Confidence
Estimation of the models
While classifying breast cancer

Group #3

Smitha Ramakrishna sramakr1@kent.edu

Varun Reddy Pisati vpisati@kent.edu

Sai Varshini Palle spalle2@kent.edu

Pavan Kumar Reddy Gunnala pgunnala@kent.edu

1. Introduction

Cancer is a disease where cells grow out of control. Of the variations of cancer, breast cancer is the second most vulnerable one, for which deaths in women are the most. The probability that a woman dies due to breast cancer is 1 in 39, which is around 2.5%. Among 264,000 being diagnosed, around 42,000 women in the U.S. die each year from breast cancer. Women whose breast cancer is detected at an early stage have a 93% survival rate in the first five years. Even though Mammogram screening is used to detect the tumor, it is not perfect. These statistics direct us to a machine-learning approach to detect breast cancer, as the traditional method is not that trustworthy. Detection of breast cancer at an early stage is possible if we solve this problem with a machine-learning approach. But the machine learning model's accuracy metrics are not everything, as there might be some data leakage and we must think about the confidence of the model to really rely on that model's predictions. So, apart from only predictions and accuracy, the confidence of the model should be checked with more importance. It's a challenge for doctors to identify each breast cancer patient. 50% of breast cancers were not detected in Mammogram screenings of women with very dense breast tissue. About 25% of women with breast cancer are diagnosed negatively within two years of screening. Experts suggest mammography screening along with machine learning can help in a more accurate diagnosis of women at risk. So, detection of cancer in the early stage is crucial, as the survival rate decreases with respect to time.

2. Project Description

- The project was created to get the accuracy of prediction obtained from different models. The goal of the project is to tune the algorithm in such a way that the accuracy of prediction is maximized.
- The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical feature detection from complex BC datasets,

machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling.

- Classification and data mining methods are effective ways to classify data. Especially in the medical field, where those methods are widely used in diagnosis and analysis to make decisions.
- Develop an understanding of the purpose of the data mining project, Obtain the dataset to be used in the analysis, Explore, clean, and pre-process the data, Reduce the data dimension, if necessary.
- Determine the data mining task (classification, prediction, clustering, etc.), Partition the data (for supervised tasks). Choose the data mining techniques to be used (regression, neural nets, hierarchical clustering, and so on).
- Use algorithms to perform the task: This is typically an iterative process—trying multiple variants, and often using multiple variants of the same algorithm (choosing different variables or settings within the algorithm). Where appropriate, feedback from the algorithm’s performance on validation data is used to refine the settings. Interpret the results of the algorithms: This involves making a choice as to the best algorithm to deploy, and where possible, testing the final choice on the test data to get an idea as to how well it will perform. (Recall that each algorithm may also be tested on the validation data for tuning purposes; in this way, the validation data become a part of the fitting process and are likely to underestimate the error in the deployment of the model that is finally chosen.).
- Deploy the model: This step involves integrating the model into operational systems and running it on real records to produce decisions or actions. For example, the model might be applied to a purchased list of possible customers, and the action might be “include in the mailing if the predicted amount of purchase is > $10.” A key step here is “scoring” the new records or using the chosen model to predict the outcome value (“score”) for each new record.

- **Challenges and technical contributions**

We want to find out how our model will work to predict breast cancer while there are three types - benign, malignant & normal in our dataset. On the other hand, we are more interested in finding the confidence of the model by changing the parameters while looking into which is the best model to rely on, in terms of confidence.

Dataset being used in this project can be found on Kaggle or you can use this link to go directly to the dataset.

- **The workload distribution for each member in your team**

There are several stages to this project. Data is first gathered and examined to establish what data is accessible or, in the lack of specifics, relevant data is offered. The user performs chat interactions according to the requirements in the second step, and the execution and presentation are done in the third. The task will be evenly distributed among the group members, and each stage will be disclosed to the others.

3. Background

- **Related papers (or surveys for graduate teams)**

  Different authors seeking to propose solutions for the early and efficient detection of breast cancer have made great contributions using data mining, machine learning, artificial intelligence and big data methods. According to the literary analysis different studies are associated with the analysis of breast cancer using different Dataset, SEER is used in where a study is presented for the prediction of breast cancer survival, a study where mining methods are compared of data DT decision trees, artificial neural networks and the statistical method of logistic regression, the best results achieved are associated with the DT method reaching a level of 93.6% accuracy, then the second best result is obtained by RNA with 91.2% and finally Logistic Regression with 89.2%. In the authors present an analysis of the prediction of the survival rate in patients using data mining methods, the methods used for the experimentation process were Naive Bayes NB, Back-Propagated RNA and DT the best results were achieved by the DT method reaching an accuracy level of 86.7%, while Back-Propagated RNA achieved an 86.5% accuracy level and finally the NB method obtained an 84.5%. In an analysis of breast cancer using statistical and data mining methods is presented, according to the authors there are three methods for the diagnosis of said disease which correspond to mammography, FNA (fine needle aspirate) and biopsy, which sometimes are usually expensive and unpleasant, the method that presents the best results is the biopsy with an accuracy level of approximately 100%, however it is possible to obtain better results easily through the implementation of integrated FNA with Data mining methods such as attribute selection, DT, AR association rules and statistical methods such as Principal component PCA analysis, PLS linear regression analysis. In it is presented to the implementation of data mining for the discovery of breast cancer patterns based on the use of RNA and multivariable adaptive regression splines, according to the authors the RNAs have been very popular for prediction tasks and classification, the basis of the analysis is, first, to use MARS to model the classification problem, then the significant variables obtained are used as input variables of the designed neural network model. To demonstrate that the inclusion of important variables obtained from MARS would improve the accuracy of the classification of networks, diagnostic tasks are performed in a breast cancer data set with fine needle aspiration cytology.

  In a novel approach to the detection of breast cancer using data mining techniques is presented, the objective of the proposed study is to compare three classification techniques using the Weka tool where the algorithms of SMO, IBK and BF Tree are used, the data set used corresponds to Breast Cancer Wisconsin, the results obtained show that SMO achieves the best 96.2% accuracy results. A comparative study between the methods of K-means and fuzzy C-Means FCM for the detection of breast cancer is presented said study is focused first on comparing the performance of K-clustering algorithms means and FCM and, secondly, the integration of different computational measures is considered that allow to improve the grouping accuracy of the aforementioned techniques, FCM obtains better results compared to K-means considering that it achieves a 97% level of accuracy compared to the other technique that achieves 92%. A study for the prediction of breast cancer recurrence using data mining techniques is presented, the study proposed proposes the use of different classification algorithms such as C5.0, KNN, Naive Bayes, SVM and as K-Means, EM, PAM, Fuzzy C-means clustering method, the experimentation performed evidence that the best results are achieved by C5.0 with an accuracy level of 81.03%.

