**BREAST**

**CANCER**

**CLASSIFICATION**



**BREAST CANCER CLASSIFICATION**

A Project Report

Submitted for the partial fulfilment for the award of the degree of

**Bachelor of Technology**

Submitted by

**Devishi Rastogi (2016141)**

**Dharana Sharma (2016142)**

**Pooja Sharma (2016222)**

**Prabhuta Chaudhary (2016223)**

Under the supervision of

**Dr. Amrita Upadhyay**

**Name of the Mentor**

**Name of Coordinator(s)**

Prof. Saurabh Mukherjee

Dr. Neelam Sharma

Dr. Pooja Asopa

Dr. Deepak Kumar

Dr. Sneha Asopa

Dr. Ashok Kumar

Dr. Rahul Vijay

Dr. Amrita Upadhyay



**Department of Mathematics & Computing**

**Banasthali Vidyapith**

**Banasthali - 304022**

**Session: 2022-23**

**Certificate**

Certified that **Devishi Rastogi (2016141), Dharana Sharma (2016142), Pooja Sharma (2016222), Prabhuta Chaudhary (2016223)** have carried out the project work titled “Breast Cancer Classification” from January 2023 to April 2023 for the award of the **Bachelor of Technology** from Banasthali Vidyapithunder my supervision. The thesis embodies result of original work and studies carried out by students themselves and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else.

**Dr. Amrita Upadhyay**

Designation: Assistant Professor

Place: Banasthali Vidyapith

Date: 12-Apr-2023

**ACKNOWLEDGEMENT**

First and foremost, we would like to thank our mentor, Amrita Upadhyay ma'am, for her constant support and guidance throughout the project. Her expertise and experience have been invaluable in shaping the direction of our research and helping us stay on track.

We would also like to thank Dr. Dipanwita Thakur ma'am for teaching us the essential concepts of Machine Learning and guiding us through the process of building a model. Her efforts and dedication have helped us gain a deeper understanding of the subject and develop the necessary skills to carry out the project.

We would also like to thank our Head Of Department Dr. C.K. Jha for the time he had took to review and offer constructive feedback.

Lastly, I would like to thank the university Banasthali Vidyapith for providing us with the opportunity to undertake this project and apply the knowledge gained in our coursework.

**ABSTRACT**

Over 1 crore people died from cancer last year, with breast cancer accounting for 22.6% of those deaths, making it one of the leading causes of death worldwide. In India, Breast Cancer accounts for 14.7% of all cancer cases and is the most prevalent malignancy in women. Numerous studies have been done on early Breast Cancer detection, which can aid with timely treatment initiation and lower mortality rates. Only roughly 86% of cases are appropriately diagnosed out of all those that are. Correctly classifying benign tumours can spare patients from unnecessary treatments. Thus, the correct diagnosis of Breast Cancer and classification of patients into malignant (cancerous) or benign (non-cancerous) groups is the subject of much research. Machine learning (ML) is widely acknowledged as the preferred methodology in Breast Cancer pattern classification and forecast modeling due to its distinct advantages in essential features discovery from complex Breast Cancer datasets. In this research we will compare several Machine Learning algorithms such as Random Forest, Logistic Regression, Decision Trees, Naive Bayes Classifier, Support Vector Machine and thus identify the algorithm which gives an effective and accurate breast cancer diagnosis and we’ll do so with the help of evaluation metrics.

**TABLE OF CONTENTS**

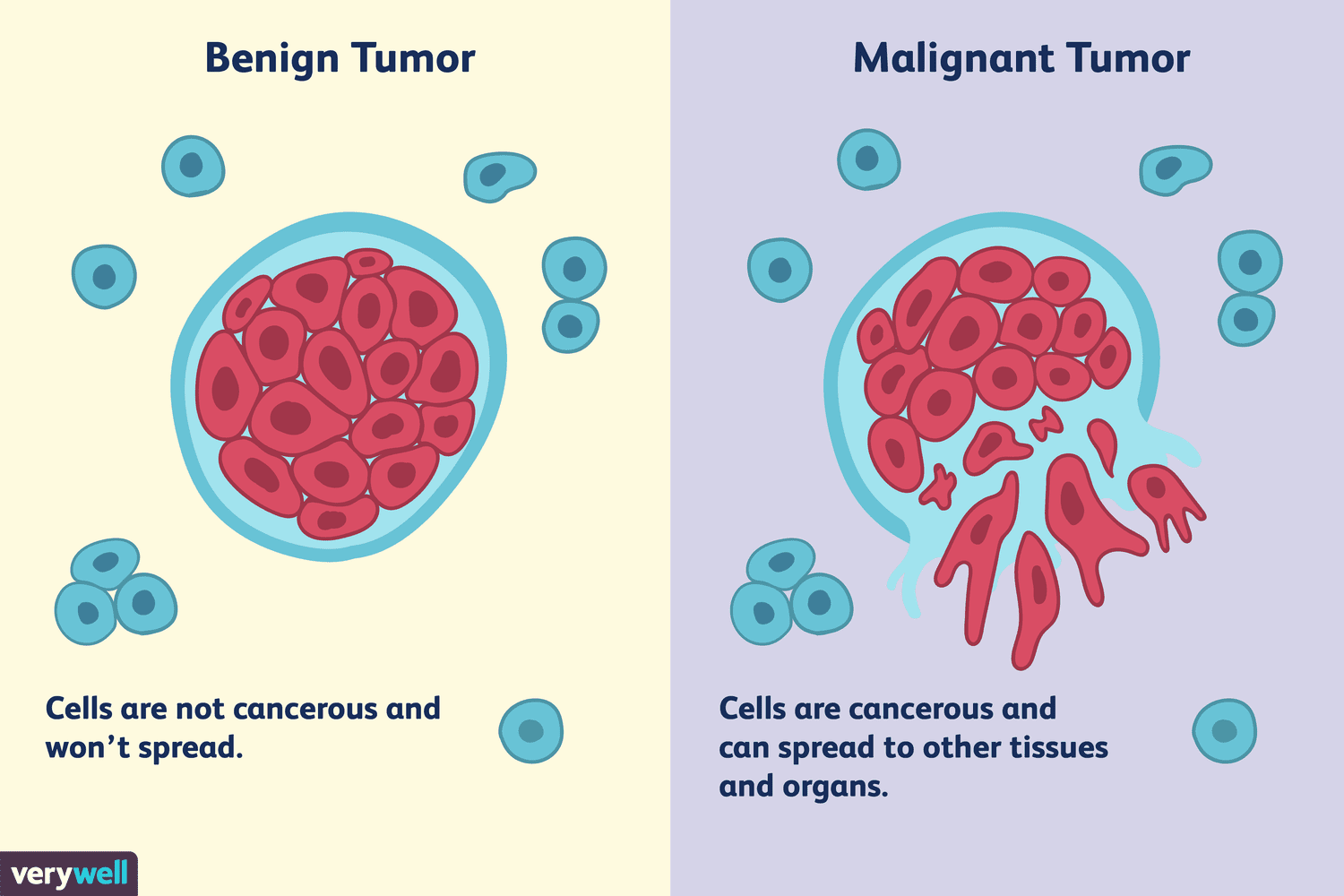
* INTRODUCTION
* LITERATURE REVIEW
* METHODOLOGY
* DISCUSSION
* RESULT
* CONCLUSION
* APPENDICES
* REFERENCES

**INTRODUCTION**

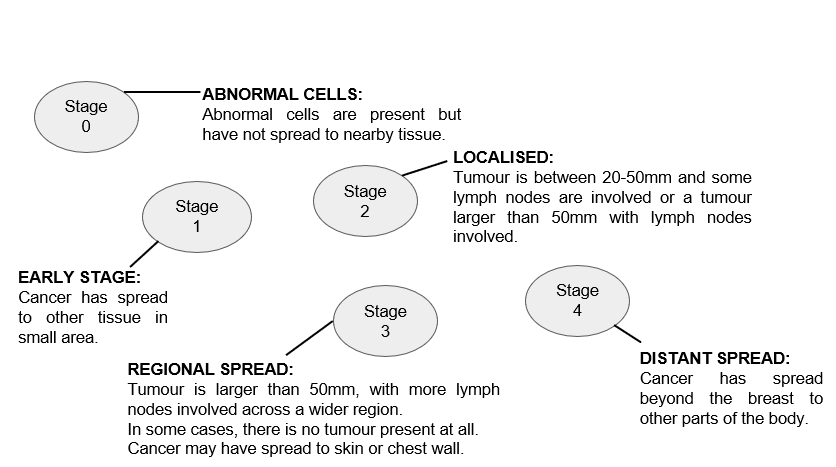
According to estimates from the World Health Organization (WHO), over 627,000 women died last year. 685 000 people worldwide died in 2020 as a result of breast cancer, which affected 2.3 million women. The most common cancer in the globe as of the end of 2020 was breast cancer, which had been diagnosed in 7.8 million women in the previous five years. Additionally, this organization projects that by 2030, there may be 2.7 million cases worldwide.

The stage of the malignancy at the time of diagnosis has a significant impact on the survival rate of Breast Cancer. The disease's late detection and challenging therapy are the key contributors to the low survival rate. Early diagnosis is necessary to provide patients with the correct care and to lower the mortality and morbidity rates. A High-performance diagnostic will be beneficial for a medical professional to support them in the diagnosis and adoption of appropriate treatment for various types of cancer. To determine the different forms of breast cancer tumours, a model based on a clinical dataset gathered via digital survey and hospitals will be used to the processed data. With the aid of machine learning algorithms, clinicians will be able to diagnose patients with the help of effective medications, thorough tests, and gracious treatment.

The study will categorize the main issues with breast cancer as follows: (Benign tumour and Malignant tumour). A benign tumour is a cancer that no longer spreads outside of the host or invades the tissue around it. Another type of cancer that can spread throughout the host's body or infiltrate the tissue around it is a malignant tumour. On rare occasions, benign cancers can also cause death, but generally speaking, they are no longer nearly as awful as the malignant cancers.



**Stages of Breast Cancer**



**Stage 0**

This stage addresses DCIS (Ductal Carcinoma Insitu) and other non-invasive breast cancers. It is a stage in which there is no sign of cancer cells developing on any area of the breast or infecting nearby healthy tissue.

**Stage 1**

Breast tumours in stage I are portrayed as spreading into the healthy breast tissue around them. The tumour is up to two centimetres in size at this point, and there are no lymph nodes implicated. In stage I breast cancer, a microscopic invasion is also possible. Microscopic invasion occurs when cancer cells first begin to penetrate tissue beyond the duct or lobule liner, however they do not extend beyond a diameter of 1 mm.

**Stage 2**

The breast cancer is still controlled at this stage because it has only spread to the nearby lymph nodes. However, it is still growing. Stage 2A and Stage 2B are the two firms that make up this stage. The tumour's size and whether the breast cancer has progressed to the lymph nodes influence the difference.

**Stage 3**

This stage denotes the spread of the cancer from the breast to nearby lymph nodes, the breast's skin, or the chest wall. Additionally, it is known as locally advanced breast cancer. A cancer's stage reveals its size and the extent of its dissemination. It aids the doctor in selecting the best course of treatment for the patient.

**Stage 4**

This type of stage IV invasive breast cancer has spread to other organs of the body, such as the lungs, distant lymph nodes, skin, bones, liver, or brain, in addition to the breast and nearby lymph nodes. Level IV breast cancer is referred to as "advanced" and "metastatic."

A form of biopsy process is **fine needle aspiration**. In fine needle aspiration, a little needle is introduced into a region of tissue or bodily fluid that appears aberrant. It aids in the diagnosis or excludes diseases like cancer. The most frequent site for a fine needle aspiration is a lump or swelling that is just under the skin.

A lump may be felt during a doctor's examination. Or it may be discovered on an imaging test such as:

* CT scan
* mammogram
* ultrasound

Imaging tests may also discover abnormal spots deeper inside the body.

Doctors may recommend fine needle aspiration for areas such as:

* cysts (fluid-filled lumps)
* nodules or masses (solid lumps)
* enlarged lymph nodes

Most fine needle aspirations are done on these areas:

* breast
* thyroid gland
* lymph nodes in the neck, groin, or armpit

After a fine needle aspiration, serious consequences are uncommon. At the location of the biopsy, there may be slight bleeding under the skin. A hematoma, which is a sore, swelling region, may develop as a result. Endoscopy poses a somewhat increased risk of complications from fine needle aspiration. For the majority of individuals, it is still very low.

It can take a long time, be tedious, and be subject to human mistake to manually classify cancer into benign and malignant forms. Based on the information gathered during the examination, the suggested approach may automatically classify various tumour kinds into risky (malignant) and safe (benign) categories. This computer performs this function using a machine learning technique. This new system's scope is as follows: Early disorder analysis could be performed without human mistake and classification errors might be considerably reduced. To locate and classify breast cancers, we'll employ machine learning.

