Analysis of the fundamental factors leading to breast cancer using data mining techniques

Harini Shree Bhaskaran (B00928615)
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
hr478842@dal.ca

Ayushi Sharma (B00940205)
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
ay847821@dal.ca

Abstract

One of the most predominant and critical disease in the recent trend is evidently the breast cancer. Although experienced radiologists can perform successful detection of malignant tumors yet these is glitch in the diagnosis of type of cancer. The primary reason behind this trend topic because of its mortality rate. Researches around the globe have proposed various methodologies for detection of this disease, however it still requires improvisation for accurate and efficient detection. Data mining algorithm can provide great assistance in determination and treatment of BC at early stage. The purpose of this research is to determine how precise these data mining techniques forecast the likelihood of recurrence of BC in patients with the defined parameters. The large volumes of data play a fundamental role according to the literature consulted, a great variety of data set oriented to the analysis of the disease has been generated, in this research the Wisconsin Breast Cancer Database (WBCD) was used, the purpose of the proposed research is to submit comparison of the performance of various clustering and classification algorithms and conclude the most accurate.This study will help future researchers in breast cancer field to continue their research and focus on improving the performance of specific algorithms.

keywords: breast cancer, prediction, data mining, recurrence, accuracy

1 INTRODUCTION

Breast Cancer (BC) is the world's most prevalent cancer after skin cancer in women. According to World Health Organization, 2.3 million women were diagnosed with breast cancer in 2020 out of which 685000 were reported dead by the end of the year [1]. Breast cancer is considered to be fatal for half of the women affected by it. The fifth major motivator for women death is BC in comparison to all forms of cancer. Breast cancer originates in the breast tissue. It occurs when breast cells mutate (change) and grow out of control, creating a mass of tissue (tumor). Like other cancers, breast cancer can invade and grow into the tissue surrounding the breast. It can also travel to other parts of your body and form new tumors. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed..According to the most recent data, the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis. Early detection of this cancer improves the chances of survival for patients suffering from it. Many biological techniques can be used to detect breast cancer early and take preventive measures. Cancer cells in a benign tumor do not have the tendency to grow beyond or spread to other parts of the body, whereas cancer cells in a malignant tumor have the tendency to grow outside of the breast tissue and spread to other parts of the body. Therefore, malignant tumors

are very dangerous and should be predicted at early stages so that it can be cured at low cost and less pain with less damage being caused to the surrounding area. Early detection necessitates an accurate and dependable diagnosis procedure that allows doctors to differentiate benign from malignant breast tumours without requiring a surgical biopsy. Data mining methods can help to successfully detect breast cancer recurrence.In this research, we focus on analyzing the various factors contributing to the cancer and conclude the evident reason and predicting in advance the possibility of recurrence of breast cancer.

2 BREAST CANCER

The second most frequent malignancy in women after skin cancer is breast cancer. Over-50-year-old women are the ones most likely to be affected. Although uncommon, breast cancer can also strike men. Male breast cancer affects about 2,600 males annually in the US, accounting for fewer than 1% of total cases.

2.1 Types of Breast Cancer

There are several different types of breast cancer, including:

2.1.1 Infiltrating (invasive) ductal carcinoma

Starting in your milk ducts of your breast, this cancer breaks through the wall of your duct and spreads to surrounding breast tissue. Making up about 80% of all cases, this is the most common type of breast cancer.

2.1.2 Ductal carcinoma in situ

Also called Stage 0 breast cancer, ductal carcinoma in situ is considered by some to be precancerous because the cells haven't spread beyond your milk ducts. This condition is very treatable. However, prompt care is necessary to prevent the cancer from becoming invasive and spreading to other tissues.

2.1.3 Infiltrating (invasive) lobular carcinoma

This cancer forms in the lobules of your breast (where breast milk production takes place) and has spread to surrounding breast tissue. It accounts for 10% to 15% of breast cancers.

2.1.4 Lobular carcinoma in situ

It is a precancerous condition in which there are abnormal cells in the lobules of your breast. It isn't a true cancer, but this marker can indicate the potential for breast cancer later on. So, it's important for women with lobular carcinoma in situ to have regular clinical breast exams and mammograms.

2.1.5 Triple negative breast cancer

Making up about 15% of all cases, triple negative breast cancer is one of the most challenging breast cancers to treat. It's called triple negative because it doesn't have three of the markers associated with other types of breast cancer. This makes prognosis and treatment difficult.

2.1.6 Inflammatory breast cancer

Rare and aggressive, this type of cancer resembles an infection. People with inflammatory breast cancer usually notice redness, swelling, pitting and dimpling of their breast skin. It's caused by obstructive cancer cells in their skin's lymph vessels.

2.1.7 Paget's disease of the breast

This cancer affects the skin of your nipple and areola (the skin around your nipple).

2.2 Signs and Stages of Breast Cancer

Breast cancer symptoms can vary for each person. Possible signs of breast cancer include:

- A change in the size, shape or contour of your breast.
- A mass or lump, which may feel as small as a pea.
- A lump or thickening in or near your breast or in your underarm that persists through your menstrual cycle.
- A change in the look or feel of your skin on your breast or nipple (dimpled, puckered, scaly or inflamed).
- Redness of your skin on your breast or nipple.
- An area that's distinctly different from any other area on either breast.
- A marble-like hardened area under your skin.
- A blood-stained or clear fluid discharge from your nipple.

The staging process explains the extent of the cancer in your body. The size, location, and extent of the tumour, as well as whether the disease has spread to other parts of your body, all play a role in this decision. The primary phases of breast cancer are:

Stage 0: The disease is non-invasive. This means it hasn't broken out of your breast ducts.

Stage I: The cancer cells have spread to the nearby breast tissue.