- **Software tools (DBMS, GUI, IDE, existing library, …) used**

  Jupyter Notebook
  Python 3: Python Programming skills
  Libraries: warnings, NumPy, pandas, matplotlib.pyplot (for visualization of data), seaborn, sklearn.model (all ml-algorithms are available in this library)

- **Required hardware**
  Minimum i3 configuration
  4 GB RAM
  256 ROM

4. Problem Definition

In a system for the automatic diagnosis of breast cancer based on the AR Association Rules method as attribute reduction technique and Neural Network NN as classification technique is presented, the data set used corresponds to Wisconsin Breast Cancer During the training and validation process, the 3-fold cross-validation method was used, the findings resulting from the experimentation carried out indicate that the AR + NN method achieves a correct classification rate of 95.6%. In the same data set is used, where the authors propose a model for the prediction of benign and malignant ash cancer through the implementation of the Naive Bayes NB, RBF Network and J48 algorithms, the results obtained indicate that NB is the best predictor with 97.3 accuracy while RBF Network obtained 96.77% and finally the j48 algorithm achieves an accuracy level of 93.41%, during the experimentation process cross validation with 10 folds was used. In an application of ML machine learning algorithms using the breast cancer Wisconsin data set for breast cancer detection is presented, for which 6 ML algorithms were submitted for comparison which correspond to the proposed experimentation the data set is divided into 70% for the training phase and 30% for the test phase, the algorithms that obtained the best results It corresponds to MLP who achieved an accuracy level of 99.04%

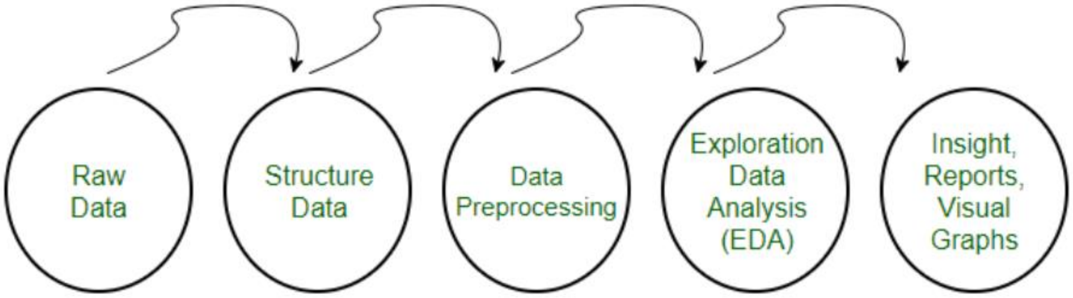
Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we preprocess our data before feeding it into our model.

- Handling Null Values
- Standardization
- Handling Categorical Variables
- One-Hot Encoding
- Multicollinearity

Importance of data preprocessing:

Due to their uneven origin, the majority of real-world datasets used for machine learning are very likely to contain missing data, inconsistent results, and noise**.** Data mining methods would not produce high-quality results when applied to this noisy data because they would be unable to successfully find patterns. Therefore, data processing is important to raising the general level of data quality.

- Missing or duplicate values could present an inaccurate picture of the data's overall statistics.
- False predictions are frequently the result of outliers and inconsistent data points disrupting the model's overall learning process.
- Quality data is required for quality decisions. To obtain this high-quality data, data preprocessing is necessary; otherwise, it would be a case of garbage-in, garbage-out.


Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Data preprocessing involves:

Data quality assessment: It is used to look on out data an gives an idea about the whole quality of our data. There are different types of data anomalies and inherent problems which are involves to look on our data. They are:

- **Mismatched data types:** You could receive data in various formats when you gather information from numerous sources. While reformatting your data for machines is the goal of the entire procedure, you must start with similarly prepared data.

- **Mixed data values**: Various sources utilize different descriptors to describe features, such as man or male. All these value descriptions must be uniform.
- **Data outliers:** The outcomes of data analysis might be significantly impacted by outliers.
- **Missing data:** Look for empty text boxes, missing data fields, or unresolved survey questions. This can be the result of inaccurate or missing data. You must undertake data cleaning to address any missing data.

Data Cleaning: There may be a lot of useless information and gaps in the data. Data cleansing is completed to handle this portion. It entails dealing with erroneous data, noisy data, etc.

- **Missing Data**: when some data is missing in the data. It can be handled in various ways. Some of them are:
- Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
- Fill in the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value
- **Noisy Data**: Data that is noisy has no meaning and cannot be understood by computers, It may be produced as a result of poor data gathering, incorrect data entry, etc.

- **Binning Method**: This technique uses sorted data to smooth it out. The entire set of data is separated into equal-sized pieces before the task is finished using a variety of techniques. Each segment is dealt with independently. To finish the operation, one can use boundary values or replace all the data in a segment with its mean.
- **Regression**: In this case, smoothing the data involves fitting it to a regression function. There are two types of regression that can be used: multiple or linear (having multiple independent variables).
- **Clustering:** This method creates groupings of related data; The outliers might not be noticed or they might be outside of the clusters.

Data Transformation:

This method is used to change the data into formats that are appropriate for the mining process.

This entails the following:

- Normalization is the process of scaling data values to fit inside a predetermined range (-1.0 to 1.0 or 0.0 to 1.0).
- Selection of Attributes: To aid the mining process, new attributes are created from the existing set of attributes in this technique.
- Discretization: This process substitutes interval levels or conceptual levels for the raw values of a numerical attribute.
- Concept Hierarchy Generation: In this step, qualities are raised in the hierarchy from a lower level to a higher level. As an illustration, the attribute "city" can be changed to "country."

Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

- Data Cube Aggregation: For the purpose of building a data cube, an aggregate operation is applied to the data.
- Attribute Subset Selection: The most relevant attributes should be used; the remaining attributes can all be disregarded. Level of importance and the attribute's p-value can be used to choose attributes. The property that has a p-value greater than the level of significance can be eliminated.
- Numerosity reduction: This makes it possible to save data models rather than the entire dataset, such as regression models.
- Dimensionality reduction: uses encoding techniques to lower the amount of the data.