### Aim and Objective of the Study

This study's primary goal is to use and compare various machine learning models to identify and categorize breast cancer into benign and malignant forms.

### Scope of Study

The theoretical and practical implications of this investigation are significant. It will act as a database for computer science researchers who would be interested in learning about the technological approach to treating breast cancer as well as for future researchers who might start a similar line of inquiry. Practically speaking, it will work as a resource for governmental and non-governmental groups, decision-makers in the field of breast cancer policy, and organizers of media campaigns. This study will make a significant contribution to the medical field's everyday clinical diagnosis method because using machine learning to classify which sort of disease will be more accurate and time-saving.

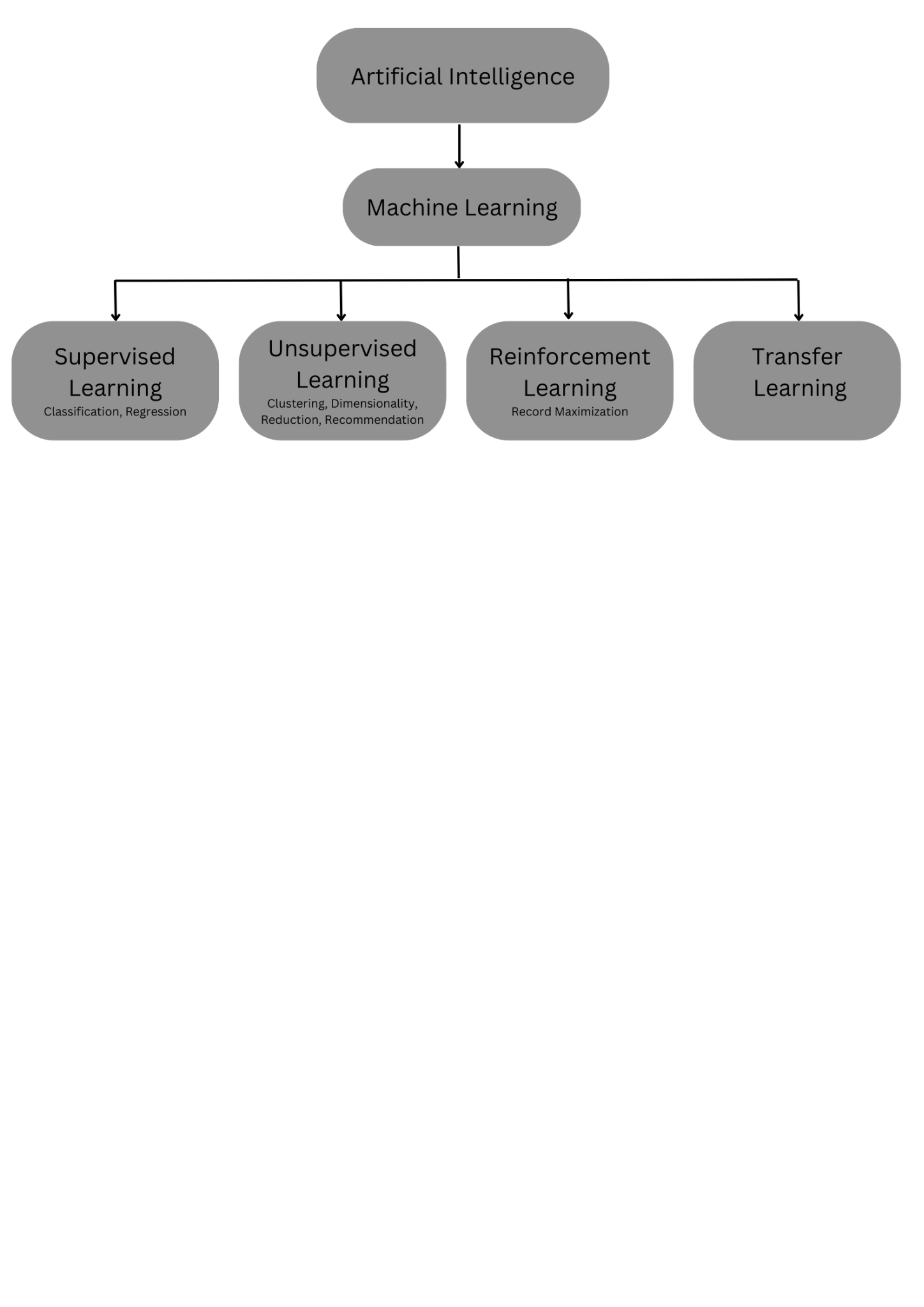
### 

### Definition of terms

**Breast cance**r: A malignant tumour that develops from a woman's breast tissue is known as breast cancer.

**Machine Learning:** It is a set of methods that allow computers to learn from data to make and improve predictions. It is a technique for altering or adjusting computer behaviour (whether these acts involve making predictions or making an attempt to direct a robot) in order to improve the accuracy of those actions, where accuracy is measured by how well the selected actions mirror the appropriate ones. It is further divided into following categories-

1. **Supervised Learning**: Supervised learning is a sort of machine learning in which the output is predicted by the machines using well-labelled training data that has been used to train the machines. The term "labelled data" refers to input data that has already been assigned the appropriate output. The method of supervised learning involves giving the machine learning model the right input data as well as the output data. Finding a mapping function to link the input variable (x) with the output variable is the goal of a supervised learning algorithm (y).
2. **Unsupervised Learning:** Unsupervised learning is a subcategory of machine learning in which models are trained using unlabelled datasets and are free to operate on the data without being checked by a human observer. From the provided data, models themselves uncover hidden patterns and insights.
3. **Reinforcement Learning:** Reinforcement learning is a type of machine learning method where an intelligent agent (computer program) interacts with the environment and learns to act within that. It is a feedback-based strategy where an agent learns how to act in a given environment by acting out scenarios and observing how those actions play out. The agent receives compliments for each positive activity, and is penalised or given negative feedback for each negative action. In contrast to supervised learning, the agent learns automatically utilising feedback without any labelled data.
4. **Transfer Learning:** Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.

**

## 

## 

## **Types of supervised Machine learning Algorithms:** Supervised learning can be further divided into two types of problems:

## 

1. **Regression:**If there is a correlation between the input and output variables, regression procedures are applied. It is a method for determining how independent traits or variables affect a dependent feature or outcome. It is a technique for machine learning predictive modelling, where an algorithm is used to forecast continuous outcomes.
2. **Classification:** In situations when the output variable is categorical, classification algorithms are applied. The primary objective of a classification algorithm is to determine the category of a given dataset, and these algorithms are primarily employed to forecast the results for categorical data.

The algorithm which implements the classification on a dataset is known as a **classifier**. There are two types of Classifications:

* **Binary Classifier:** If the classification problem has only two possible outcomes, then it is called Binary Classifier.
* **Multi-class Classifier:** If a classification problem has more than two outcomes, then it is called Multi-class Classifier.

**Data mining for classification work:** Large amounts of information are initially sorted in the data mining process before patterns are found and associations are created to carry out data analysis and address problems. Identification of a model that recognises and distinguishes data classes and concepts is the process of classification.

**Classifier**: A binary classifier is used to categorise patients into cancer patients and healthy patients. The majority of characteristics used in classification come from training dataset features. A proper and ideal training dataset is necessary for accurate classification.

Some classification algorithms:

* **Random Forest:** Instead of relying on one decision tree, the random forest takes the prediction from each tree and bases its prediction of the final output on the majority votes of predictions. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Higher accuracy and overfitting are prevented by the larger number of trees in the forest. 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 principle of the Random Forest algorithm can be comprehended by following the sequential steps enumerated below:

Step 1: First, start with the selection of random samples from a given dataset.

Step 2: Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3: In this step, voting will be performed for every predicted result.

Step 4: At last, select the most voted prediction result as the final prediction result.

* **Naive Bayes:** The Naive Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur.

Steps to implement NB Classifier are as follows:

Step 1: Load the data

Step 2: Split data into train and test

Step 2: Use the training data to estimate the probability of each feature given each class label using Bayes Theorem.

Step 3: Evaluate the performance of model on the testing data by calculating metrics such as accuracy, precision, recall.

Step 4: Visualize the model to gain insights into the model and its decision-making process.

Step 5: Use the Naive Bayes model to predict the class labels of new data based on its features.

* **Decision Trees:** 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. The given dataset's features are used to execute the test or make the decisions. It is a graphical depiction for obtaining all feasible answers to a choice or problem based on predetermined conditions. A decision tree only poses a question and divides the tree into subtrees according to the response (Yes/No). The procedure begins at the tree's root node to predict the class of the input dataset. This algorithm follows the branch and jumps to the following node by comparing the values of the root attribute with those of the record (real dataset) attribute. For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further. It continues the process until it reaches the leaf node of the tree.

Steps to implement Decision Tree Classifier are:

Step 1: Load the data

Step 2: Split data into train and test

Step 3: Train and validate the model. Evaluate the performance on the testing data by calculating metrics such as accuracy, precision, recall.

Step 4: Visualize the decision tree to gain insights into the model. This can help identify overfitting and interpret the results.

* **Logistic Regression:** In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. It has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

The methodology of the Logistic Regression algorithm may be comprehended by following the steps enumerated below -

Step 1: Load the data

Step 2: Logistic Regression measures the relationship between the dependent variable and the independent variables, by estimating probabilities using its underlying logistic function.

Step 3: These probabilities must then be transformed into binary values in order to actually make a prediction using the sigmoid function.

Step 4: The Sigmoid-Function takes any real-valued number and map it into a value between the range of 0 and 1.

Step 5: This values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier.

* **Support vector Machines:** A straightforward approach for classification and regression tasks is the Support Vector Machine. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors are used to describe these severe situations. It can deliver great precision quickly and with little computational effort. We are utilising the Linear SVC due of the substantial number of features. It found out that lowering overfitting and raising the quality of the prediction when the regularisation parameter C=0.0001 is used.

We can implement the SVM model by following the below steps:

Step 1: Load the data

Step 2: Split data into train and test

Step 3: First, it finds decision boundaries that correctly classify the training dataset.

Step 4: Pick the decision boundary which has maximum distance from the nearest points (supported vectors) of these two classes as the best one.

Step 5: Predict values using the SVM algorithm model

Step 6: Calculate the accuracy and precision.

**LITERATURE REVIEW**

First, we looked into a number of papers and discussions and debates on the current hot topics in healthcare machine learning. Breast cancer is a malignant tumour that develops in the breast tissue. It is a cancer of the glandular breast tissue in which the tissues are damaged as a result of an excessive growth of cancer cells, which then causes the destruction of surrounding tissues and other organs through blood circulation (Russel et al., 2013). This literature review focuses on the numerous methods and algorithms that have been used to predict breast cancer, with genetic algorithms serving as the classifier for the provided dataset.

We also made references to related works previously done on breast cancer diagnosis by researchers using different machine learning for a more comprehensive discussion and its long-term complications.

**David A. Omondiagbe, Shanmugam Veeramani, Amandeep S. Sidhu** used the Wisconsin Diagnostic Breast Cancer Database to examine the effectiveness of Support Vector Machine, Artificial Neural Network, and Naive Bayes.

By combining these machine learning approaches with feature selection and feature extraction techniques, the best results were obtained for the (WDBC) Dataset. According to the simulation findings, because of its lengthy calculation time, SVM-LDA was preferred above all the other approaches.

**S. Vasundhara, B.V. Kiranmayee, and Chalumuru Suresh** ***et al*** proposed employing different machine learning methods to automatically classify mammography pictures as benign, malignant, or normal. Support Vector Machines, Convolutional Neural Networks, and Random Forest were compared and contrasted. The simulation results showed that CNN produces intuitive classification of digital mammograms using filtering and morphological procedures, making it the best classifier.

**Anji Reddy Vaka, Badal Soni and Sudheer Reddy K *et al*** By utilising machine learning techniques such as Naive Bayes classifier, SVM classifier, bi-clustering Ada Boost techniques, RCNN classifier, and Bidirectional Recurrent Neural Networks (HA-BiRNN), they revealed a unique method to diagnose breast cancer. The proposed methodology (Deep Neural Network with Support Value) and machine learning techniques were compared, and the simulated results showed that the DNN algorithm was superior in terms of performance, efficiency, and image quality - factors that are critical in today's medical systems - while the other techniques didn't work as expected.

**Kalyani Wadkar, Prashant Pathak, and Nikhil Wagh *et al*** In order to better handle the dataset, they conducted a comparison study between ANN and SVM and included a number of classifiers, including CNN, KNN, and Inception V3. According to the experimental findings and performance analysis, ANN performed more efficiently than SVM, making it a superior classifier.