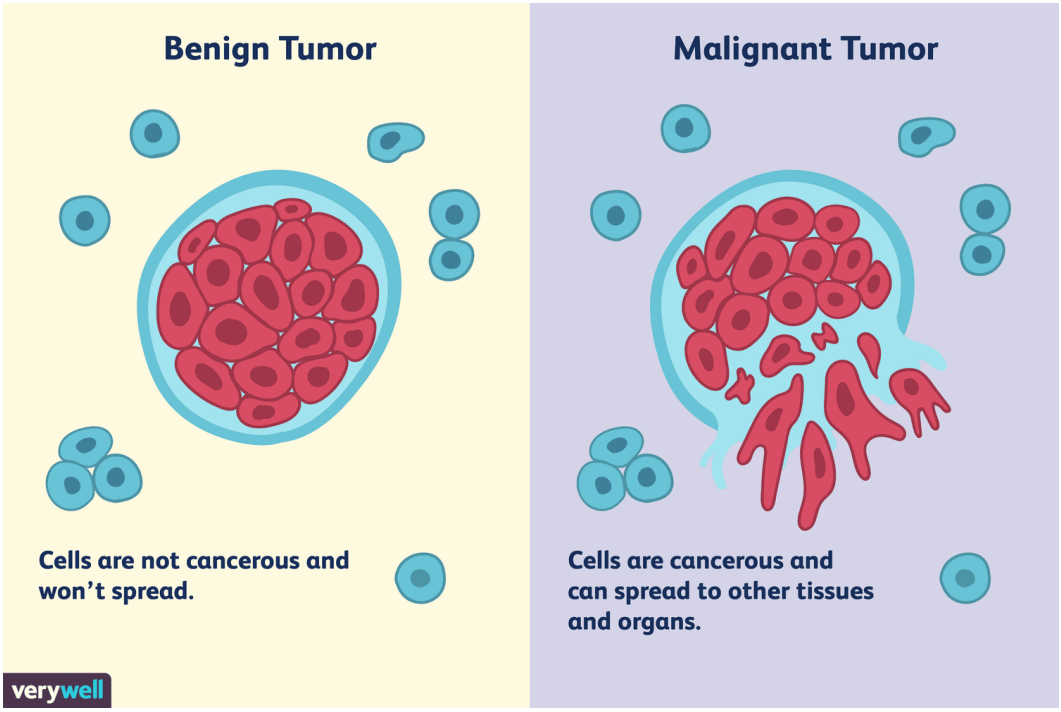
Stage II: The tumor is either smaller than 2 centimeters across and has spread to underarm lymph nodes or larger than 5 centimeters across but hasn't spread to underarm lymph nodes. Tumors at this stage can measure anywhere between 2 to 5 centimeters across, and may or may not affect the nearby lymph nodes.

Stage III: At this stage, the cancer has spread beyond the point of origin. It may have invaded nearby tissue and lymph nodes, but it hasn't spread to distant organs. Stage III is usually referred to as locally advanced breast cancer.

Stage IV: The cancer has spread to areas away from your breast, such as your bones, liver, lungs or brain. Stage IV breast cancer is also called metastatic breast cancer. To prevent or lower the chances of breast cancer we can first look into the factors that are associated with it and which contribute to the development of breast cancer. Firstly, being a woman in itself is the biggest risk factor of developing breast cancer. Although men also have the risk of developing breast cancer, it is comparatively less than a woman since only less than 1% of the breast cancer cases belong to male patients. Secondly, age is a huge contributor to increased risk of breast cancer. According to CDC, most of the patients with breast cancers lie in the age group of 50 and above. Thirdly, if there has been a family history of breast cancer and a female relative in the family has had an experience with it, the chances of developing a breast cancer get increased significantly. Genetics can also be a leading factor in the development of breast cancer. If a defective gene has been passed down from a parent or a gene mutation has occurred in the body, then that person will have a higher chance of developing breast cancer. There are many other factors as well that lead to increased risk of developing breast cancer such as one's own personal history of breast cancer, exposure to radiation at a young age to treat another disease, development of benign tumors in breast tissues, one's race or ethnicity, being overweight, history of breastfeeding, history of pregnancy that is having a child after the age of 30, menstrual history that is getting periods before the age of 12 and experiencing menopause after the age of 55, getting hormone replacement therapy, drinking alcohol, smoking, lack of exercise, having Vitamin D deficiency, light exposure at night, not getting enough or proper nutrients and eating unhealthy food, exposure to toxic chemicals or chemicals found in food, lawns, gardens, plastic such as Bisphenol A(BPA), sunscreens, grilled and pre-prepared foods and water.

2.3 Types of Tumour

Figure I: Tumour Cells[18]



Benign tumors are non-cancerous and generally less harmful than malignant tumors. They can grow to large sizes – potentially causing pain and other problems by putting pressure on the area around the tumor – but do not invade other tissues or organs.

Malignant tumors are cancerous, with the capacity to spread and establish new tumors in other tissues and organs in a process called metastasis. During metastasis, primary malignant tumors spread to secondary sites. Cancer cells break away from the tumor and spread via the blood or lymph system – commonly establishing metastases in the liver, lungs, brain and bones. Because of this, malignant tumors quickly require treatment to avoid spreading. Treatment usually entails surgical removal, chemotherapy, radiotherapy or a combination of these depending on how advanced the cancer is.

3 LITERATURE SURVEY

To identify breast cancer different machine learning methods have been proposed by various researchers. In this work we have been discussed some of the state of the art breast cancer diagnosis methods. The main purpose of literature review to identify the problems in existing methods and provide a reliable solution.Also we will be using WEKA tool to identify the best classifier.

Wang et. al,[3] in their research compared four data mining models support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree, to predict the effective way to predict breast cancer. It was concluded that principal component analysis (PCA), as a dimension reduction technique, manifests some advantages in terms of prediction accuracy and efficiency. However PCA is a linear feature reduction model which cannot be generalized as some non linear feature reduction techniques such as K-Means could also be used.

Shomona et al. [4] have compared the various classifier algorithms on the WBCD dataset. Their results show that the Random Tree and C4.5 classification algorithm produce 100% accuracy. It is observed that they have used'Time' attribute (Time to recur/Disease-free Survival) along with other parameters to predict the outcome of recurrence or non-recurrence of breast cancer among the patients. However the 'Time' attribute has not been relied upon for prediction of recurrence of the disease.