  It may be lossless or lossy. Such reductions are referred to as lossless reductions if the original data can be recovered after being compressed; otherwise, they are referred to as lossy reductions. Wavelet transforms and PCA are the two most effective dimensionality reduction techniques (Principal Component Analysis).

5. The Proposed Techniques

- **Framework (problem settings)**

  This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether breast cancer is benign or malignant. To achieve this, we have used machine learning classification methods to fit a function that can predict the discrete class of new input.

  Programming language: Python – programming

  Dataset: The dataset used in this story is publicly available and was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA.

  Input: Dataset

  Output: Confidence level of model estimations.

- **Details of major techniques (e.g., pruning methods in lemmas/theorems; illustrated with toy examples)**
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier
  - Support Vector Machine (SVM) Classifier
  - Gaussian Naive Bayes Algorithm Model
  - Stochastic Gradient Descent Classifier
  - Gradient Boosting Classifier

- **Query processing algorithms (pseudo code) and query optimizations**

  Data Standardization: The process of converting data to a common format to enable users to process and analyze it.

  In our project we used “sklearn.preprocessing.StandardScaler”

  By eliminating the mean and scaling to unit variance, characteristics are made uniform. A sample x's average score is determined as follows:

  z = (x - u) / s

  where s is the training samples' standard deviation or one if with std=False, and u is the training samples' mean, or zero if with mean=False. By calculating the pertinent statistics on the samples in the training set, centering and scaling are applied independently to each feature. Then, for usage with later data using transform, the mean and standard deviation are recorded.

- **This section can be split into multiple sections if you have many contents to present**

- Importing Dependencies (library & packages)
- Data Preparation --> (Load And Check Data)
  Upload the dataset in CSV format using Panda, we have 32 data features in our dataset.

  Variable/Attribute Description \

  Target--> **(M= malignant, B = Benign) ___**

  1. Ten real-valued features are computed for each cell nucleus:

  - 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
  - fractal dimension ("coastline approximation" - 1)

  The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

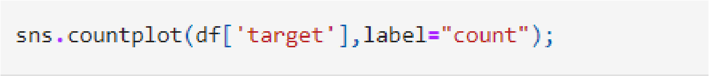
1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant).

We look at the data need for standardization, if there are big differences between the data, standardization is required.

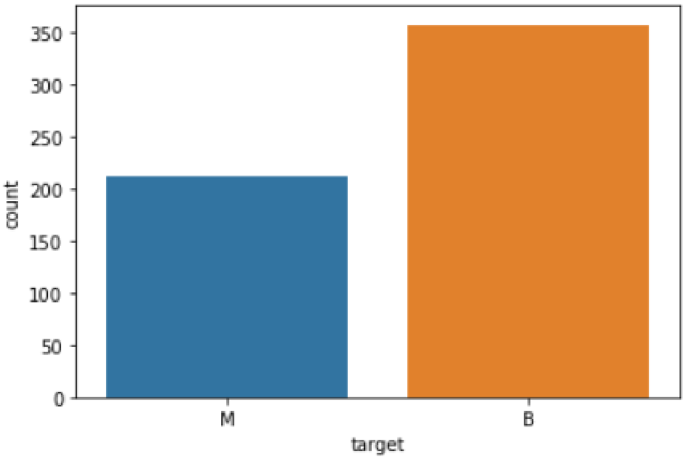
6. Visual Applications

- **GUI design**

There are 32 features in our dataset, In the below graph we have target i.e. Benign and Malignant count



```
sns.countplot(df['target'],label="count");
```



Here, in the above picture we have visualised our data set from the data.csv file and known how many values constitute to benign cancer and malignant cancer. Before applying any machine learning model on that we are making sure what the data we are dealing with, which makes our analysis easy to have a view on what would be the output. This would help us in getting the accuracy or confidence level estimation.

Data Exploration & Analysis

Visualization of data is an imperative aspect to understand data and also to explain the data to another person. Python has several interesting visualization libraries such as Matplotlib, Seaborn etc.

Correlation and heatmaps:

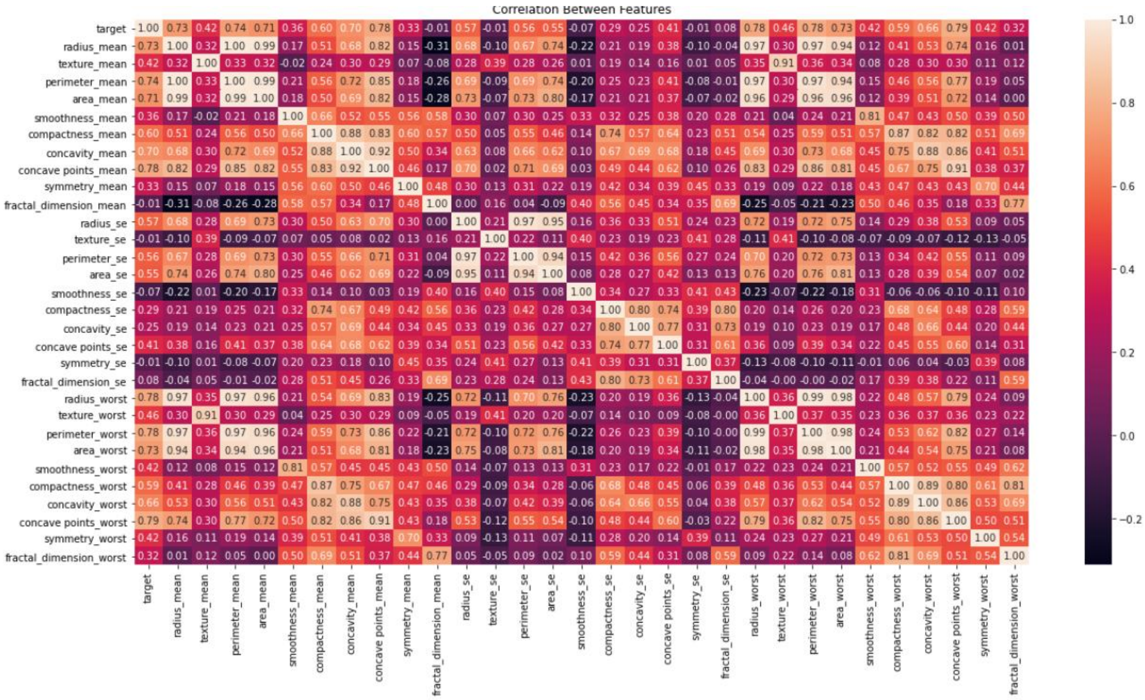
Correlation: It shows the relation between one set of data to another set. The two types of correlation is positive and negative. Positive correlation means the two variables are travel in same direction. if one variable value increases then another variable value also increases. Negative correlation means the two variables travel in opposite directions if one variable value increases than other variable value decreases.