**Sivapriya J, Aravind Kumar V, Siddarth Sai S, and Sriram S *et al*** compared SVM, Logistic Regression, Naive Bayes, and Random Forest. The comparison is carried out using the Wisconsin Breast cancer dataset. The Random Forest method had the best accuracy (99.76%) and lowest mistake rate, according to the results of the experiments that were conducted. All of the studies were run in a virtual environment using the ANACONDA Data Science Platform.

**Muhammet Fatih Ak *et al*** used the dataset from Dr. William H. Walberg of the University of Wisconsin Hospital. This dataset was subjected to data visualisation and machine learning methods such as logistic regression, k-nearest neighbours, SVM, naive Bayes, decision tree, random forest, and rotation forest. These machine learning methods and visualisation were implemented using R, Minitab, and Python. All the techniques were compared in a comparative analysis. The best classification accuracy (98.1%) was obtained using the logistic regression model with all features included, and the suggested method demonstrated improved accuracy performances.

**Arpita Joshi and Dr. Ashish Mehta *et al*** compared KNN, SVM, Random Forest, and Decision Tree (Recursive Partitioning and Conditional Inference Tree) classification results using Wisconsin Breast Cancer dataset from UCI repository. KNN was the top classifier, according to the simulation results, followed by SVM, Random Forest, and Decision Tree.

**Abdullah-Al Nahid and Yinan Kong** ***et al*** presented CNN method for breast image classification. Conventional Neural Networks (NN), Random Forest (RF) algorithm, Support Vector Machines (SVM), and Bayesian methods were used to detect breast cancer by image classification. Since Convolutional Neural Network (CNN) techniques often extract the features globally with the help of kernels and these Global Features were used for image classification, the CNN method showed to be the best for the detection of breast cancer.

**K.Anastraj, Dr.T.Chakravarthy, K.Sriram** ***et al*** They conducted a comparison of the machine learning algorithms: back propagation network, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), and Support Vector Machine (SVM) on the Wisconsin Breast Cancer (original) datasets. For feature extraction and analysis of benign and malignant tumours, ALEXNET was utilised in conjunction with deep and convolutional neural networks. According to the simulation results, SVM is the best strategy and has produced superior outcomes (94%).

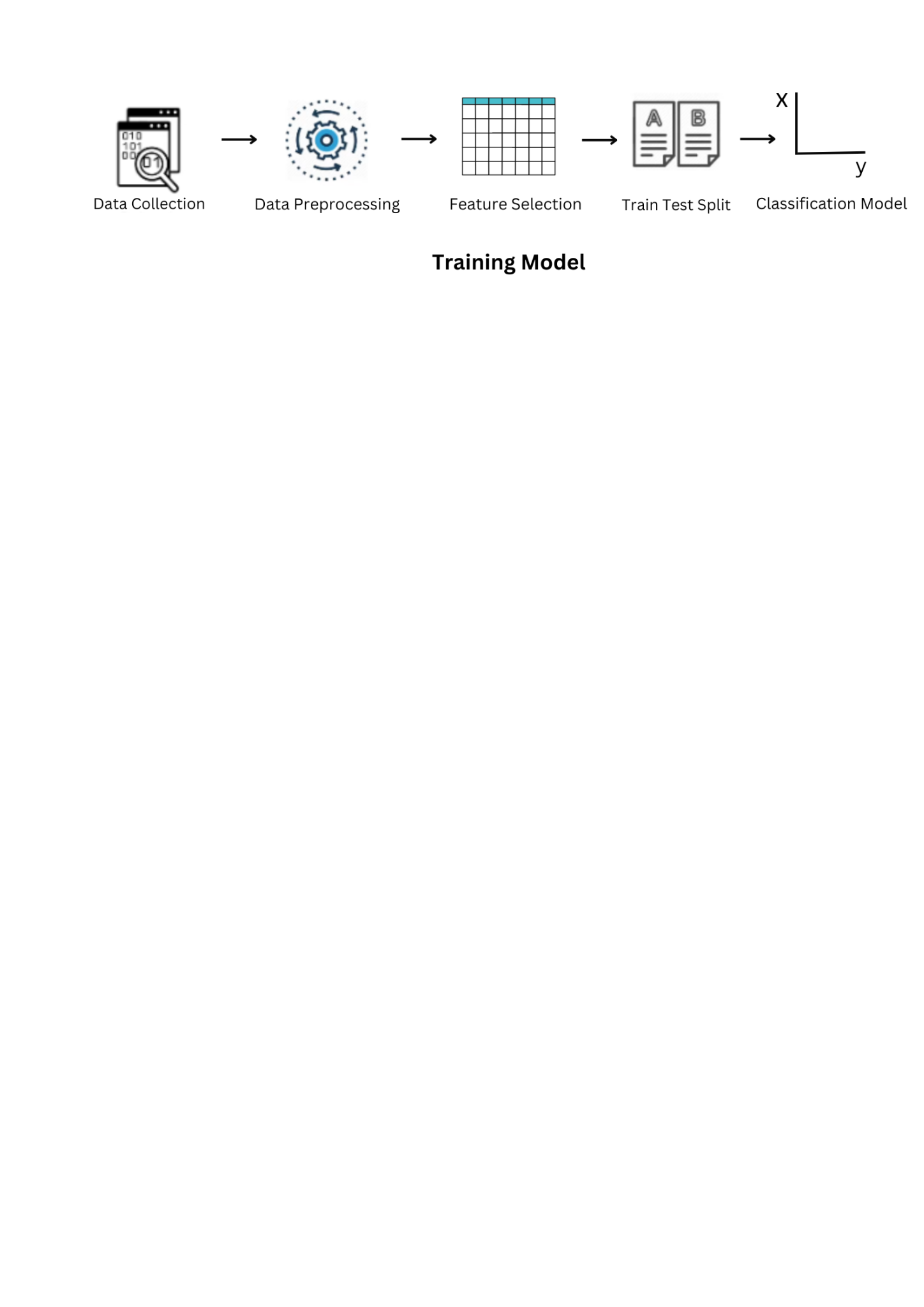
**Saeed S, Jhanjhi NZ, Naqvi M, Humyun M, Ahmad M. *et al*** They discovered that the optimum measure of roughness for regular elements is Fractal Dimension (FD). The breast is one of the ideal sites to use FD since breast lumps are irregular and can range from benign to cancerous. On the other hand, a brand-new categorising method is the SVM. They used two methods—FA: SVM and Box Count Method (BCM)—in separate operations that were successful in their respective industries. To extract features, the BCM is employed. The 42-image input dataset's difficulty is evaluated by the extracted feature FD. The SVM classifier is then utilised to process the generated FD and categorise benign and cancerous cells. They are 98.13% accurate on average.

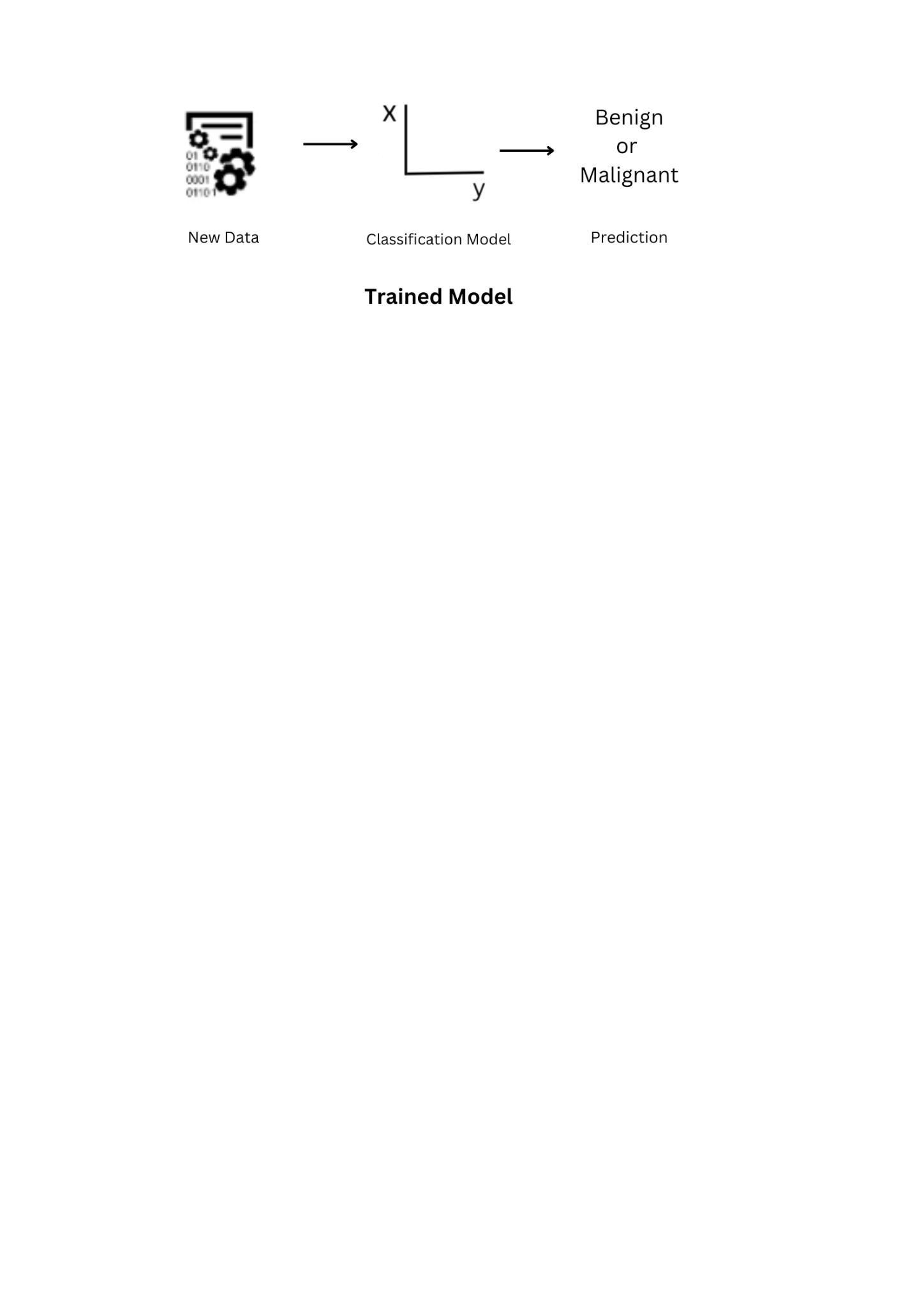
**Monica Tiwari, Rashi Bharuka, Praditi Shah, and Reena Lokare *et al*** Using machine learning techniques like Logistic Regression, Random Forest, K-Nearest Neighbour, Decision Tree, Support Vector Machine, and Naive Bayes Classifier as well as deep learning techniques like Artificial Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks, presented a novel method to detect breast cancer. The comparison of machine learning and deep learning methods revealed that the accuracy of the CNN and ANN models (99.3% and 97.3%, respectively) was higher than that of the machine learning models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S. No.** | **Author** | **Year** | **Title of Paper** | **Findings** |
| **1.** | **David A. Omondiagbe, Shanmugam Veeramani, Amandeep S. Sidhu** | **2019** | **Machine Learning Classification Techniques for Breast Cancer Diagnosis** | **SVM-LDA was preferred above all the other approaches** |
| **2.** | **S. Vasundhara, B.V. Kiranmayee, and Chalumuru Suresh** | **2019** | **Machine Learning Approach for Breast Cancer Prediction** | **CNN produces intuitive classification of digital mammograms** |
| **3.** | **Anji Reddy Vaka, Badal Soni and Sudheer Reddy K** | **2020** | **Breast Cancer Detection by Leveraging Machine Learning** | **DNN algorithm was superior** |
| **4.** | **Kalyani Wadkar, Prashant Pathak, and Nikhil Wagh** | **2019** | **Breast Cancer Detection Using ANN Network and Performance Analysis with SVM** | **ANN performed more efficiently than SVM** |
| **5.** | **Sivapriya J, Aravind Kumar V, Siddarth Sai S, and Sriram S** | **2019** | **Breast Cancer Prediction using Machine Learning** | **Random Forest method had the best accuracy rate** |
| **6.** | **Muhammet Fatih Ak** | **2019** | **A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications** | **Logistic Regression model had the best accuracy rate** |
| **7.** | **Arpita Joshi and Dr. Ashish Mehta** | **2017** | **Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer** | **KNN was the top classifier** |
| **8.** | **Abdullah-Al Nahid and Yinan Kong** | **2017** | **Involvement of Machine Learning for Breast Cancer Image Classification: Asurvey** | **CNN method showed to be the best for the detection of breast cancer** |
| **9.** | **K.Anastraj, Dr.T.Chakravarthy, K.Sriram** | **2019** | **Breast Cancer detection either Benign or Malignant Tumour using Deep CNN With ML Techniques** | **SVM is the best strategy** |
| **10.** | **Monica Tiwari, Rashi Bharuka, Praditi Shah, and Reena Lokare** | **2020** | **Breast Cancer Prediction using Deep learning and Machine Learning Techniques** | **SVM had 98.13% accuracy rate** |
| **11.** | **Saeed S, Jhanjhi NZ, Naqvi M, Humyun M, Ahmad M.** | **2022** | **Optimized breast cancer pre-mature detection method with computational segmentation** | **SVM had 98.13% accuracy rate** |

**METHODOLOGY**

The dataset is subjected to a rigorous examination to identify any null values and purged accordingly. In the event that the classification data is encoded as string values, they are transformed into integer format. Once the data is sanitized, an in-depth analysis is performed to extract its characteristic features. Subsequently, the pristine data is bifurcated into two distinct sets, namely train and test datasets. Upon successful partitioning, the classifier model is executed, and the accuracy of the test dataset is predicted.