Lambrou et al presented a Conformal Predictor in light of Genetic Algorithms, and applied to (WBCD) problem [5]. A rule-based Genetic Algorithms (GAs) was utilized as a strategy for building a Conformal Prediction (CP). The subsequent algorithm was applied to the problem of breast cancer diagnosis for 683 records without missing values from WBCD dataset. The error rates affirmed the legitimacy of their CP for any given confidence level 1-e, where e is the error rate.

A recent study by Azzam et. al,[6] compares and evaluates the performance and accuracy of the key supervised and semi-supervised machine learning algorithms for breast cancer prediction. It was concluded that semi-supervised algorithms provided promising and competitive approach to solve a problem involving diagnosis of tumor type.

Subrata kumar Mandal et. al,[7] narrowed down the features which can accurately classify the breast cancer as malignant and benign. Comparative study of various approaches of classifiers i.e. Naive Bayes(NB), Logistic Regression (LR) , Decision Tree on the basis of time complexity and accuracy. Experiments were conducted by the author and found that the Logistic Regression (LR) has maximum accuracy with minimum time complexity.

In this study,[8] compared multi-layer perception(MLP), Decision tree(J48) and Sequential Minimal Optimization (SMO). Three datasets have been used i.e WBC, WDBC, WPBC on which 10-fold cross validation method is applied. Results shows that fusion of SMO, MLP, J48 is higher than other classifier in WPBC datast and the fusion of MLP and SMO is better in WDBC datase.

The authors Aruna et.al, compared Naive Bayes, Spport Vector Machine, Radial Basis Neural Network, J48 and simple CART using WEKA. The authors used multiclass and binary dataset and compared all these classifiers on the basis of precision, sensitivity and specificity. Results shows that SVM-RBF kernel has high performance percentage when compared to other classifiers.[9]

In the study proposed by Shilpa et. al [10], the breast cancer dataset is analyzed and factors determining cancer such as benign and malignant are conisedered and machine learning algorithms such as Naïve Bayes, J48, Sequential Minimal Optimization (SMO) and Instance- Based for K-Nearest neighbor (IBK) are used.It is concluded that the proposed IBK algorithm gives the maximum accuracy of 100% in contrast to the other three algorithms, which results in selecting the best classification algorithm for the breast cancer prediction and can be used for detection and treatment.

The study by Sruthi et.al, [11] compares the various machine learning algorithms across different types of cancers.The breast cancer model as a classification job as is the development of the Support Vector Machine (SVM) approach to classify breast cancer as benign or malignant. Random forest classifier was employed in the lung cancer and prostate cancer prediction systems. The main aim of this paper was to develop a classifier that could predict the likelihood of a person developing lung or prostate cancer based on a set of common factors.

The method proposed by [12] is to identify pattern recognition and prediction modelling using Logistic Regression, K-Neighbours Classifier, Support Vector Classifier (SVC) linear, Gaussian Naive Bayes, and Decision Tree Classifier in the breast cancer dataset. The model SVC linear can determine and diagnose whether or not a patient has cancer, with an output metric of around 97.19 percent.

Rovsheno et. al, [13] aimed to classify benign and malignant breast cancer image features. Artificial Neural Network, Support Vector Machine and Random Forest algorithms were used to classify features obtained from images. Experiments were performed on the Wisconsin Breast Cancer data set. Experimental evaluation shows that 99% of the most successful results were achieved with the Artificial Neural Network algorithm. According to experimental findings, the classification technique can identify breast cancer in its early stages.

Yifan et.al, [14] in their study integrated Random Forest and AdaBoost algorithms, for breast cancer classification prediction model that can give a diagnosis result of benign or malignant. It was compared the model with single Support Vector Machine, Logistic Regression, K-Nearest Neighbor, Decision Tree algorithms. The test results have shown that the ensemble model's prediction accuracy has been increased by 4.3% on average compared to the single algorithm models, with the highest increase up to 9.8%, which has provided a new reference model for breast cancer prediction.

A 2017 research by Vikas et al [15], helps estimate the survivability of breast cancer patients. To develop prediction models, multiple data mining algorithms were used such as Naïve Bayes, RBF Network, J48. Then they made use of the 10-fold cross-validation methods using a stratified sampling

technique to measure the estimate of the three prediction models to compare each of the algorithm's performance. To analyse and evaluate the performance of data mining techniques they made use of the Waikato Environment for Knowledge Analysis (WEKA) tool. Their research indicated that the Naive Bayes algorithm gave the best results out of the three prediction models with an accuracy of 97.36% followed by RBF Network which gave an accuracy of 96.77% and at last the J48 which gave the least accuracy out of the three algorithms which was of only 93.41%.