Heat map: It is a graphical representation of data where the values are represented by color. It provides visual context to analyze the problem very easily. Heat maps are an excellent choice for high-throughput data presentation because of their dense and simple layout. On a screen, hundreds of rows and columns may be seen. Heat maps' primary components are color encoding and meaningful row- and column-rearranging.

Correlation and Heat maps: A form of graphic called a correlation heatmap shows the strength of correlations between numerical variables. To determine which variables are related to one another and how strongly they are associated, correlation graphs are utilized.

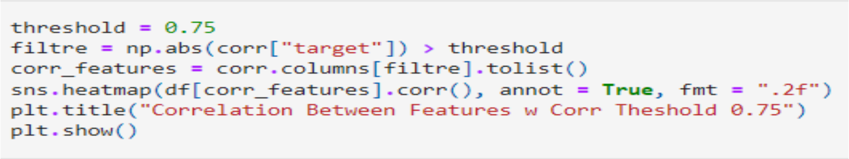
Heatmaps are a quick way to look at 1 on 1 relationships. For a small number of features, you might be able to look at multi-variable linear dependencies. Again, such features should be removed too but they are hard to detect from heatmaps.

Heatmap correlation of the dataset: A heatmap that displays a 2D correlation matrix between two discrete dimensions and uses colored cells to represent data from typically a monochromatic scale is called a correlation heatmap. The first dimension's values are displayed as the table's rows, while the second dimension's values are displayed as columns.

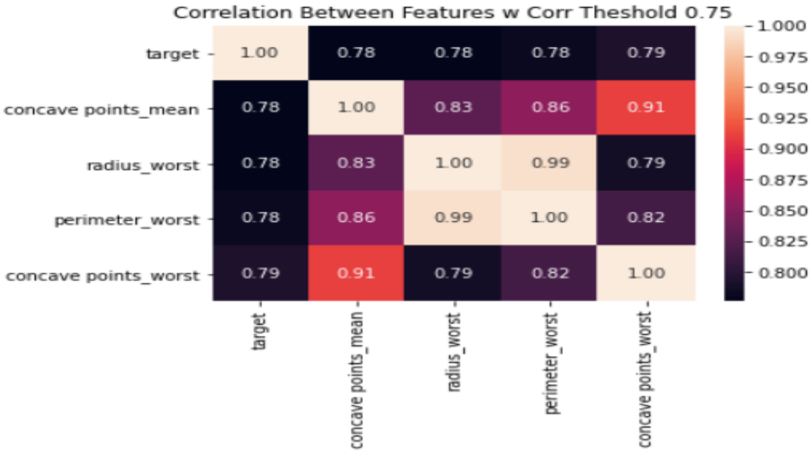


In the above heatmap we have visualized the correlation between the attributes. Here we have took the mean of ten real valued features around the nucleus of the cancer to compute the correlation. We have mean, se and worst for all ten attributes which we constituted as features and visualized the heatmap in the above image.

- First, we set a limit value. Here we set it to 0.75. We bring the ones whose relationship between properties is greater than 0.75.



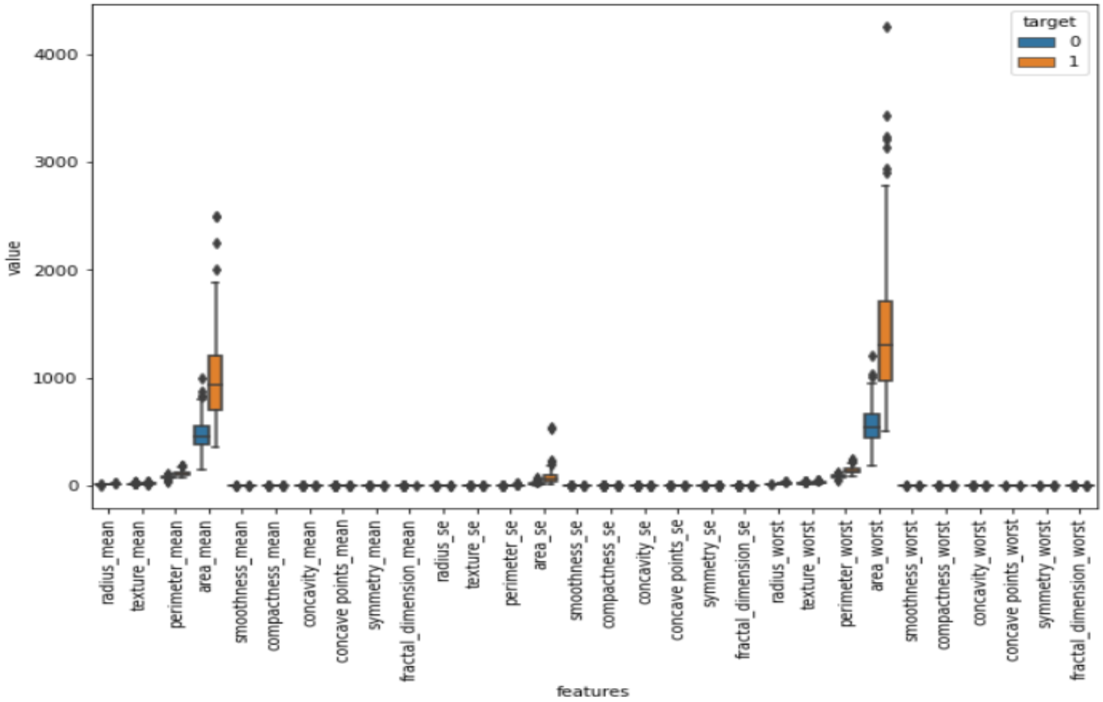
```
threshold = 0.75
filtre = np.abs(corr["target"]) > threshold
corr_features = corr.columns[filtre].tolist()
sns.heatmap(df[corr_features].corr(), annot = True, fmt = ".2f")
plt.title("Correlation Between Features w Corr Theshold 0.75")
plt.show()
```



Here we generated an heatmap with correlation between features with respect to correlation threshold. The threshold which we took here is 0.75 and we also rounded the values upto two decimal points.