1. **Data Collection:** The fundamental process in the machine learning pipeline is gathering data for training the ML model. The gathered information must be pertinent to the problem statement, should not have any blank or missing values in the prediction columns, and should not be biased. Target data is furthermore required for supervised learning, which can include speaking with subject-matter authorities and devoting a large amount of time. For machine learning algorithms to be effective, a large amount of noise-free data is necessary. Because processing costs rise as dataset sizes do, it is frequently impossible to pinpoint the precise point at which data volume and computing cost are optimally balanced.
2. **Data pre-processing:** Data pre-processing is a data mining technique used to turn the raw data into a format that is both practical and effective. Data cleaning is necessary before data analysis because it enables you to spot patterns in the data.

**Steps Involved in Data Pre-processing:**

* **Data Cleaning:** There may be a lot of useless information and gaps in the data. Data cleaning is completed to handle this portion. It entails dealing with erroneous data, noisy data, etc.
  + 1. **Missing Data:** Values missing in the dataset gives rise to the missing data.
    2. **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.
* **Data Transformation:** This method is used to change the data into formats that are appropriate for the mining process.
* **Data Reduction:** Data mining technique is used to manage vast amounts of data. Analysis in these situations grew more difficult when working with large amounts of data. We use a data reduction approach to get rid of this. It attempts to lower the price of data storage and analysis while increasing storage efficiency.

1. **Feature selection:** Finding the best set of features that makes it possible to create effective models is the aim of feature selection strategies in machine learning. We looked at prospective traits that might be helpful in cancer early detection. It requires identifying the most advantageous traits for the current problem.
2. **Train-test split:** The train test split method is used to gauge how well machine learning algorithms perform when making predictions on data that was not used to train the model. The dataset is split in half, 80:20, into a training set and a test set for testing.
3. **Algorithm:** Machine learning uses a variety of strategies for binary classification. This work should be able to prepare you to choose the best method (or methods) given the dataset because this is exactly what is required—understanding the core concepts underlying each algorithm and real-world applications.

The most common are:

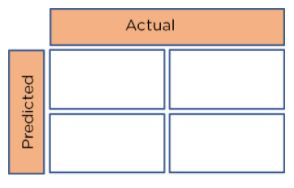
* + Logistic Regression
  + Support Vector Machines
  + Naive Bayes
  + Decision Trees
  + Random Forest

1. **Parameter and Model Selection:** Numerous algorithms involve parameters that must be manually set or tested to obtain the correct values.
2. **Training Given:** Simply using computer resources to build a model of the data in order to predict the outcomes on new data given the dataset, algorithm, and parameters should constitute training.
3. **Evaluation:** A system needs to be checked for accuracy and other assessment metrics prior to deployment.

**Evaluation Metrics:**

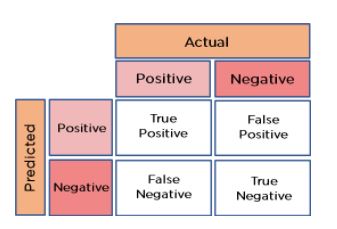
1. **Confusion Matrix**

A confusion matrix presents a table layout of the different outcomes of the prediction and results of a classification problem and helps visualise its outcomes. It plots a table of all the predicted and actual values of a classifier.



Basic layout of a Confusion Matrix

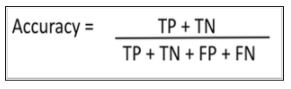
We can obtain four different combinations from the predicted and actual values of a classifier:



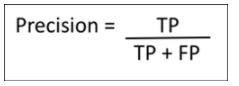
* **True Positive:** The number of times our actual positive values are equal to the predicted positive. You predicted a positive value, and it is correct.
* **False Positive:** The number of times our model wrongly predicts negative values as positives. You predicted a negative value, and it is actually positive.
* **True Negative:** The number of times our actual negative values are equal to predicted negative values. You predicted a negative value, and it is actually negative.
* **False Negative:** The number of times our model wrongly predicts negative values as positives. You predicted a negative value, and it is actually positive.

Just from looking at the matrix, the performance of our model is not very clear. To find how accurate our model is, we use the following metrics:

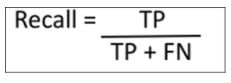
1. **Accuracy**   
   The accuracy is used to find the portion of correctly classified values. It tells us how often our classifier is right. It is the sum of all true values divided by total values.



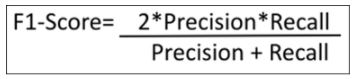
1. **Precision**   
   Precision is used to calculate the model's ability to classify positive values correctly. It is the true positives divided by the total number of predicted positive values.



1. **Recall**It is used to calculate the model's ability to predict positive values. "How often does the model predict the correct positive values?". It is the true positives divided by the total number of actual positive values.



1. **F1-Score**It is the harmonic mean of Recall and Precision. It is useful when you need to take both Precision and Recall into account.



1. **ROC-Curve**

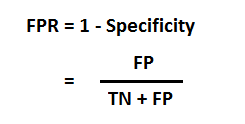
ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

* True Positive Rate
* False Positive Rate

**True Positive Rate** (**TPR**) is a synonym for recall and is therefore defined as follows:



**False Positive Rate** (**FPR**) is defined as follows:



It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

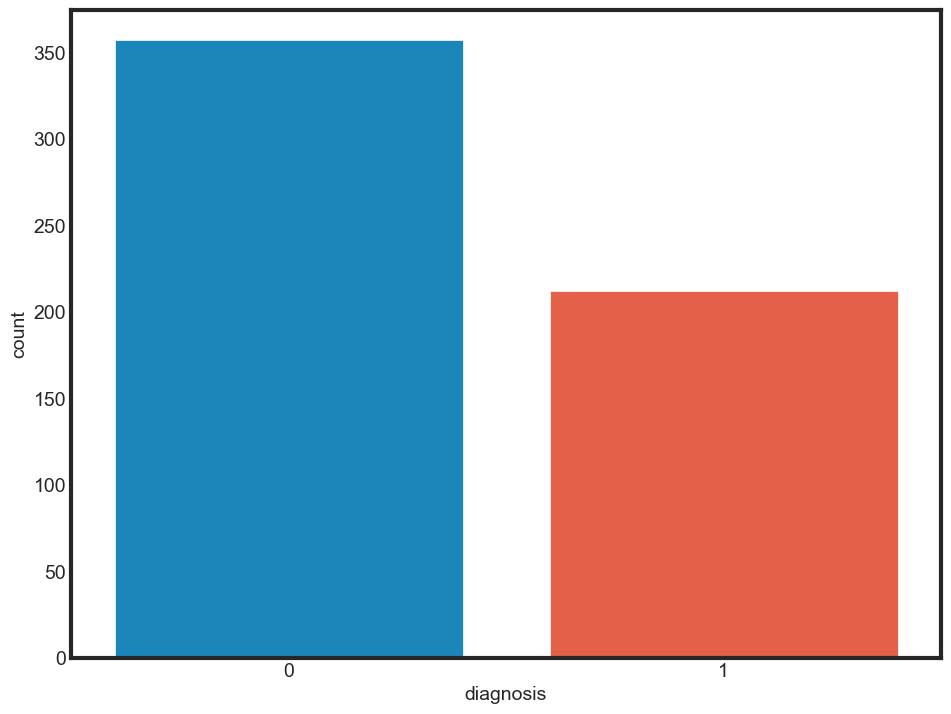
**DISCUSSION**

## **Unimodal Data Visualizations**

One of the main goals of visualizing the data here is to observe which features are most helpful in predicting malignant or benign cancer. The other is to see general trends that may aid us in model selection and hyper parameter selection.

We have applied 3 techniques that we can use to understand each attribute of our dataset independently.

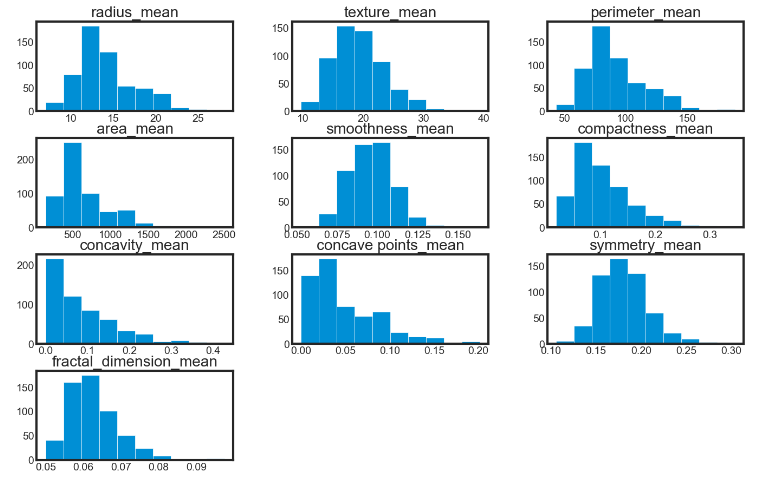
* Histograms.
* Density Plots.
* Box and Whisker Plots.



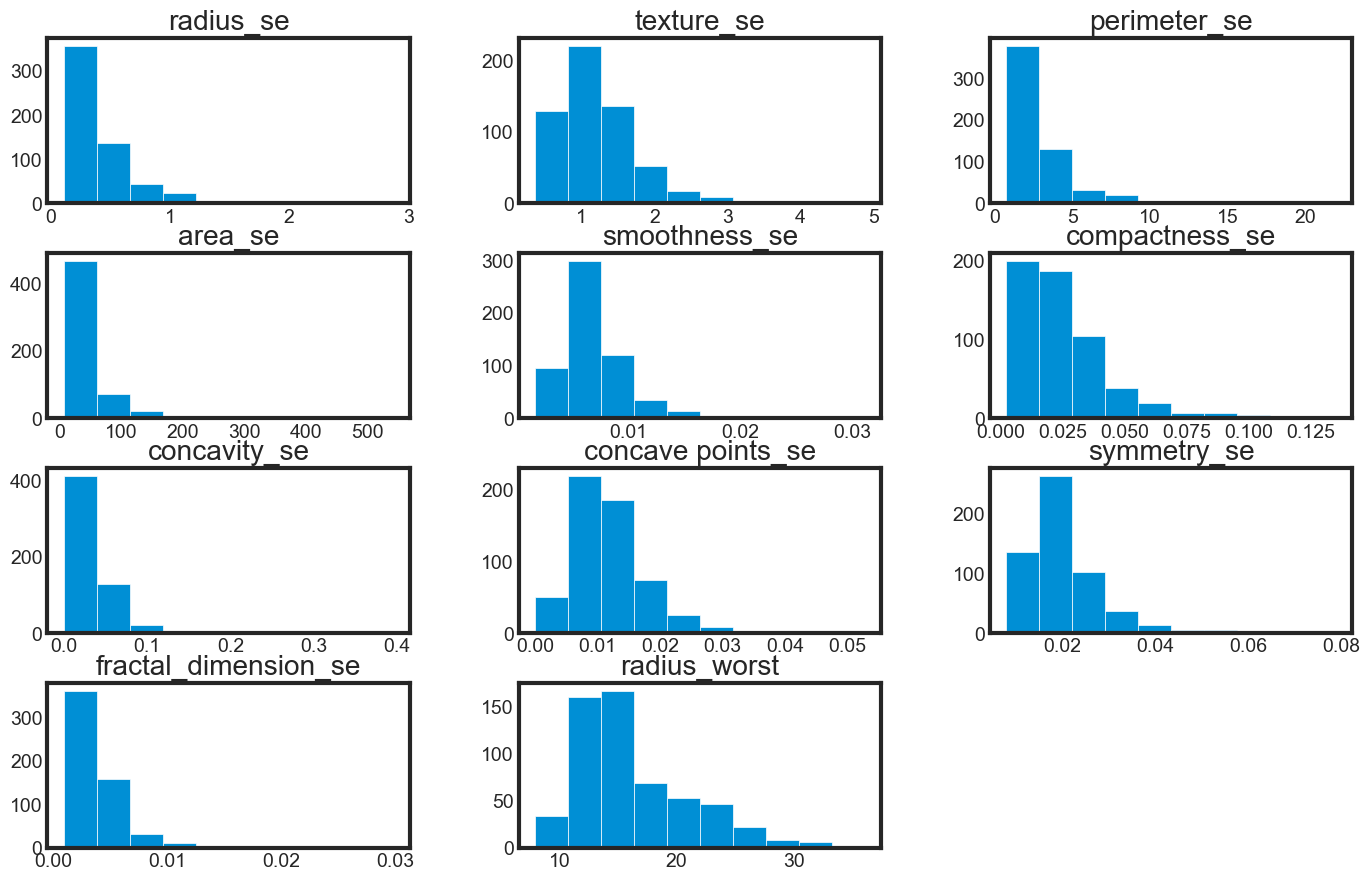
## **Visualizing distribution of data via histograms**

Histograms are commonly used to visualize numerical variables. A histogram is similar to a bar graph after the values of the variable are grouped (binned) into a finite number of intervals (bins). Histograms group data into bins and provide you a count of the number of observations in each bin.