4 DATASET

4.1 Dataset Description

4.1.1 Primary Dataset Description and Exploration

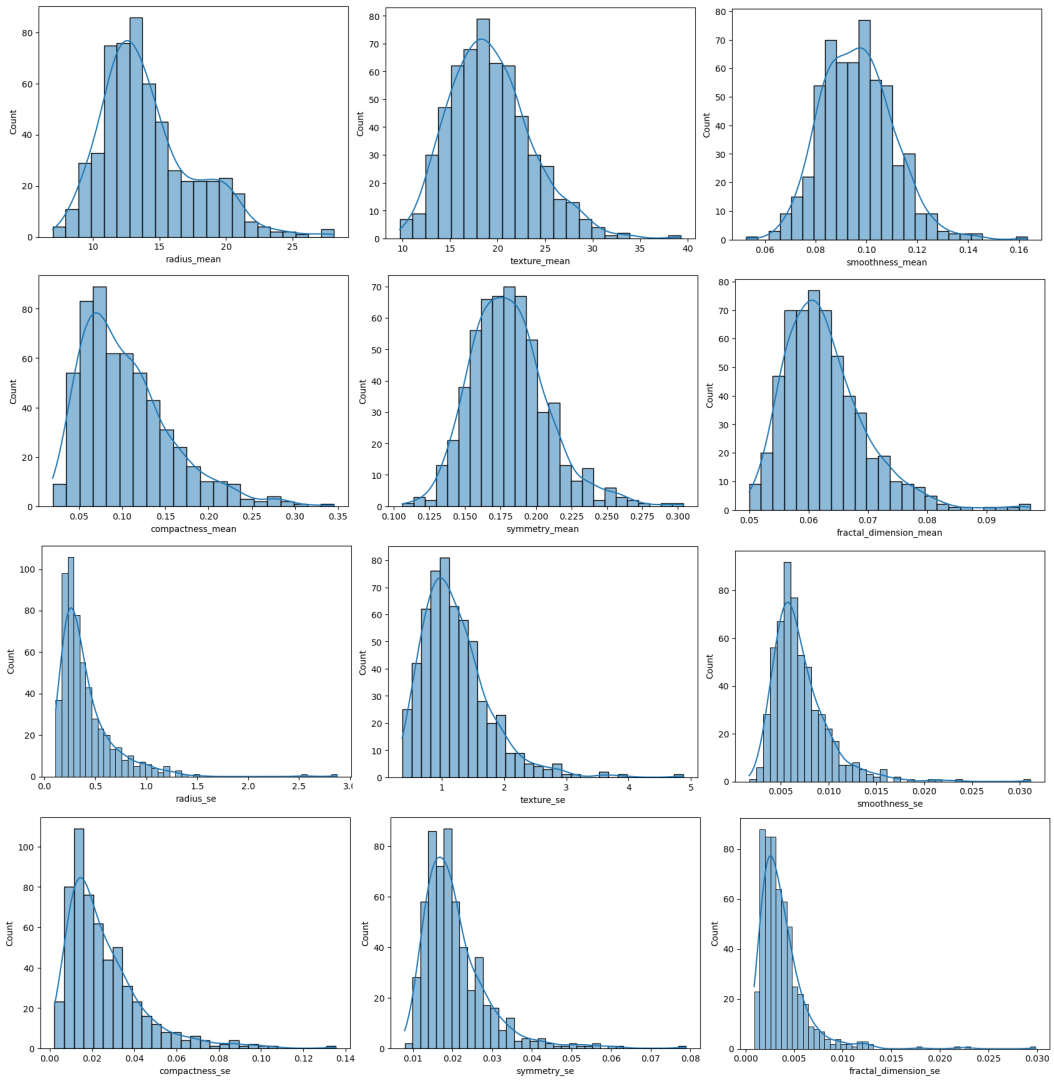
For our proposed research project, we will use the well-known Wisconsin Breast Cancer Database (WBCD) [2] provided at the University of California, Irvine's Machine Learning repository. It is a classification data set, which records the measurements for breast cancer cases. The larger the data set, more accurate are the results. The considered data set is around 50KB in size. The data set contains two primary attributes namely malignant and benign.The malignant class of this data set is down sampled to 21 points, which are considered as outliers, while points in the benign class are considered inliers. Our data set consists of a total of 569 samples. There are a total of 32 features that characterize our samples, the first of which is the ID of the sample, the second is its class, and the remaining 30 are features that contain various information about the cells.

Figure II: Data Visualization

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

The class label of our samples can be malignant (M) or benign (B). These are medical terms that refer to the benign and malignant tumor cells we talked about earlier. There are no missing values for the properties. Of our samples, 357 are benign and it is distributed to be 212 malignant. Also, it comprises of various features of the cell nucleus. These features are computed for each cell nucleus are radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness ($perimeter^2$ / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry and fractal dimension ("coastline approximation" - 1). We can see that the primary data set follows a normal distribution curve for most of its features.

Figure III: Data Exploration



4.1.2 Secondary Dataset Description

This dataset of breast cancer patients is obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics. This dataset involves female patients with infiltrating duct and lobular carcinoma breast cancer diagnosed in 2006-2010. A total of 4024 patients are included. The dataset is around 44 KB in size. There are a total of 16 features including their age, race, marital status, t-stage, n-stage, 6th stage, differentiate, grade, a-stage, tumor size, estrogen status, progesterone status, regional node examined, regional node positive, survival months, patient's living status.

4.2 Data Preprocessing

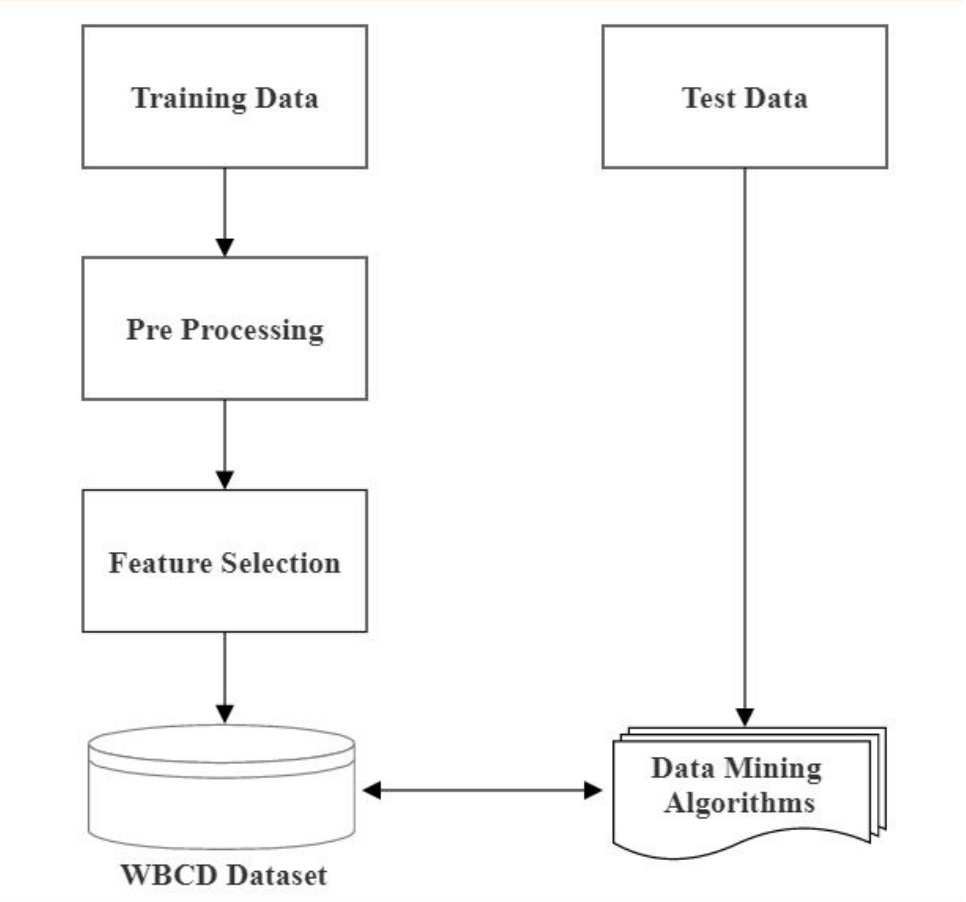
Many types of attributes and values can be found in datasets. Data may, for instance, have noisy values, duplicate data, or missing values. Data preparation in this scenario must be performed. Data preprocessing must therefore reduce lost values and noisy data. [16].