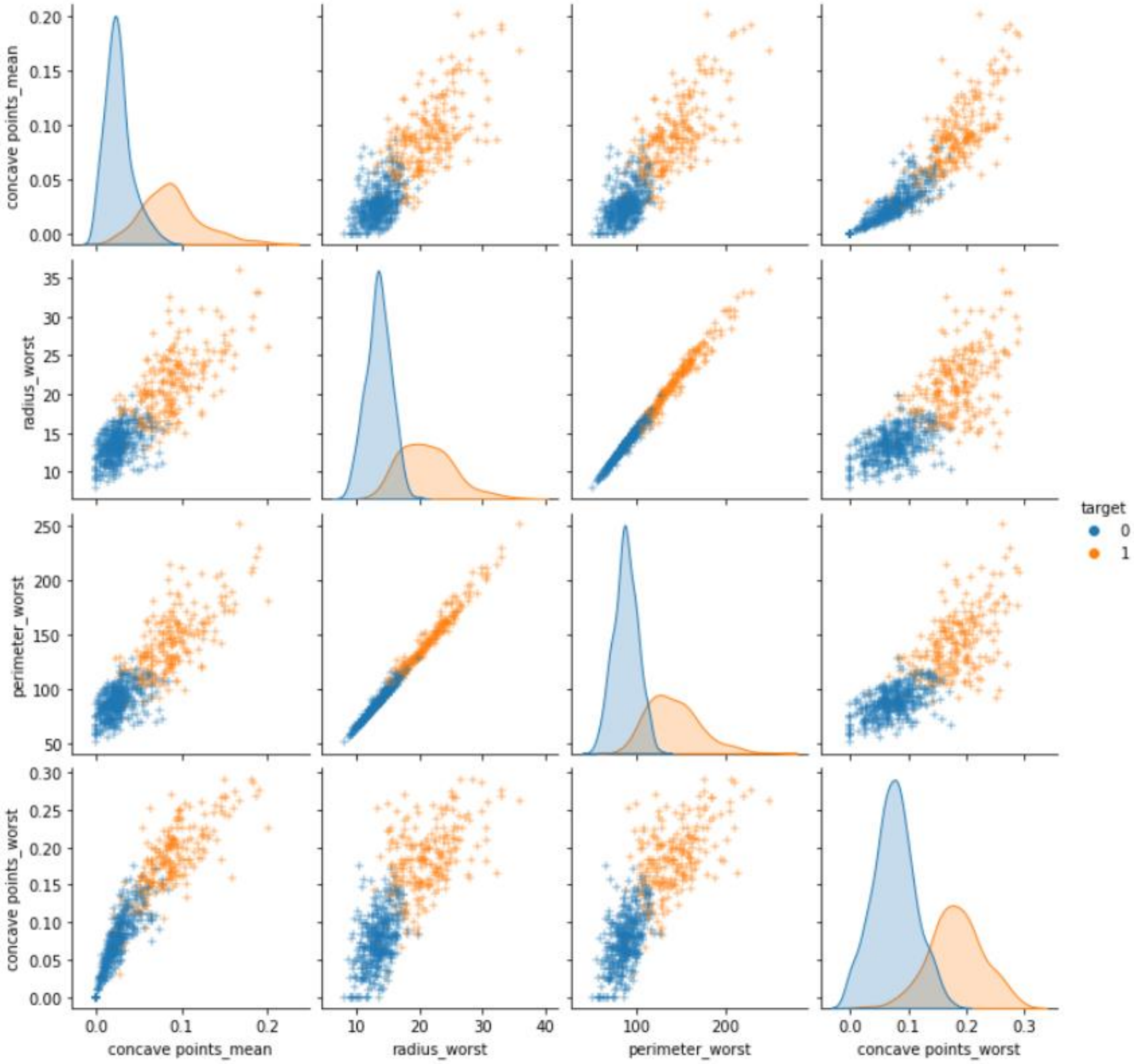
- **Design modules (with descriptions, figures, and/or flowcharts)**

Plotting boxplot:



Here In this boxplot we have visualized our dataset according to the values with respect to attributes we used for the analysis. The orange box plotting describes the value we got is 1 which means the cancer is Malignant and the blue box constitutes for the value of 0 which means the cancer is benign.

Plotting data:



Now we are plotting the dataset in order to know the types of cancer constitutions in the dataset. Here we have plotted these graphs for the following reasons

a) analyze the properties of real-world graphs,

(b) predict how the structure and properties of a given graph might affect some application, and

(c) develop models that can generate realistic graphs that match the patterns found in real-world graphs of interest.

We have plotted these graphs for various attributes like concave points, perimeter, radius and concave points of and cancer.