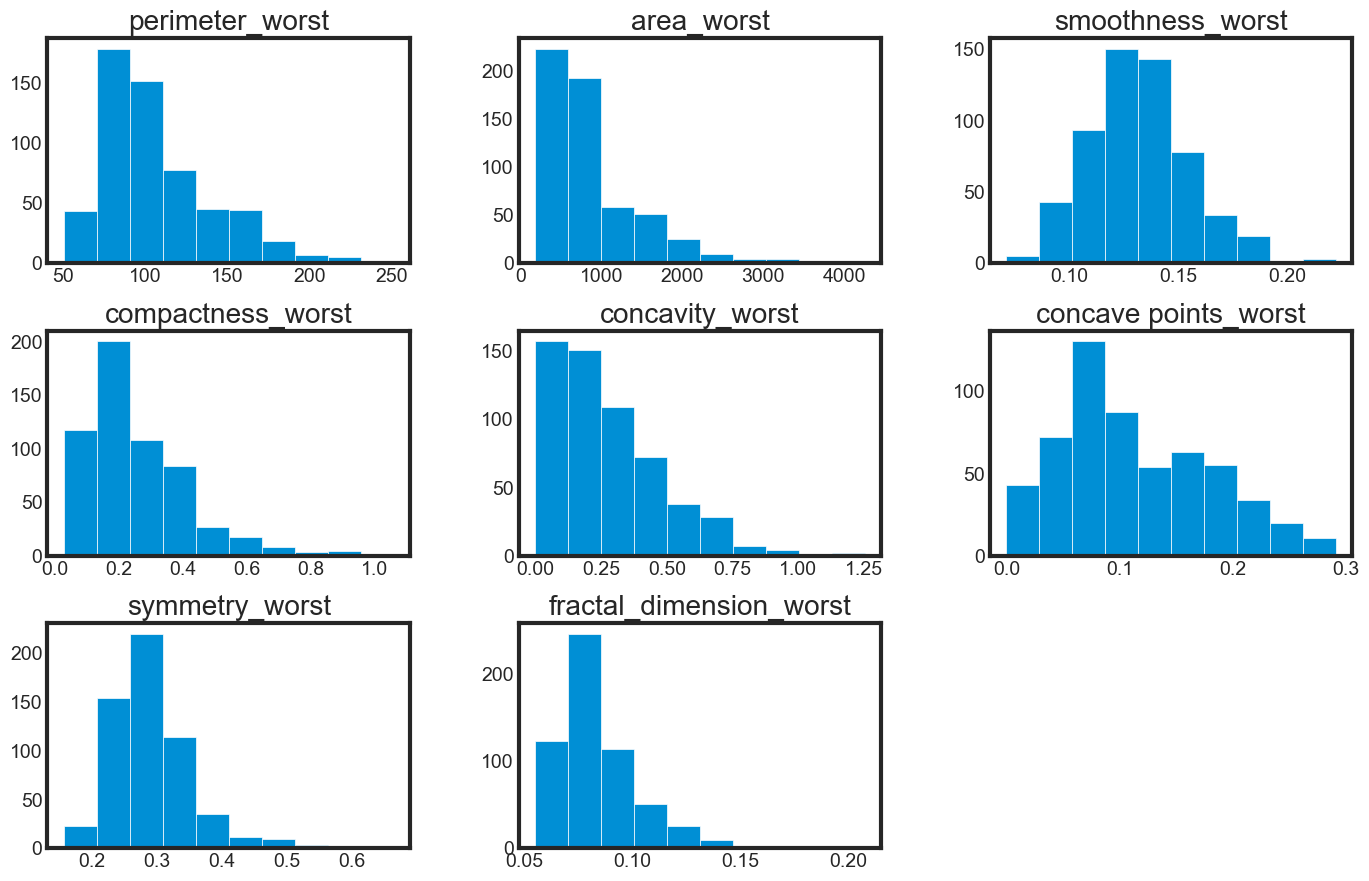
Histogram of  \_mean suffix

****

Histogram of \_se suffix



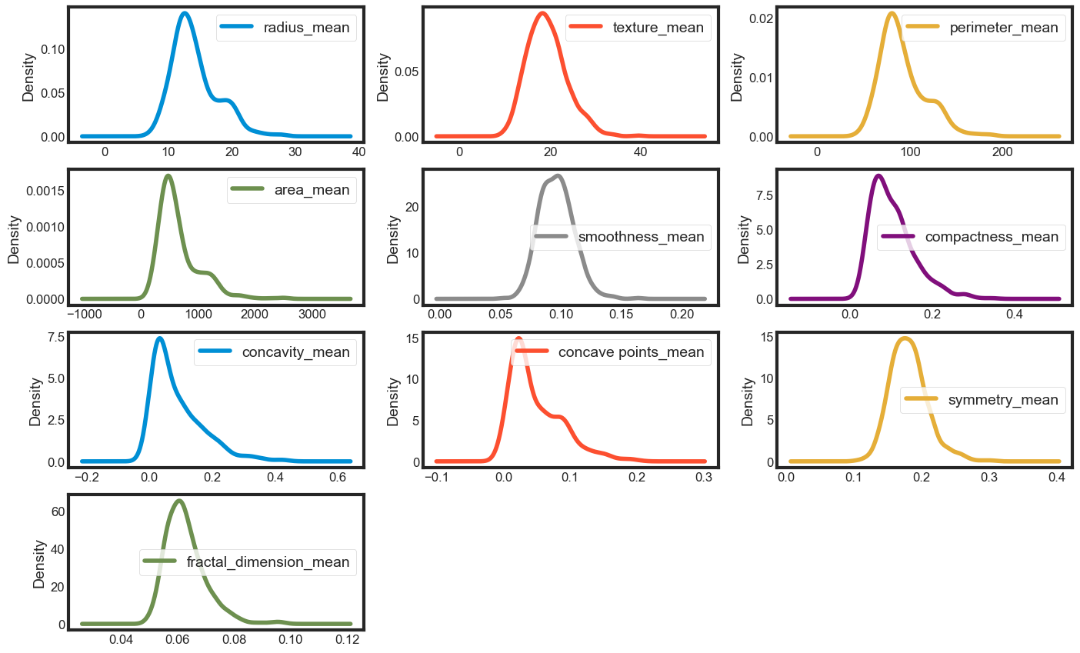
Histogram of \_worst suffix



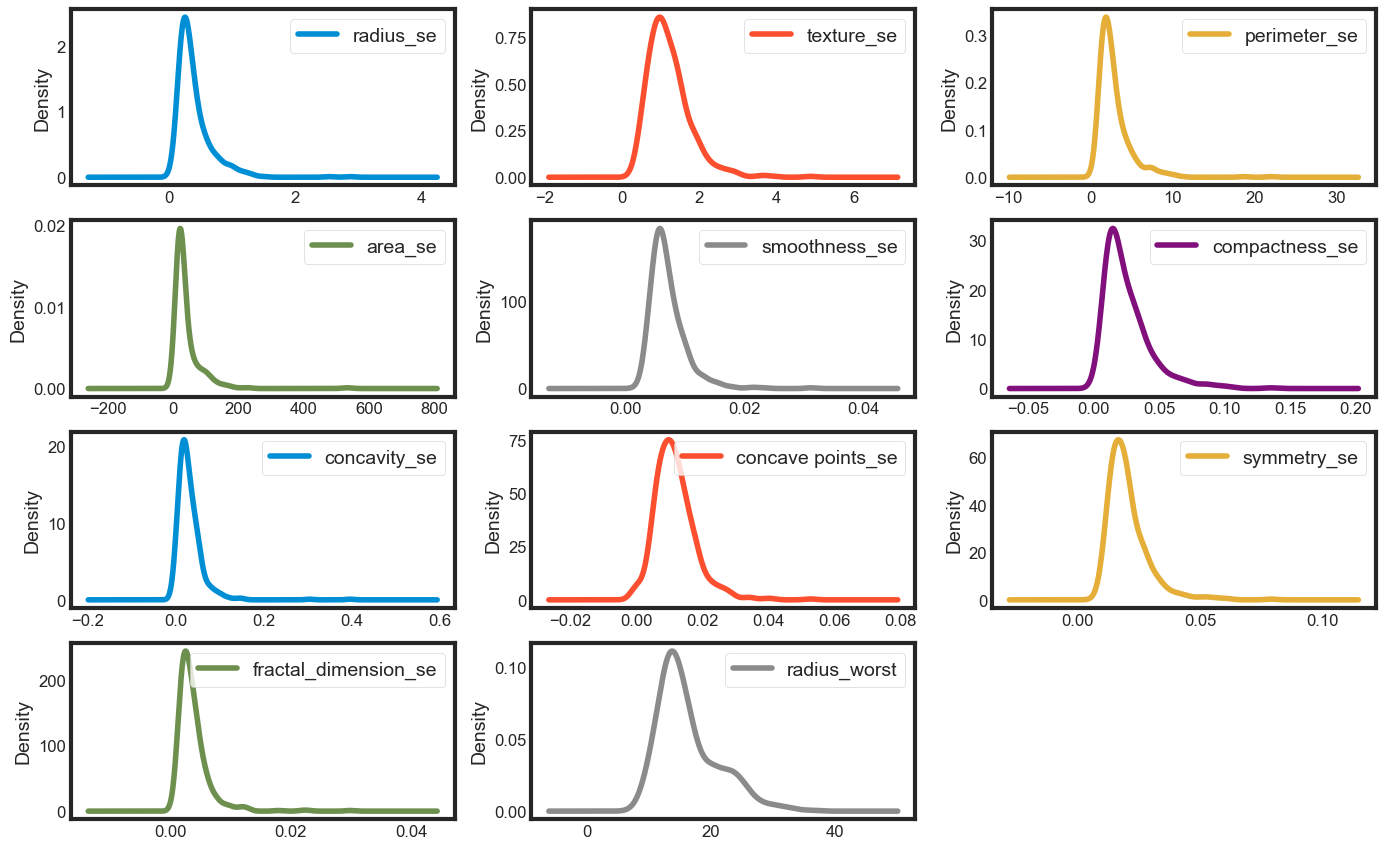
We can see that perhaps the attributes **concavity**, and **concavity\_point** may have an exponential distribution. We can also see that perhaps the texture and smooth and symmetry attributes may have a Gaussian or nearly Gaussian distribution. This is interesting because many machine learning techniques assume a Gaussian univariate distribution on the input variables.