Data preprocessing, data normalisation, and the translation of categorical data into numerical data were all done. From various data sets, a standard data format is produced through data standardisation. It deals with the transformation of datasets after the data are obtained from multiple sources and before they are integrated in the chosen study subject. The sklearn.preprocessing was employed.

4.3 Model Flow

The considered WBCD dataset is split into test and train sets. Data is pre processed to remove null values, inconsistency in the data and any noise in the data.

Figure IV: Flowchart

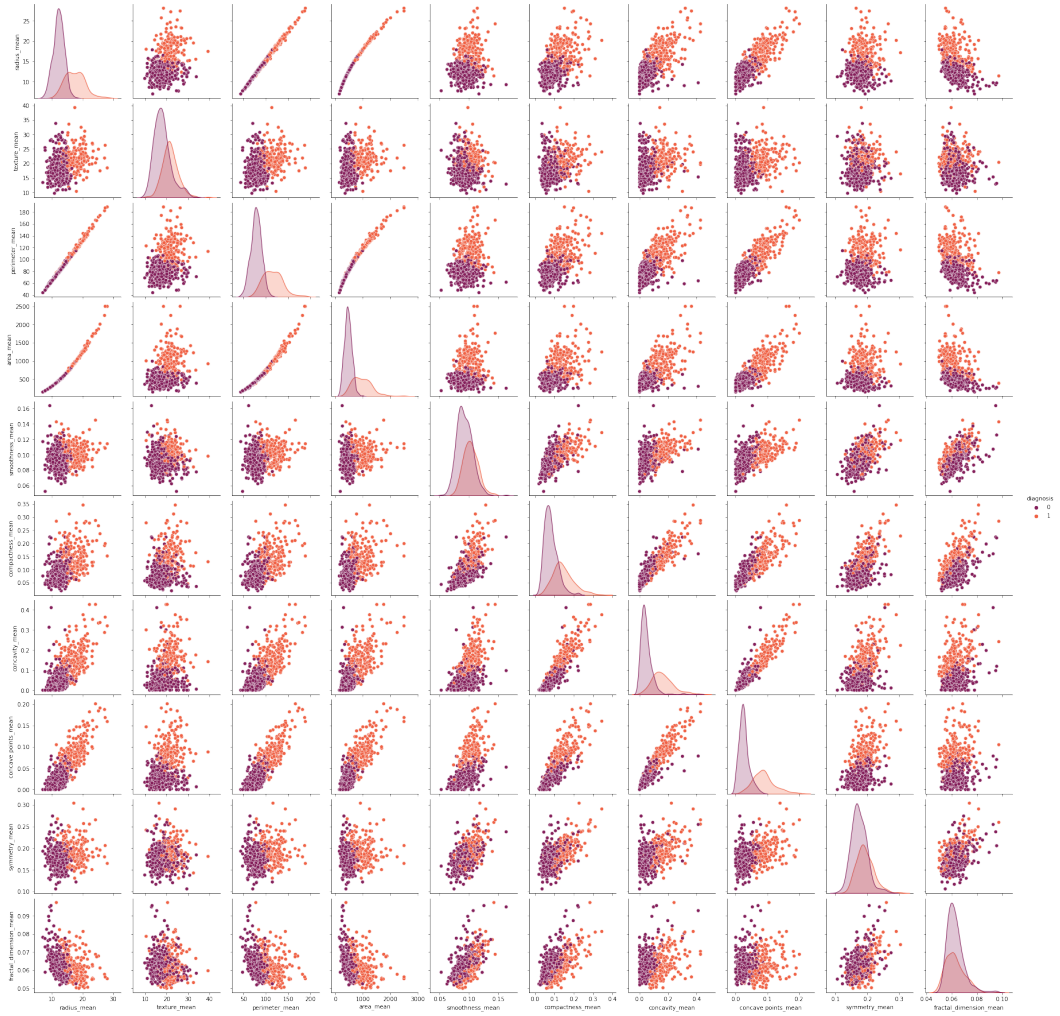


Further, the sklearn library's StandardScaler structure is used to standardise the data. Data with extremely large values and points with extremely small values were compressed using the fit transform() function in the StandardScaler class.

$$y = (x - mean)/standarddeviation$$

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data The following is the scatter plot matrix with mean columns.