Categorical data

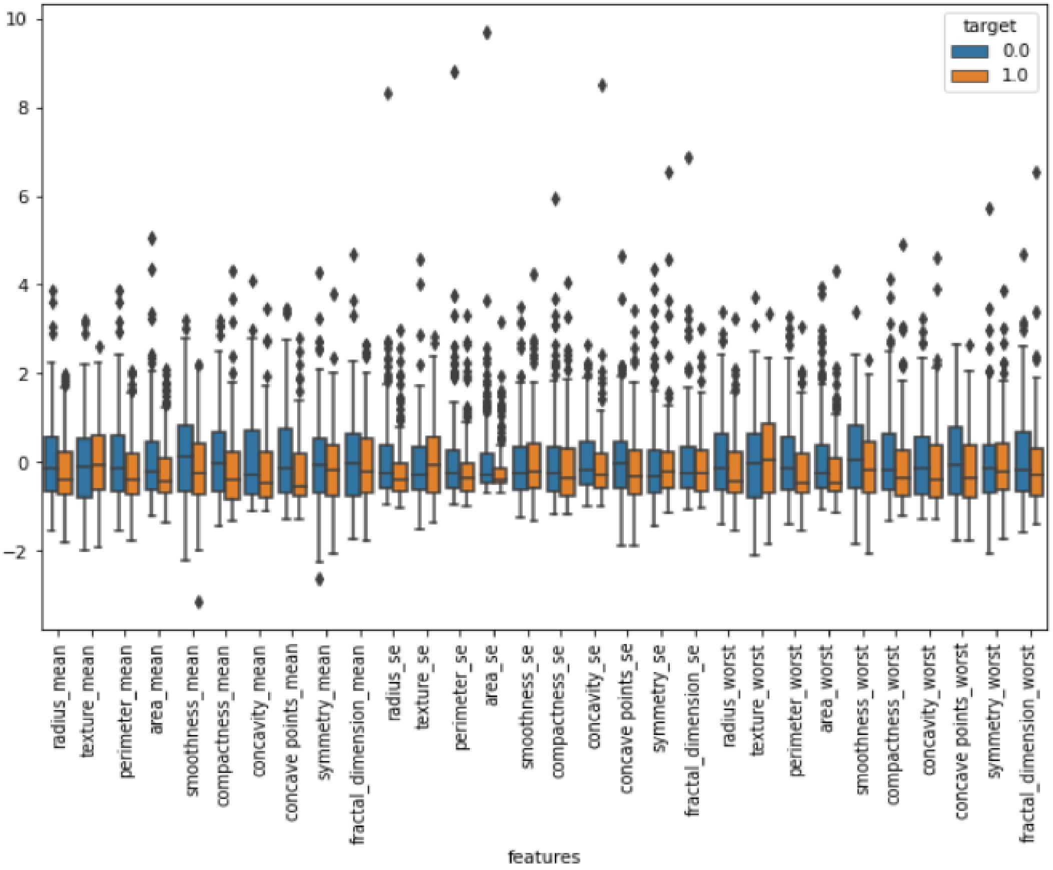
Categorical data are variables that contain label values rather than numeric values.The number of possible values is often limited to a fixed set. There are three types of categorical data they are Binary, Nominal and ordinal variables.

Splitting the dataset :

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. We will do this using SciKit-Learn library in Python using the train_test_split method.

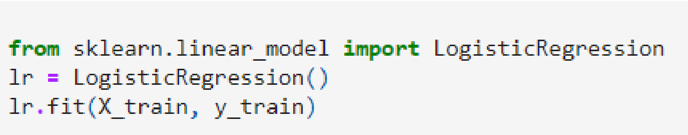
Feature scaling:

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling. This means that you're transforming your data so that it fits within a specific scale, like 0–100 or 0–1.
We will use StandardScaler method from SciKit-Learn library.



Model Selection:

- **Logistic Regression:** It is a supervised learning algorithm. It is used to compute prediction and classification problems. In this we are used this algorithm to predict that the patient is having the benign or malignant cancer. It is calculated based on the given dataset.



```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train, y_train)
```

The above code tells us how we imported the model, trained and testing our dataset. All the other machine learning models in this project followed the same way as this, in order to run their processes and to find the accuracy of that model. The same way be followed by other methods to attain their confidence levels.

Logistic Regression is three types they are:

Binary Logistic Regression: When we have Two possible Outputs like True or False it showed in binary form. For example, does this patient have cancer or not.

Multinomial logistic regression: This means that we have multiple possible outputs are there, for example, The patient is having the Ductal carcinoma in situ, Invasive breast cancer, Triple-negative breast cancer, Inflammatory breast cancer etc. in Breast Cancer.

Ordinal logistic regression: When the results are ordered. In this we are using this to check the severity of breast cancer. That the cancer is mild, moderate and severe.

Training data assumptions for logistic regression:

The assumptions are given below:

1. The anticipated result is categorically binary or dichotomous. (This applies for binary logistic regression).
2. The variables that determine the outcome, or factors, are independent of one another. In other words, the independent variables' multicollinearity is either minimal or nonexistent.
3. The independent variables and the log odds may be related linearly.
4. Sufficiently large sample sizes.

Mathematical Calculation for Logistic Regression:

The probability is constantly between 0 (does not occur) and 1. (happens). Using our Covid-19 example, the likelihood of testing positively and not testing positively will total to in the case of binary categorization.

In logistic regression, probability is calculated using the logistic function or the sigmoid function. A straightforward S-shaped curve called the logistic function is used to translate data into a value between 0 and 1.

$$h\Theta(x) = \frac{1}{1+e-(\beta 0+ \beta 1)}$$

hΘ(x) is the output logistic functions, where $0 \leq h\Theta(x) \geq 1$

β1 is the slope

β0 is the y- intercept

(β0 + β1* X) derived from the equation of a line Y(predicted) = (β0 + β1* X)+Er

- **Decision Tree Classifier:** It is a non-parametric supervised algorithm. It is used to compute classification and regression. This is used to learn simple decision rules derived from the data features in order to build a modal that predicts the value of a target value. 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**.** We are using this to compare the predictive results classifying breast cancer patients using specific data. The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.

Given certain parameters, it provides a graphic representation of all possible solutions to an issue or decision.

Uses of Decision Tree:

1. Decision trees are typically designed to mimic how people think when making decisions, making them simple to comprehend.
2. Because the decision tree displays a tree-like structure, the logic behind it is simple to comprehend.

How does the Decision Tree algorithm Work?

Step 1: S suggests starting the tree from the root node, which has the entire dataset.

Step 2: Use Attribute Selection Measure to identify the dataset's best attribute (ASM).

Step 3: Subsets of the S that include potential values for the best attributes should be created.

Step 4: To determine the best attribute in the dataset, use Attribute Selection Measure (ASM).

Step 5: Using the subsets of the dataset generated in step 3, repeatedly design new decision trees. Continue along this path until you reach a point when you can no longer categorize the nodes and you refer to the last node as a leaf node.

Attribute Selection Measure:

The fundamental problem that arises while developing a decision tree is how to choose the best attribute for the root node and for sub-nodes. So, a method known as attribute selection measure, or ASM, can be used to tackle these issues.

There are two best Techniques to perform ASM, which are:

1. Information Gain
2. Gini Index

Information Gain:

1. Following the segmentation of a dataset based on an attribute, information gain is the measurement of changes in entropy.
2. It figures out how much knowledge a feature gives us about a class.
3. We divide the node and create the decision tree based on the value of the information gained.
4. A node or attribute with the largest information gain is split first in a decision tree algorithm, which always seeks to maximize the value of information gain. Using the formula below, it can be calculated:

- **Random Forest Classifier:** It is a powerful and versatile supervised algorithm. It grows and is used to combine several decision trees and form as a "Forest". Instead of depending on a single decision tree, the random forest uses predictions from each tree and predicts the result based on the votes of most predictions. This classifier combines multiple data and forms as trees and gives the result that the patient is having the beast cancer or not. This classifier modal ha 98% of accuracy rate.

