in PH 'staggering' – health expert. *Philippine Daily Inquirer*.

13. Nag, S. & Nag, J. (2021). A comparative analysis of machine learning approaches for prediction of breast cancer. *Journal of Emerging Investigators, 3*, 1-9.

14. Pazzibugan, D.Z. (2023, July 10). Only 1% of PH women screened for breast, cervical cancer. *Philippine Daily Inquirer*.

15. Reza Rabiel, S.M.A., Sohrabel, S., Esmaeli M., Atashi, A. (2022). Prediction of breast cancer using machine learning approaches. *Journal of Biomedical Physics & Engineering, 12* (3).

16. Salod, Z. & Singh, Y. (2019). Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. *Journal of Public Health Research, 8* (4), 112-118.

17. Sandeep, D. & Beena Bethel, G.N. (2021). A Survey on accurate breast cancer detection and classification using machine learning approach. *E3S Web of Conference, 309*, 01116. https://doi.org/10.1051/e3sconf/202130901116

18. Tapas, A. (2023). Only 1% of Filipino women screened for breast, cervical cancer study. *Manila Times*. https://www.manilatimes.net/2023/07/10/news/only-1-of-filipino-women-screened-for-breast-cervical-cancer-study/1899964

19. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. https://doi.org/10.3390/designs2020013

APPENDIX

Online Survey Form

BREAST CANCER DETECTION IN THE PHILIPPINES USING MACHINE LEARNING APPROACHES: A PILOT STUDY

The purpose of this study is to detect breast cancer using various machine learning techniques. This could help the patients and those in the medical sector determine whether the patients are vulnerable to breast cancer at an early stage. The questions are based from Reza Rabiel et al.'s paper, "Prediction of Breast Cancer Using Machine Learning Approaches," which was published in 2022.

Thank you for taking part in this research. Rest assured that your responses will be treated anonymously.

Sincerely,

Maria Maura S. Tinao
University of the Philippines Open University
mstinao@up.edu.ph
Ruth B. Rodriguez
University of the Philippines Open University

Eunelfa Regie Calibara
University of the Philippines Open University
Email address:

DATA PRIVACY CONSENT

☐ I hereby give the researchers permission to collect, record, organize, update or alter, retrieve, consult, utilize, consolidate, erase, or destroy my personal data as part of my information. I hereby affirm my right to be informed, to object to processing, to access and rectify my personal data, to suspend or withdraw it, and to be indemnified in the case of a breach of the provisions of the Republic Act No. 10173 of the Philippines, Data Privacy Act of 2012, and its corresponding Implementing Rules and Regulations.

INSTRUCTIONS: Select the answer that corresponds to your criteria.

1. Age at latest mammographic test.

- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55-64
- ◯ 65-74
- ◯ 75 and above

2. Marital Status

- ◯ Single
- ◯ Married
- ◯ Separated
- ◯ Widowed

3. What is the stressful incident in your life?

- ◯ Death of loved one/s
- ◯ Separation
- ◯ Family problem/s
- ◯ Work related problem/s
- ◯ None
- ◯ Others

4. Do you have personal breast cancer history?

- ◯ Yes
- ◯ No

5. What is your personal other cancer history?
   - ◯ Ovary
   - ◯ Endometrium
   - ◯ Colon
   - ◯ Meningioma
   - ◯ Lymphoma
   - ◯ None
   - ◯ Others

6. Do you have a family history of breast cancer?
   - ◯ Yes
   - ◯ No

7. Do you have a family history of other cancer?
   - ◯ Yes
   - ◯ No

8. Are you a smoker?
   - ◯ Yes
   - ◯ No
   - ◯ Not sure

9. Breast density
   - ◯ Fatty tissue
   - ◯ Glandular and fibrous tissue
   - ◯ Dense
   - ◯ Heterogeneously dense
   - ◯ Extremely dense
   - ◯ No information, need to consult

10. Biopsy

- ◯ No malignancy detected
- ◯ Malignancy detected
- ◯ Not applicable