Figure V: Scatter Plot

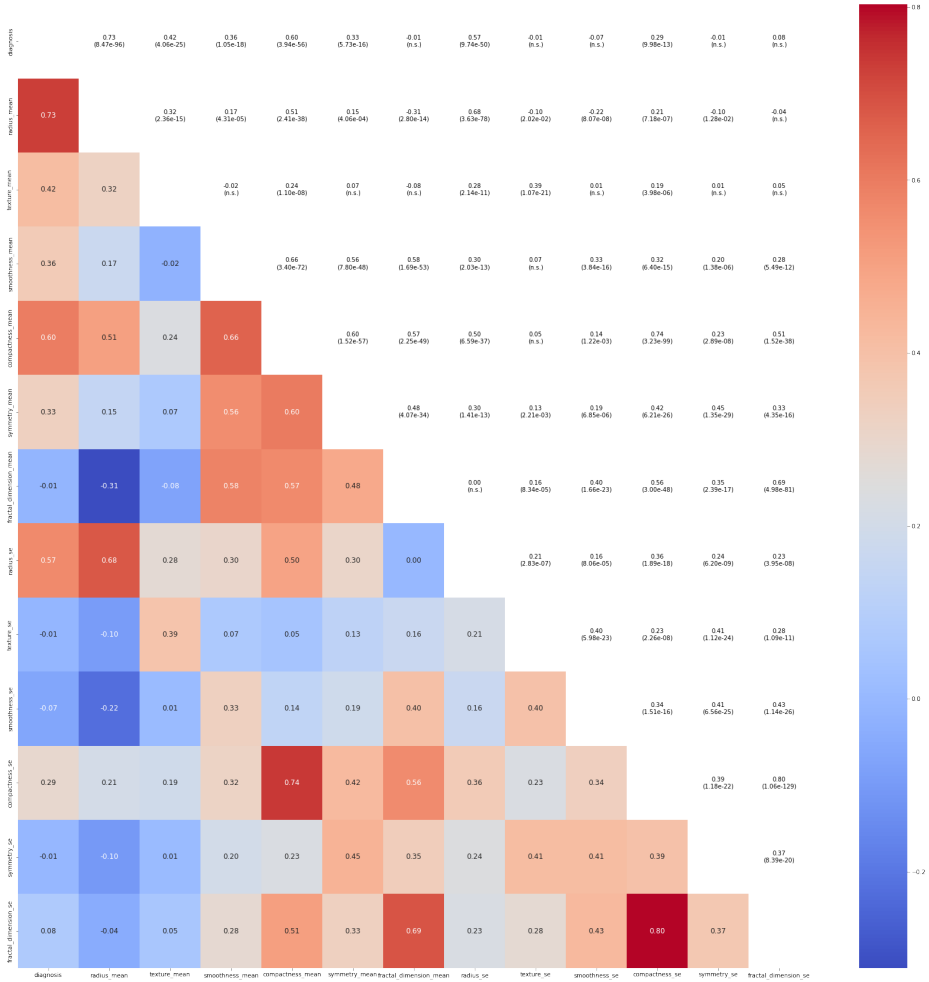


From the plot it can be concluded that almost perfectly linear patterns are observed between the radius, perimeter and area attributes alongside hinting at the presence of multicollinearity between these variables. (they are highly linearly related) Another set of variables that possibly conclude multicollinearity are the concavity, concave points and compactness.

Along with finding the correlation coefficient of the various features of datasets we also found the p-value of the features. When the p-value is more than 0.05 the correlation coefficients are considered

to be non-significant and are considered significant only when the p-value is more than 0.05. The following is the correlation coefficients and p-values for each column.

Figure VI: Correlation Heatmap



4.4 Machine Learning Algorithms

The model was created using machine learning algorithms such the Logistic Regression, Decision Tree, Random Forest, k Nearest Neighbour, Support vector machines and Naive Bayes .

The chance that a sample belongs to a certain class is calculated using the linear classification technique known as logistic regression. The goal of logistic regression is to identify the best decision boundary for class separation. It represents algorithms for the True or false or yes or no categories.

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result.

The Random Forest algorithm is one of the supervised classification algorithms. It is used in both classification problems and regression. The technique generates many decision trees in order to improve classification performance during computation. A random forest algorithm selects the decision tree with the highest score out of numerous ones that operate independently of one another. The following is the entropy formula:

$$E = -\sum P_i * log_2(P_i)$$

kNN stands for k-Nearest Neighbours. It is a supervised learning algorithm. This means that we train it under supervision. We train it using the labelled data already available to us.

$$y* = \max_{i \in \mathcal{N}(x*)} \sum (y_i = c)$$

$$\hat{y}_q = \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{k} w_i \cdot [y_i = y]$$

Support vector machines (SVMs) are vector space-based machine learning techniques that determine which class is farthest away from the other in training data. Data from binary classification are primarily discriminated using SVM. The loss function reduced through SVM is denoted by the below formula:

$$f_i(\theta) = max(1 - \theta^T x, 0)$$

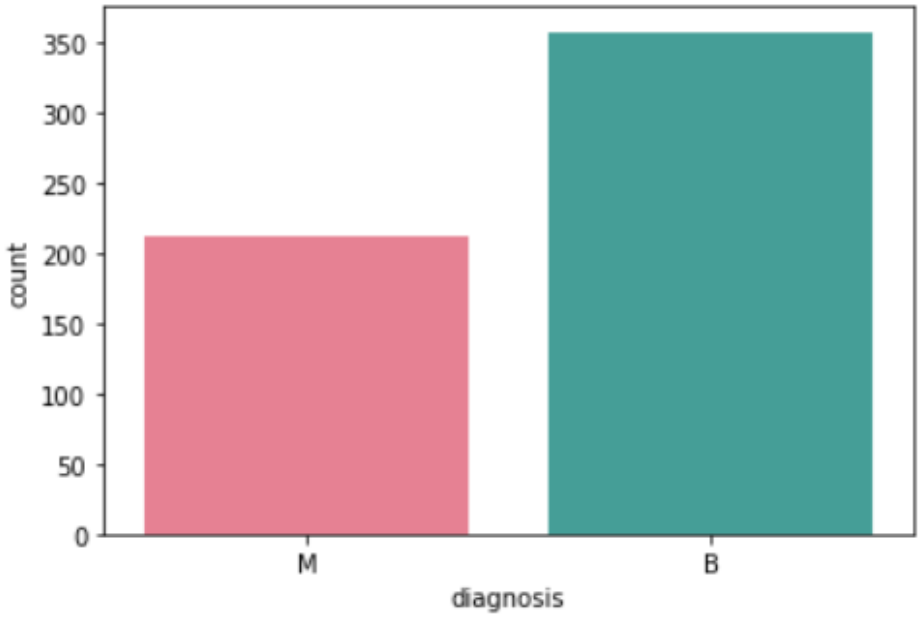
The likelihood that a particular feature vector will be linked to a label is determined by the Naive Bayes classification method, which is based on the Bayes theorem. Since the algorithm expects all features to be independent, which is not always the case, Naive Bayes makes the naive assumption of conditional independence for each feature.