Uses of Random Forest Algorithm.

1. In comparison to other algorithms, it requires less training time.
2. Even with the enormous dataset, it operates effectively and predicts the outcome with a high degree of accuracy.
3. When a significant amount of data is lacking, accuracy can still be maintained.

How random Forest Algorithm works:

First, N decision trees are combined to generate the random forest, and then predictions are made for each tree that was produced in the first phase.

The working process is:

Step1: Pick K data points at random from the training set.

Step2: Create the decision trees linked to the chosen data points (Subsets).

Step3: Choose the number N for the decision trees that you want to build.

Step4: Repeat step3 and step4.

Step5: Find each decision tree's forecasts for any new data points, then place them in the category that receives the most votes.

Advantages:

The forecasts from each decision tree for any new data points are located and then placed in the category with the highest number of votes.

Support Vector Machine (SVM) Classifier: The importance of using this classifier is to analyze the data for finding patterns which are used for classification and regression. An SVM modal can classify new text after being given sets of labeled training data for each category. In this we are used this algorithm to classify whether the Cancer is benign and malignant.

Gaussian Naive Bayes Algorithm Model: It is a probabilistic classification algorithm. This algorithm is based on Bayes Theorem with strong independence assumptions. Here we are used this classifier to predict the cancer tumor with 98% accuracy rate.

Stochastic Gradient Descent Classifier: This classifier is used to identify the modal parameters that perfectly meet the predicted and actual outputs.

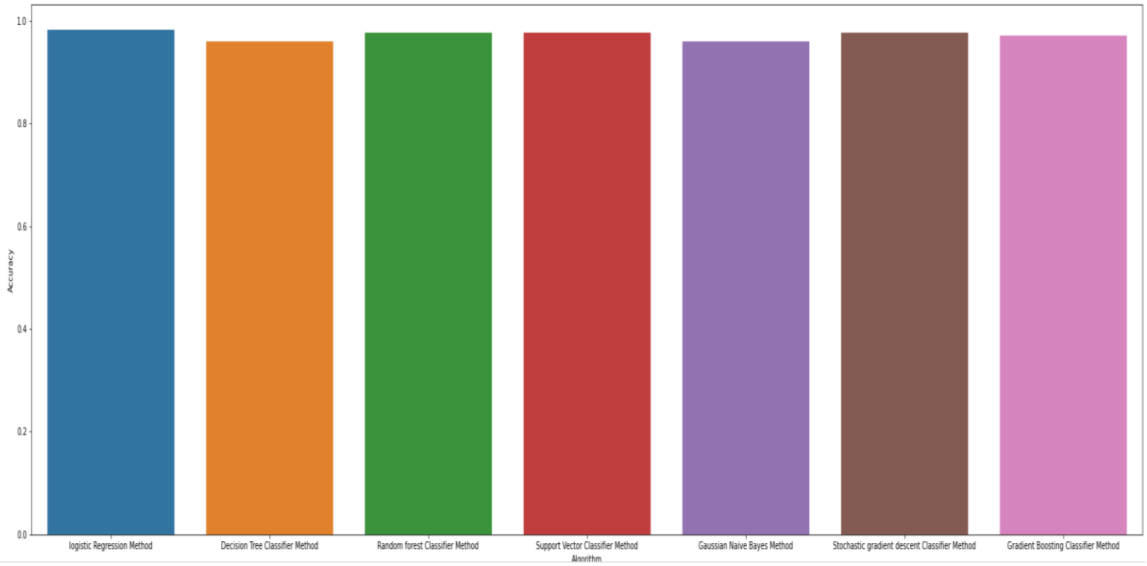
Gradient Boosting Classifier: It is a machine learning algorithm which is used to combine the number of week signals to produce a powerful predicting modal. During this algorithm process, random samples are initially created from the training data. For this sample a classifier was trained, and the whole training set was put to the test. For every sample estimate, an error was determined. If the sample is incorrectly identified, its weight is increased and a new sample. is madeup the system achieves a high level of accuracy, these procedures are repeated.

Perform Comparative Analysis of each & every 7-algorithms & then Conclude to the best-mode.

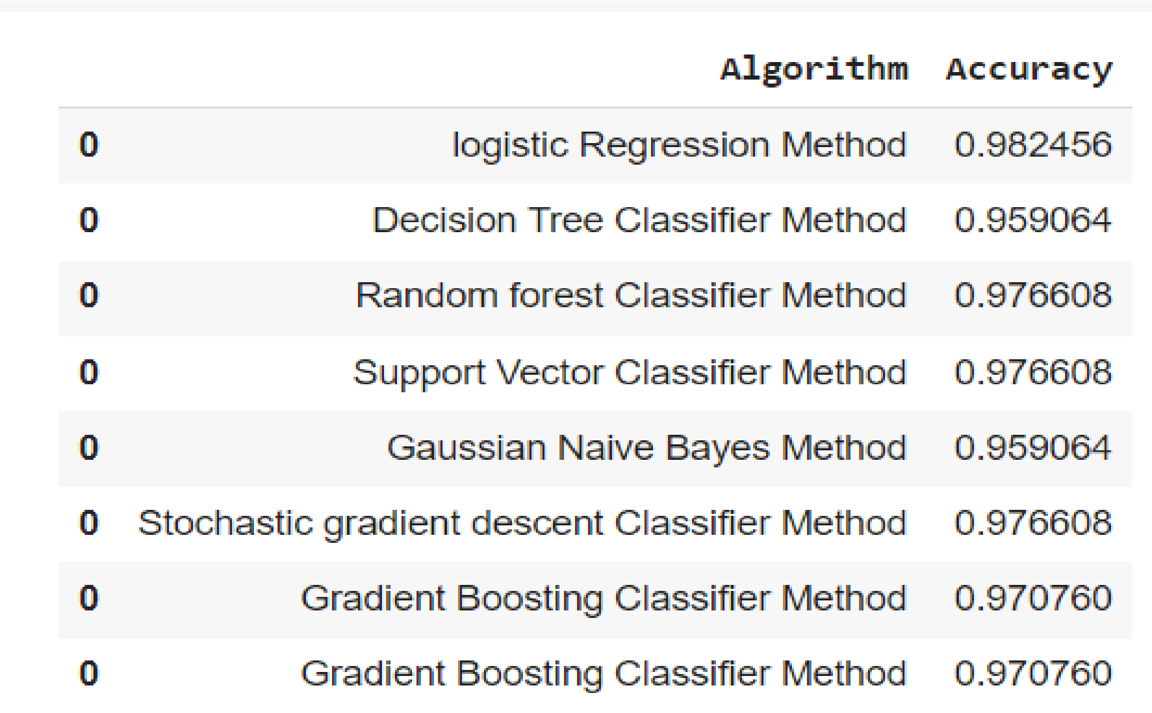
Comparative analysis: It is used for comparison of two or more processes, documents, datasets or other objects. Here we are using this analysis to compare the above modals result and finally we have to choose the one modal which gave the result more accurately.