* **Visualizing distribution of data via density plots**

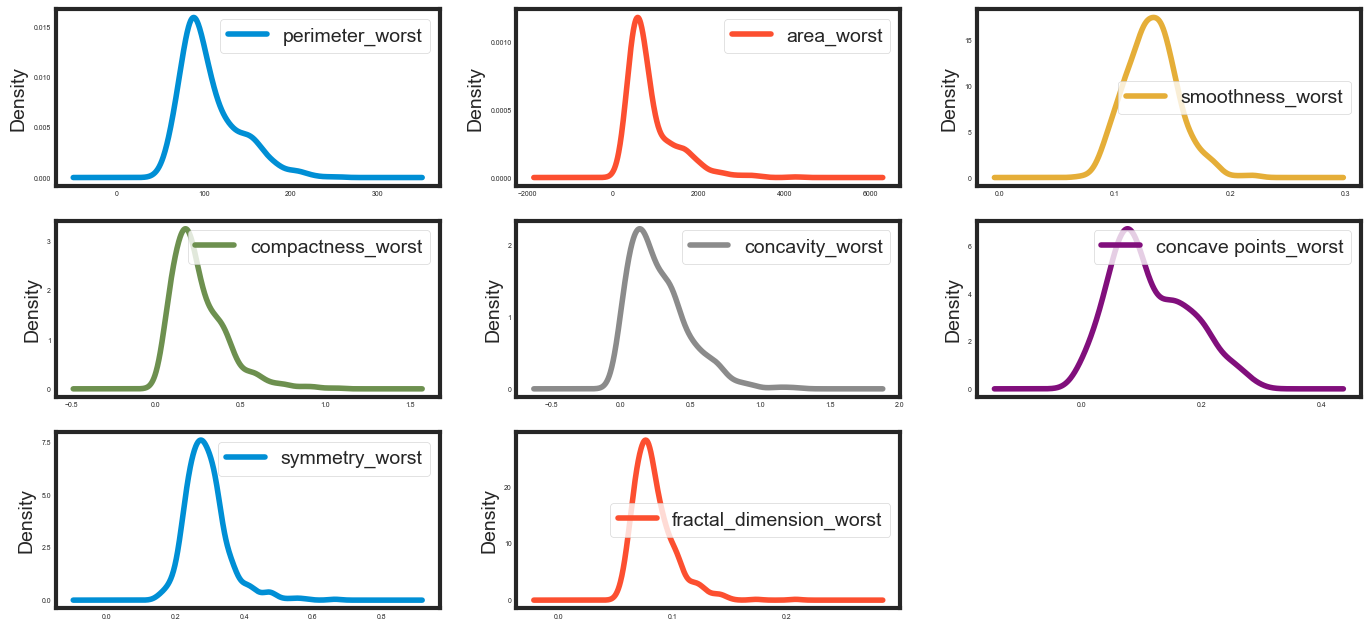
Density plots of \_mean suffix

****

Density plots of \_se suffix

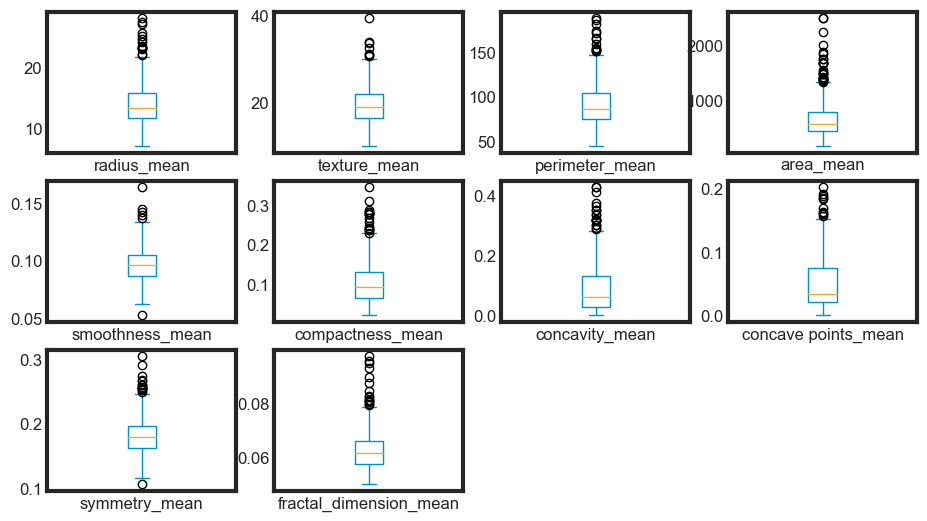
****

Density plots of \_worst suffix

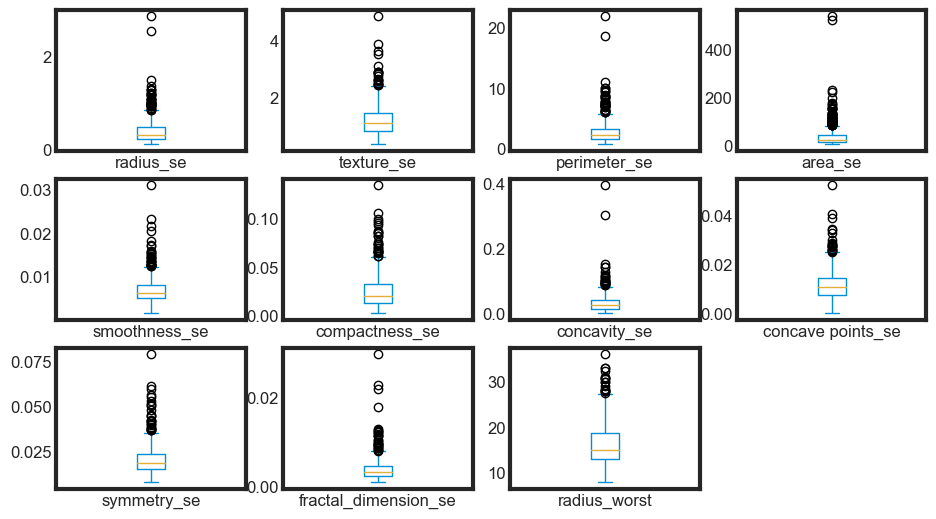


## **Visualizing distribution of data via box plots**

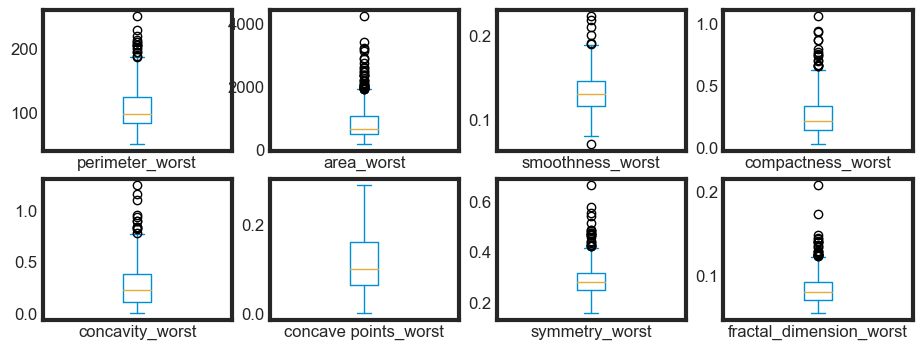
Box plot of \_mean suffix



Box plot of \_se suffix



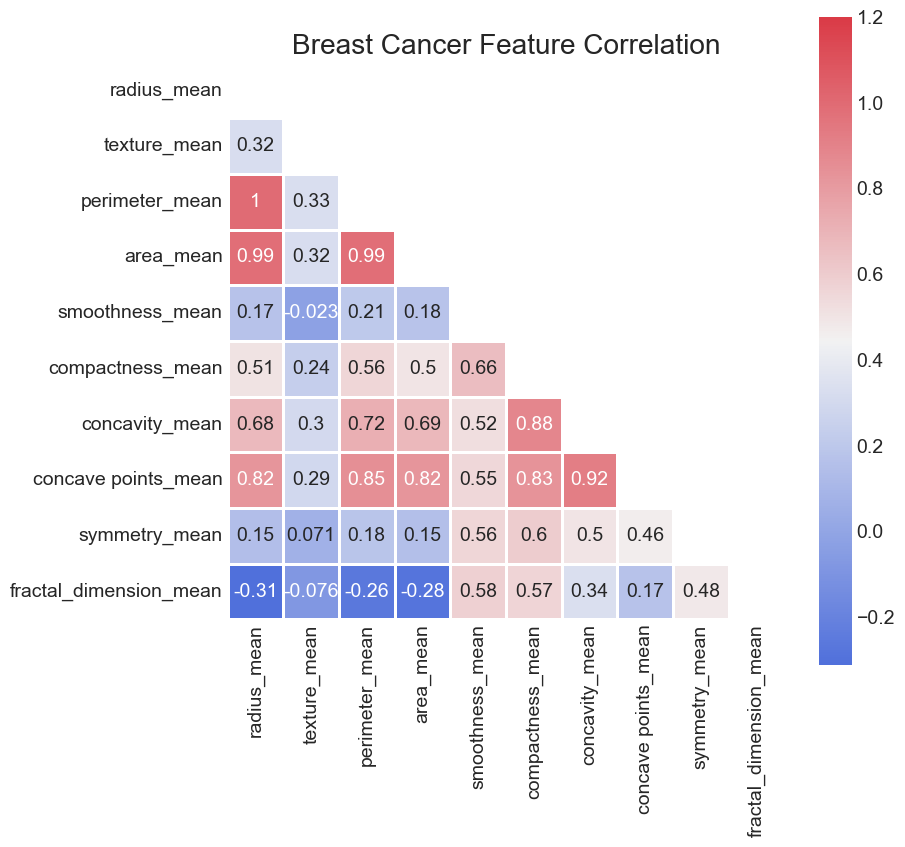
Box plot of \_worst suffix



## **Multimodal Data Visualizations**

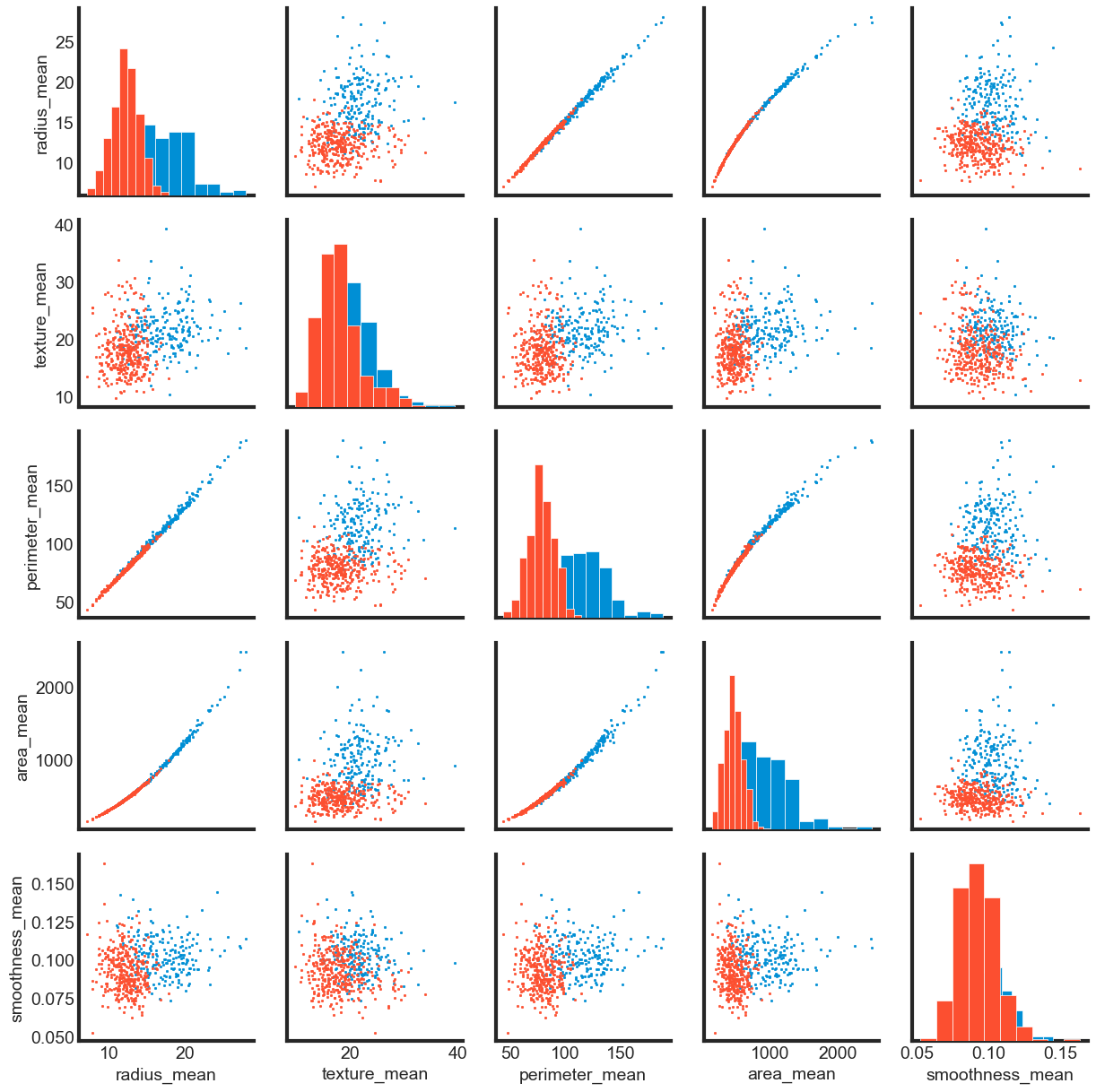
* Correlation matrix
* Scatter plots

**Correlation matrix**

****

We can see strong positive relationship exists with mean values parameters between 1 to 0.75.