Cross Validation allows us to compare different machine learning methods and see how well they will work in practice. The data set is divided into two categories, training set and testing set respectively. In machine learning, estimating the parameters is called 'training the algorithm' and evaluating a method is called 'testing the algorithm'. In cross validation the whole data set is divided into n number of blocks from which 1 block is used for testing and n-1 blocks are used for training. Therefore there exists many types of cross validation methods such as Four - Fold Cross Validation, Ten - Fold Cross Validation and Leave One Out Cross Validation. In practice it is common to divide data into ten blocks, that is, Ten - Fold Cross Validation. We use Cross Validation to determine how many misclassifications and observations to allow inside of say, the soft margin in an svm, to get the best classification.

4.5 Evaluation Metrics

We shall be using precision, recall, F-measure, and accuracy values. Precision is the proportion of estimated positive values. Recall value is a metric that shows how many transactions need to predict as positive. The harmonic mean of recall and accuracy is known as the F measure. The most used metric for accuracy is the proportion of properly identified samples to the total number of samples method.

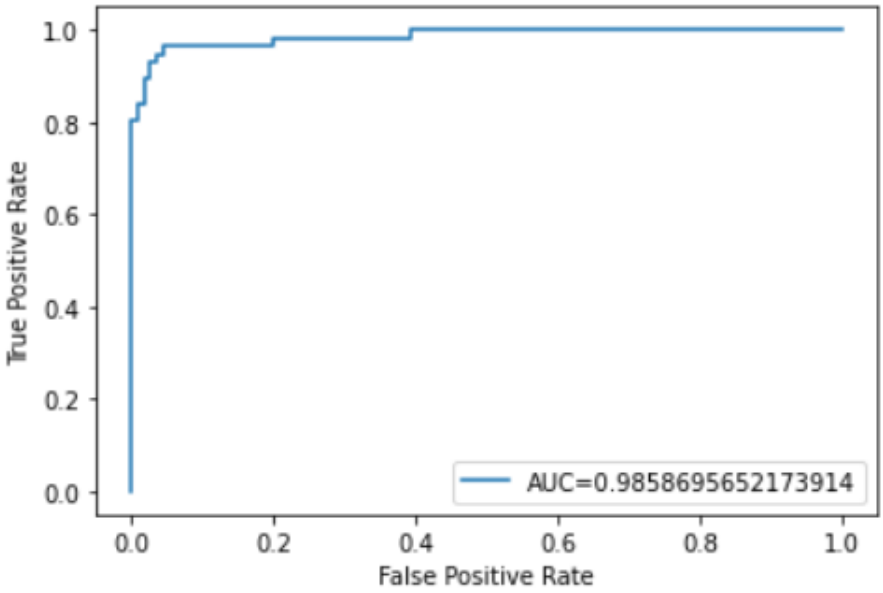
Figure VII: Class Attribute Frequency



4.6 Accuracy and Model Selection

It is seen that the model has 0.98 value for AUC which implies the considered data set is good and satisfactory for the parameters of specificity and sensitivity. It can be concluded that k Nearest Neighbour and Support Vector Machine gives the best results for our dataset. We shall perform cross validation on these two algorithms.

Figure VIII: AUC



5 EXPERIMENTAL RESULTS

5.1 Model Accuracy

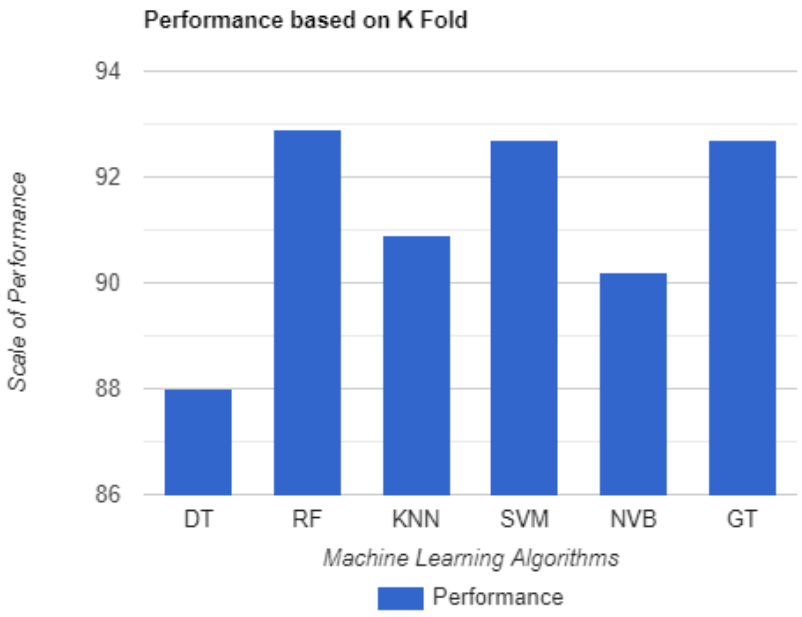
A set of experiments were performed using Anaconda Navigator, Jupyter IDE in python programming language. To diagnosis breast cancer, data mining methods (DT, SVM, KNN, NVB, RF, GT are performed. The comparative analysis of these machine learning techniques is performed in terms of Accuracy, Precision, Recall is mentioned in Table 1. Here Precision, recall and R2 performed on all attributes.