7. Experimental Evaluation

Outputs:



Here we have pictorially represented the final output with a bar graph where we have visualized all the six method’s accuracy. The final output we got after training the data set is that we have best accuracy in the blue colored bar graph which is logistic regression model.



	Algorithm	Accuracy
0	logistic Regression Method	0.982456
0	Decision Tree Classifier Method	0.959064
0	Random forest Classifier Method	0.976608
0	Support Vector Classifier Method	0.976608
0	Gaussian Naive Bayes Method	0.959064
0	Stochastic gradient descent Classifier Method	0.976608
0	Gradient Boosting Classifier Method	0.970760
0	Gradient Boosting Classifier Method	0.970760

The final output we got in this project is that we got the best confidence level with logistic regression model with highest accuracy of 0.98. Here we have tested the final output using the test data which we had divided prior to the project.

The results obtained suggest the possibility of using intelligent computational tools based on data mining methods for the detection of breast cancer recurrence in patients who had previously undergone surgery. In this investigation, the Breast Cancer data set taken from was used, which was pre-processed for the validation of the data quality where it was necessary to perform data balancing, the algorithms implemented for the classification process.

8. Future Work

In future we will be dealing with predicting the breast cancer with a smaller number of attributes, here in this project we have used many attributes to attain the good confidence level but while collecting data it would not be possible to collect all the data for all attributes so in future we will be planning to attain best confidence with using less number of attributes.

9. References:

1.X. Jia, W. Meng, S. Li, Z. Tong and Y. Jia, "A rare case of intracystic Her-2 positive young breast cancer," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 2598-2602, doi: 10.1109/BIBM52615.2021.9669897.

2. V. E. Orel et al., "Computer-assisted Inductive Moderate Hyperthermia Planning For Breast Cancer Patients," 2020 IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), 2020, pp. 474-477, doi: 10.1109/ELNANO50318.2020.9088908.

3. B. Bılgıç, "Comparison of Breast Cancer and Skin Cancer Diagnoses Using Deep Learning Method," 2021 29th Signal Processing and Communications Applications Conference (SIU), 2021, pp. 1-4, doi: 10.1109/SIU53274.2021.9477992.

4. A. Easson, A. Pandya, J. Pasternak, N. Mohammed and A. Douplik, "Improving the patient cancer experience: Multispectral (White Light/Autofluorescence/Raman) Needle Endoscopy for cancer diagnostics in breast and thyroid," 2020 Photonics North (PN), 2020, pp. 1-2, doi: 10.1109/PN50013.2020.9166986.

5. Y. Amkrane, M. El Adoui and M. Benjelloun, "Towards Breast Cancer Response Prediction using Artificial Intelligence and Radiomics," 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), 2020, pp. 1-5, doi: 10.1109/CloudTech49835.2020.9365890.

6.M. Li, "Research on the Detection Method of Breast Cancer Deep Convolutional Neural Network Based on Computer Aid," 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2021, pp. 536-540, doi: 10.1109/IPEC51340.2021.9421338.

7. B.M Gayathri and C P Sumathi, An Automated technique using gaussian Naïve Bayes Classifier to Classify Breast Cancer. Internal Journal of Computer Applications 148(6): 16-21, August 2016.

8. Ahmad, L.G.; Eshlaghy, A.; Poorebrahimi, A.; Ebrahimi, M.; Razavi, A. Using three machine learning techniques for predicting breast cancer recurrence. J. Health Med. Inf. 2013, 4,

9. Sharma, G.N.; Dave, R.; Sanadya, J.; Sharma, P.; Sharma, K. Various types and management of breast cancer: An overview. J. Adv. Pharm. Technol. Res. 2010, 1, 109.

10. Turkki, R.; Byckhov, D.; Lundin, M.; Isola, J.; Nordling, S.; Kovanen, P.E.; Verrill, C.; von Smitten, K.; Joensuu, H.; Lundin, J.; et al. Breast cancer outcome prediction with tumour tissue images and machine learning. Breast Cancer Res. Treat. 2019, 177, 41–52.

11. Assiri A. S., Nazir S., Velastin S. A. Breast tumor classification using an ensemble machine learning method. Journal of Imaging . 2020;39(6) doi: 10.3390/jimaging6060039.

12. Mohammad Monirujjaman Khan,corresponding author 1 Tahia Tazin, 1 Mohammad Zunaid Hussain, 1 Monira Mostakim, 1 Taeefur Rehman, 1 Samender Singh, 2 Vaishali Gupta, 3 and Othman Alomeir 4.

13. AbienFred M.Agarap, ”On Breast Cancer Detection:An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset’,International Conference on Machine Learning, February 2–4, 2018, Phu Quoc Island, Viet Nam.

14. Wu, Y. T., Wei, J., Hadjiiski, L. M., Sahiner, B., Zhou, C., Ge, J., ... & Chan, H. P. (2007). Bilateral analysis based false positive reduction for computer aided mass detection. Medical physics, 34(8), 3334-3344.

15. Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics,7.

16. Aswathy, M. A., & Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: Present status and future possibilities. Informatics in Medicine Unlocked, 8, 74-79.

17. Kor H. Classification of Breast Cancer by Machine Learning Methods. SETSCI Conference Proceedings 2019;4:508-11.

18. Magesh G, Swarnalatha P. Analysis of breast cancer prediction and visualisation using machine learning models. International Journal of Cloud Computing 2022;11:43-60.

19. F.Bunea,”Honest Variable Selection in Linear and Logistic Regression models” International Journal of Statistics2 (2008).

20. Safavian S. R., Landgrebe D. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics.* 1991;21(3):660–674. doi: 10.1109/21.97458**.**