* The mean area of the tissue nucleus has a strong positive correlation with mean values of radius and parameter;
* Some parameters are moderately positive correlated (r between 0.5-0.75) are concavity and area, concavity and perimeter etc
* Likewise, we see some strong negative correlation between fractal\_dimension with radius, texture, parameter mean values.

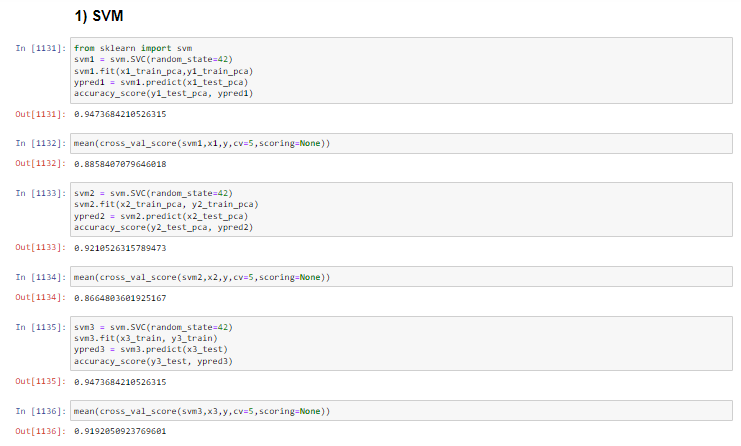


The mean values of key parameters such as cell radius, perimeter, area, compactness, concavity, and concave points can be leveraged to classify cancer. Notably, higher values of these parameters are often indicative of malignant tumours, thus providing a strong correlation. Conversely, the mean values of other parameters, such as texture, smoothness, symmetry, and fractal dimension, do not exhibit a significant preference for one diagnosis over the other.

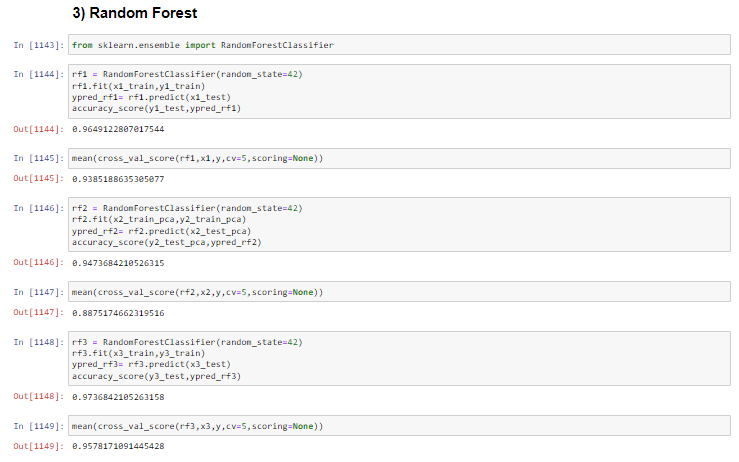
Upon inspecting the histograms, no noticeable outliers of significant magnitude are present, and therefore, no further clean-up efforts are deemed necessary.

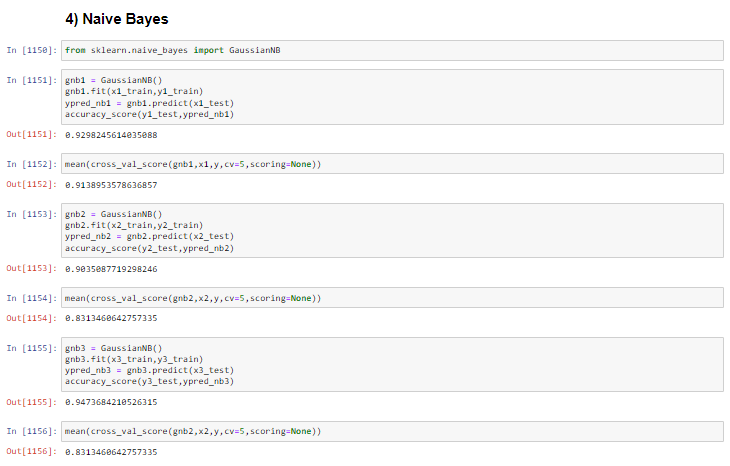
**Model Training and Evaluation**

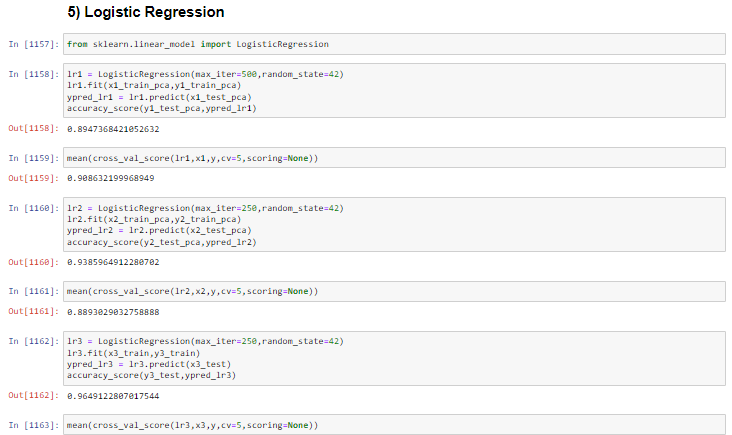
1. **Module 1 – Splitting Dataset into \_mean, \_se and \_worst values**

****

****

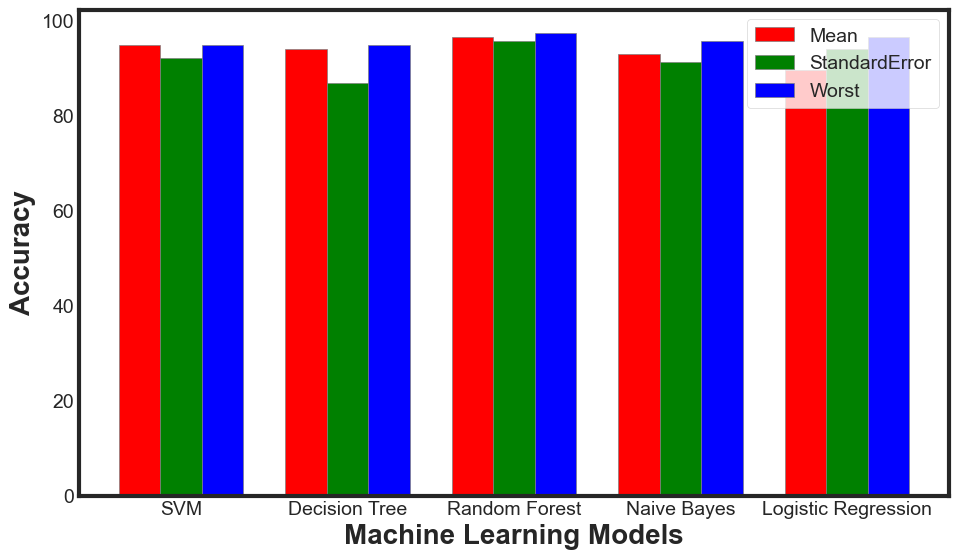
****

****

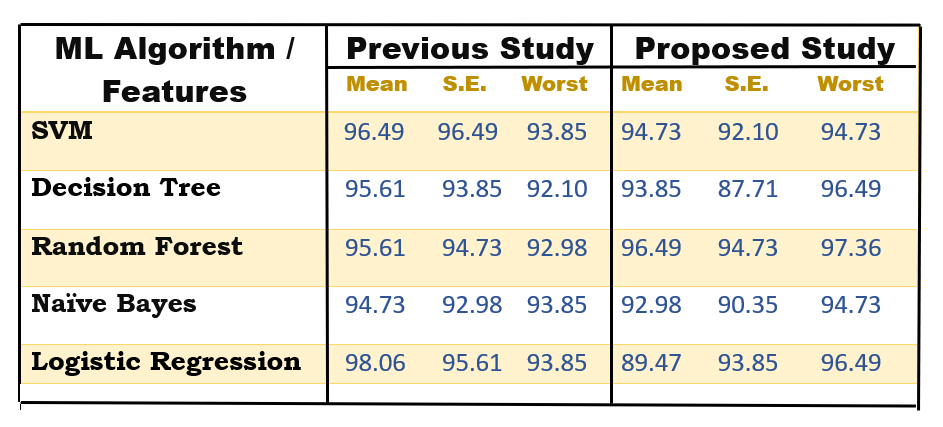
****

****

**Graphical Representation of \_mean, \_se, \_worst suffix of different ML models**

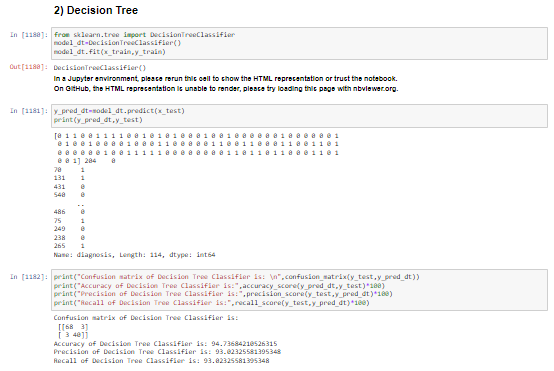


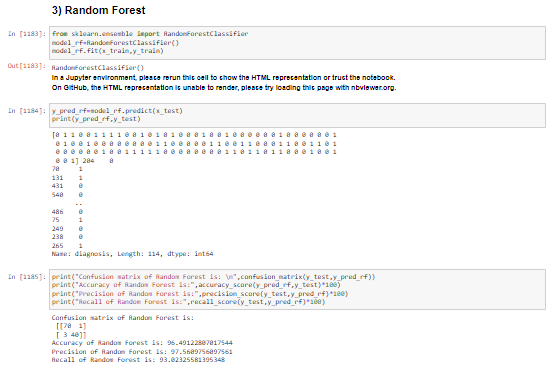
**Tabular Representation of \_mean, \_se, \_worst suffix of different ML models**

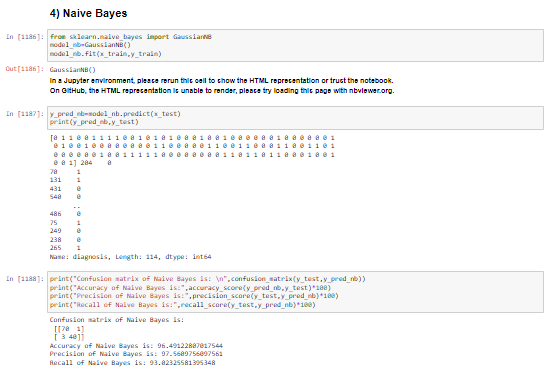


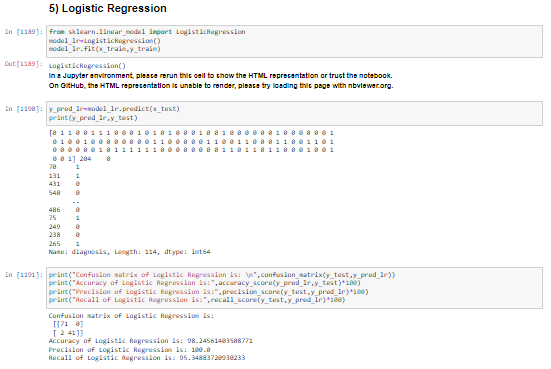
1. **Module 2 – Using Whole Dataset**
   1. **) Without Feature Engineering**