Table I: Accuracy and Model Selection

Model	Accuracy(%)
Logistic Regression	95.90
Decision Tree	89.47
Random Forest	91.81
K Nearest Neighbor	96.49
Support Vector Machine	96.49
Naive Bayes	92.98

5.2 Performance measure on the basis of "10-Fold cross validation

Figure IX: K- Fold Graph of models



Cross-validation [17] is an approach to evaluate the performance of a machine learning model on a validation set. Cross validation use some computation technique in place of mathematical analysis". In research work conducted, 10 fold cross validation is using to validate the classification model. The whole data is divided in 10 equal parts, where 9 parts are applied for training and 1 part for validation as testing the model. The process is repeated 10 times, with every time each of the 10 sub-samples uses at least once.Fig. 3 represent the performance comparison of the techniques separately on the basis of "10 fold cross validation". Here, X axis signifies ML model and Y axis signifies scale by which performance is measured. Performance of DT: 90% RF: 93% KNN: 90% SVM: 92% NVB: 90% GT: 92%

According to the graph,RF model performs better among all other models on the basis of "10-fold cross validation".

5.3 Gradient Boosting for improving model accuracy

Gradient Boosting in machine learning is a technique used in many tasks such as regression and classification. Multiple weak models are combined together to form a model that gives less prediction error and better prediction, hence, better performance. The process that it follows is that of adding models iteratively to form an ensemble, with each new resulting model attempting to make better predictions and therefore correct the errors of the previous model. The models usually used are decision trees. In Gradient Boosting we begin by training a decision tree model on the data which in this case is the WBCD data set. We split the data into two categories for training and testing and then

this model makes predictions on the testing data after it has been trained. But the initial model does not perform very well. We then make use of the Gradient Boosting technique to add decision tree models to the ensemble. The new decision tree models added are trained to predict remaining errors of the previous model.

In our dataset, the first decision tree model identifies a group of patients who are likely to develop breast cancer, but it incorrectly includes some patients who are not likely to develop breast cancer. The second decision tree added to the ensemble is trained to help in minimizing and correcting the errors found in the first decision tree model, specifically by focusing on predicting the diagnosis of the misclassified patients. This cycle of adding new models and correcting the errors of previous models will go on until the ensemble of models reach an accuracy level that is desired. Gradient Boosting optimizes the cost function. This is done by iteratively adding new models to the ensemble that results in reducing or correcting the errors of the previous model. The weighted sum of the individual model predictions helps in making the final prediction. Therefore, this way the gradient boosting technique helps in improving the accuracy of the models.

6 CONCLUSION

The main goal of the research being done is to increase the accuracy of diagnosis through improving breast cancer prediction. The majority of the studies that have been proposed over the past several years are given, with a focus on the creation of classification and machine learning-based predictive models for the diagnosis and prognosis of breast cancer.

7 FUTURE WORKS

There are many areas in which research can be done to gain more data and improve the predictions of breast cancer. Instead of using population data we can start focusing on collecting data that is personalized to a person's genetics, environmental factors where that person lives or travels to, they're medical history and if they have taken or are taking a certain type of medication. Combining their medical records and lifestyle routine can also provide us with information about a person's likelihood to get breast cancer. We can also look into developing methods to improve the quality of the data collected so that the predictive models can become more reliable.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/breast-cancer,2021.

[2] William H. Wolberg, W. Nick Street, Olvi L. Mangasarian, UCI, University of Wisconsin - 1995.

[3] Wang, Haifeng Yoon, Sang Won.'Breast Cancer Prediction Using Data Mining Method',Conference Paper, 2015.

[4] Shomona G. Jacob and R. Geetha Ramani, "Efficient Classifier for Classification of Prognosis Breast Cancer Data Through Data Mining Techniques," Proceedings of the World Congress on Engineering and Computer Science 2012, Vol. I, October 2012.

[5] A. Lambrou, H. Papadopoulos, A. Gammerman, "Evolutionary Conformal Prediction for Breast Cancer International Conference on Emerging Research in Electronics, Computer Science and Technology – 2015.

[6] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer," Annals of Medicine and Surgery, vol. 62, pp. 53–64, 2021.

[7] Mandal, Subrata Kumar. "Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree." International Journal Of Engineering And Computer Science 6, no. 2 (2017).

[8] Salama, Gouda I., M. Abdelhalim, and Magdy Abd-elghany Zeid. "Breast cancer diagnosis on three different datasets using multiclassifiers." Breast Cancer (WDBC) 32, no. 569 (2012): 2.

[9] Aruna, S., S. P. Rajagopalan, and L. V. Nandakishore. "Knowledge based analysis of various statistical tools in detecting breast cancer." Computer Science Information Technology 2 (2011): 37-45.

[10] K. Shilpa, T. Adilakshmi and K. Chitra, "Applying Machine Learning Techniques To Predict Breast Cancer," 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), Chennai, India, pp. 17-21, doi: 10.1109/ICPS55917.2022.00011, 2022.

[11] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 217-221, doi: 10.1109/ICIPTM54933.2022.9754059.

[12] Prerita, N. Sindhwani, A. Rana and A. Chaudhary, "Breast Cancer Detection using Machine Learning Algorithms," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-5, doi: 10.1109/ICRITO51393.2021.9596295.

[13] A. Rovshenov and S. Peker, "Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-6, doi: 10.1109/IISEC56263.2022.9998248.

[14] D. Yifan, L. Jialin and F. Boxi, "Forecast Model of Breast Cancer Diagnosis Based on RF-AdaBoost," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, pp. 716-719, doi: 10.1109/CISCE52179.2021.9445847.

[15] Vikas Chaurasia, Saurabh Pal and BB Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," 2018 Journal of Algorithms Computational Technology, doi: 10.1177/1748301818756225

[16] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2018, pp. 997-1002, doi: 10.1109/ICCMC.2018.8487537.

[17] Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." Statistics surveys 4 (2010): 40-79

[18] Splane, B. (2022, October 17). What Is a Benign vs. Malignant Tumor? Verywellhealth. Retrieved March 29, 2023, from https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240