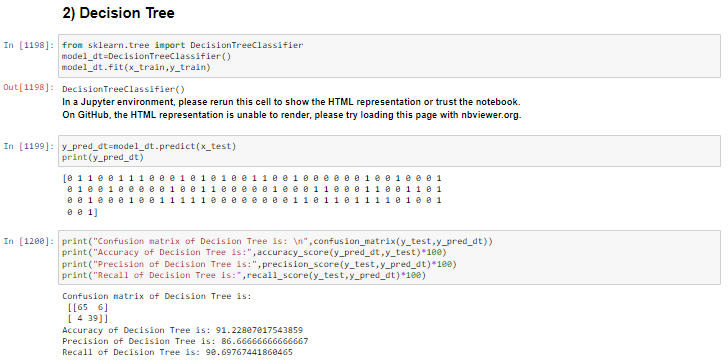


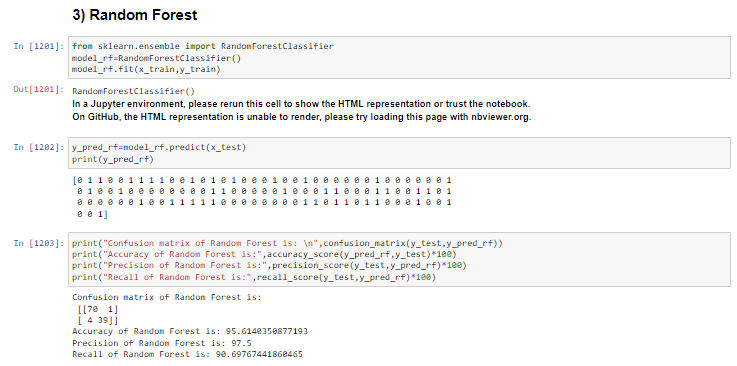


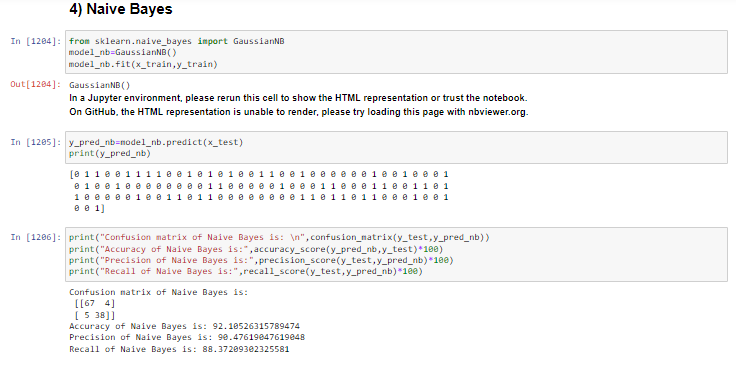


**2.2) After Applying PCA**

****

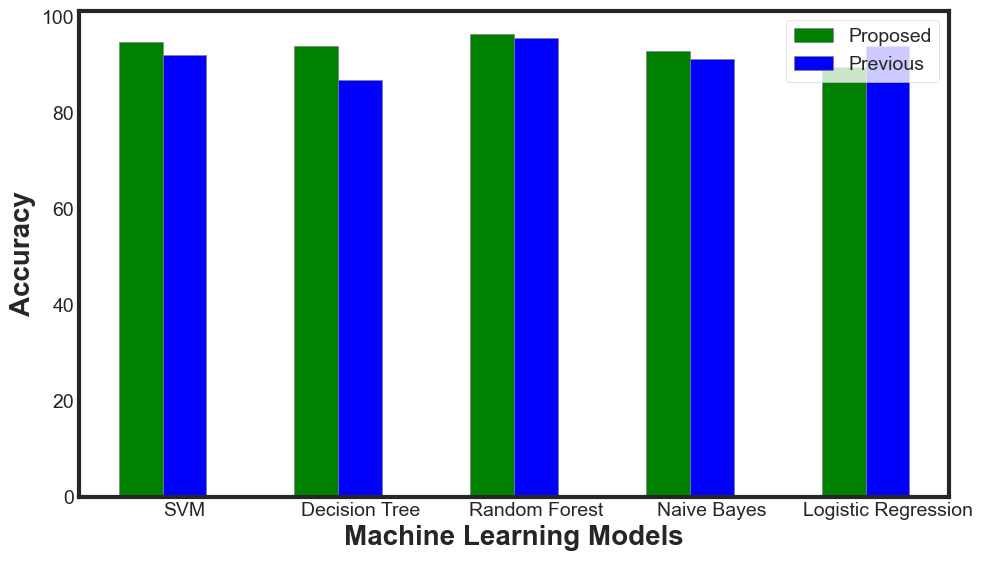
****

****

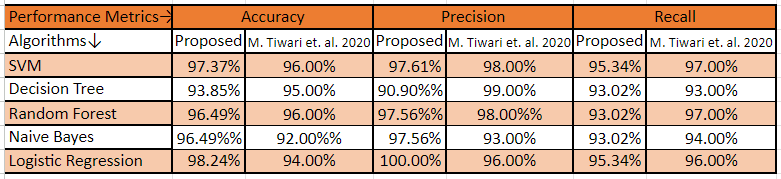
****

****

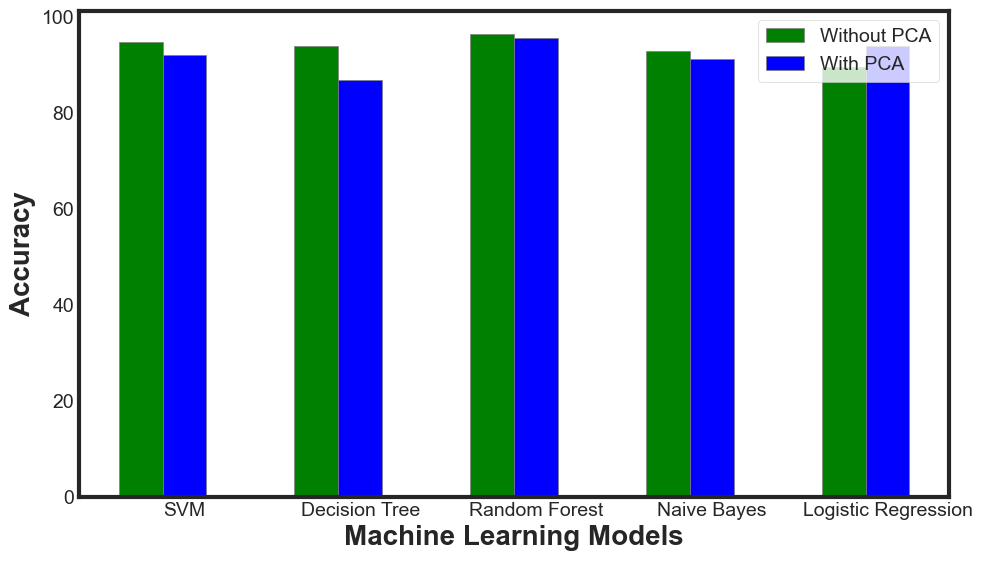
**Graphical Representation of Comparison of Accuracy of Proposed (2.1 ML Models) and Previous Models**

****

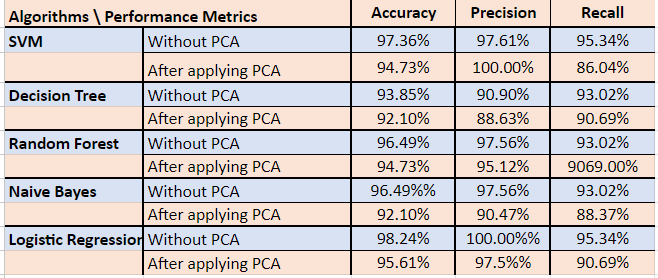
**Tabular Representation of Comparison of Accuracy of Proposed (2.1 ML Models) and Previous Models**

****

**Graphical Comparative Analysis of Accuracy of 2.1 ML Models (Without PCA) and 2.2 Models (With PCA)**

****

**Tabular Comparative Analysis of Accuracy of 2.1 ML Models (Without PCA) and 2.2 Models (With PCA)**

****

**RESULT**

After much research and analysing previous research papers as well , we found that for our proposed study -

1. **Logistic Regression Model** gave the best result – an accuracy of **98.24%** on the whole dataset i.e. considering all the **30 attributes**.
2. **Random Forest Model** gave the best result – an accuracy of **97.36%** on consideration of the **\_worst valued attributes** i.e. considering only

10 attributes.

Further, we have also deployed our Breast Cancer Classification Project to check the malignancy of the tumour and predict whether the tumour is cancerous or not.

**CONCLUSION**

* The objective of the project was to predict whether the tumour is malignant or benign.
* Different Machine Learning algorithms were used to make prediction using the Wisconsin Breast Cancer dataset.
* Logistic Regression provided highest accuracy of 98.24%.
* While Decision Tree Classifier gave the worst accuracy of 94.74% for prediction.
* Upon standardizing data, we get good accuracy results using Support Vector Machine, Logistic Regression, Random Forest Decision Tree and Naive Bayes, correctly classifying tumour into malignant or benign almost 95-98% of times.

### LIMITATION OF OUR STUDY

When using known facts and attempting to generalise to situations that aren't expressly covered in the knowledge base, issues arise due to the overfitting and overgeneralization effects. These issues also arise with techniques that make use of machine learning techniques. The present study faced significant challenges due to a lack of supplies.

**APPENDICES**

[1] [https://techvidvan.com/tutorials/breast-cancer-classification**/**](https://techvidvan.com/tutorials/breast-cancer-classification/)

[2]<https://projectgurukul.org/breast-cancer-classification-using-machine-learning/>

[3] <https://emeritus.org/in/learn/types-of-supervised-learning/>

[4] <https://www.ibm.com/in-en/topics/supervised-learning>

[5] <https://www.narayanahealth.org/blog/stage-wise-treatment-of-breast-cancer/>

**REFERENCES**

[1] David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu “Machine Learning Classification Techniques for Breast CancerDiagnosis” (2019).

[2] S.Vasundhara , B.V. Kiranmayee and Chalumuru Suresh "Machine Learning Approach for Breast Cancer Prediction" (2019).

[3] Anji Reddy Vaka, Badal Soni and Sudheer Reddy “Breast Cancer Detection by Leveraging Machine Learning” (2020).

[4] Kalyani Wadkar, Prashant Pathak and Nikhil Wagh “Breast Cancer Detection Using ANN Network and Performance Analysis with SVM” (2019).

[5] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S "Breast Cancer Prediction using Machine Learning" (2019).

[6] Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications" (2020).

[7] Arpita Joshi and Dr. Ashish Mehta “Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer” (2017).

[8] Abdullah-Al Nahid and Yinan Kong “Involvement of Machine Learning for Breast Cancer Image Classification: Asurvey” (2017).

[9] K.Anastraj , Dr.T.Chakravarthy and K.Sriram " Breast Cancer detection either Benign Or Malignant Tumour using Deep Convolutional Neural Network With Machine LearningTechniques " (2019).

[10] Saeed S, Jhanjhi NZ, Naqvi M, Humyun M, Ahmad M. Optimized breast cancer pre-mature detection method with computational segmentation: a systematic review mapping. Approach Appl Deep Learn Virt Med Care. (2022)

[11] Monika Tiwari, Rashi Bharuka, Praditi Shah and Reena Lokare “Breast Cancer Prediction using Deep learning and Machine Learning Techniques” (2020).

[12] <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)>

[13]<https://github.com/milaan9/93_Python_Data_Analytics_Projects/tree/main/007_Breast_Cancer_Prediction_with_ML>

[14] <https://www.geeksforgeeks.org/bar-plot-in-matplotlib/>

[15] <https://scikit-learn.org/stable/modules/preprocessing.html>

[16]<https://www.researchgate.net/publication/341871629_Breast_Cancer_Classification_and_Prediction_using_Machine_Learning>

[17]<https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240>

[18]<https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>

[19] <https://www.cancer.gov/types/breast>

[20]<https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps>

[21] <https://www.javatpoint.com/machine-learning-life-cycle>

[22] <https://www.mskcc.org/cancer-care/types/breast/diagnosis/stages-breast>

[23] <https://www.breastcancer.org/pathology-report/breast-cancer-stages>

[24]<https://www.google.com/search?q=stages+of+breast+cancer&rlz=1C1CHBD_enIN911IN911&sxsrf=APwXEddKlvOzKFtdyJ5XlC6HurOX9N8K-A:1681268142226&source=lnms&tbm=isch&sa=X&ved=2ahUKEwj7vo7zq6P-AhXZumMGHSFXD6cQ_AUoAXoECAEQAw&biw=1600&bih=762&dpr=1.2#imgrc=MBVQL6Kw9C43pM>

[25]<https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/>

[26]<https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>

[27] <https://scikit-learn.org/stable/modules/tree.html>

[28] <https://scikit-learn.org/stable/modules/svm.html>

[29] <https://scikit-learn.org/stable/modules/naive_bayes.html>

[30]<https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

[31]<https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

[32] <https://www.webmd.com/a-to-z-guides/fine-needle-aspiration>

[33]<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.freepik.com%2Ffree-vector%2Fbreast-cancer-awareness-month-concept_9890305.htm&psig=AOvVaw1US2D6Pjm8APjHIdAoCd8P&ust=1681355145131000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCKiG0JOuo_4CFQAAAAAdAAAAABAE>

[34]<https://www.google.com/search?q=ml%20classification&tbm=isch&tbs=rimg:CRHx_agvjlS9YRb1wzIsNwE9sgIMCgIIABAAOgQIABAAwAIA&rlz=1C1CHBD_enIN911IN911&hl=en&sa=X&ved=0CBoQuIIBahcKEwjgx8zasaP-AhUAAAAAHQAAAAAQIg&biw=1579&bih=762#imgrc=S0IzRgWEGuZmZM&imgdii=HpDpNnOQyJZNoM>