# ABBREVIATIONS

|  |  |
| --- | --- |
| **Abbreviations** | **Full Forms** |
| BC | Breast Cancer |
| DT | Decision Tree |
| KNN | K-Nearest Neighbors |
| LDA | Linear Discriminant Analysis |
| LR | Logistic Regression |
| ML | Machine Learning |
| NB | Naïve Bayes |
| SVM | Support Vector Machine |
| WBC | Wisconsin Breast Cancer |

**ABSTRACT**

Each year, countless lives are claimed by breast cancer. It remains a significant opponent in women's health. The healthcare community must select whether to focus on early detection or accuracy in diagnosis for better patient results. While our old ally, traditional mammograms, sometimes fail us in terms of their detection rate we must consider more advanced methods. This project is about using ML to detect breast cancer. LR, LDA, DT, KNN, NB and SVM are the six algorithms we used. Each algorithm has its own strengths, so we employed them as our instruments towards understanding the intricate nature of breast cancer data. Our expedition began with the Wisconsin Breast Cancer (WBC) dataset, a treasure trove of information from the UCI Machine Learning Repository. This diagnostic dataset, comprising 569 patient cases and 30 features, became the foundation of our research. We treated this data with utmost care, employing various preprocessing techniques. We standardized the features to ensure fair comparison, discretized the data using KBinsDiscretizer to capture important thresholds, and addressed the imbalance between benign and malignant cases using the Synthetic Minority Over-sampling Technique (SMOTE). We turned to PCA. This technique helped us navigate through the complexity, reducing the dimensionality while preserving the essence of our data. The first ten principal components emerged as our beacons, explaining approximately 85% of the total variance and illuminating the path forward.

With our data prepared, we set our ML models to the task. Each algorithm was trained-and-tested in a carefully crafted Python environment. To determine their effectiveness, we used different measures which included accuracy, pre-cision, re-call and F1-score. This metrices became our compass, helping us gauge the true effectiveness of each model. The results were both exciting and promising. In our initial validation, KNN and SVM emerged as frontrunners, both achieving an impressive 98.84% accuracy. LDA followed closely behind with 97.67% accuracy, showing that it too had valuable insights to offer. However, the true test came with our final evaluation on the unseen test set. LR had the highest percentage of 97.67% accuracy followed by precision and recall rates of 96.88% each then LDA plus SVM tied at 96.51% accuracy for this stage. These findings filled us with optimism. They demonstrated that ML models can enhance our ability to detect breast cancer, potentially offering a powerful complement to traditional diagnostic methods. The strong performance of LR, KNN, and SVM, in particular, hinted at their potential as valuable allies in the fight against breast cancer. But our journey doesn't end here. We see numerous paths for future exploration. Ensemble learning methods, which combine the strengths of multiple models, could potentially push our detection rates even higher. Additionally, applying these models to diverse datasets could help us understand their broader applicability and robustness. As we conclude this study, we find ourselves filled with hope and determination. The research has shown signifi-cant improvements especially related to LR, KNN and SVM models indicating that ML can be a useful additional tool for diagnosis. When used alongside conventional methods, these sophisticated techniques may help in early identification leading to more accurate diagnosis thus increased chances of survival.

*Keywords: Logistic Regression (LR), Linear Discriminant Analysis (LDA)), Decision Tree (DT), K- Nearest Neighbors, Naïve Bayes (NB), Support Vector Machine (SVM), Machine Learning (ML), Breast Cancer (BC), Wisconsin Breast Cancer (WBC), Principal Component Analysis (PCA)*

**Table Of Contents**

|  |
| --- |
| **Abbreviations V** |
| **Abstract** |
| **1 Introduction 1** |
| **2 Literature Review 3** |
| 1. **Materials and Methods 5** 2. **Results and Discussion 20** 3. **Conclusion and Future Prospective 32** 4. **References 33** |

**List of Figures**

An example of Convolutional Neural Network [7] 3

Various types of ML techniques [11] 8

Choosing and training a ML method [12] 9

Linear Discriminant Analysis Algorithm [13] 10

Fine Decision Tree Algorithm [13] 10

K-nearest neighbors algorithm [14] 11

A conceptual description of disease stratification using SVM [15] 12

Schematic representation of Confusion Matrix.[17] 13

Schematic representation of proposed method 15

Representation of Dataset before Pre-processing 20

Representation of Dataset after encoding and removal of column ‘ID’ 20

Dataset before feature standardization 21

Pie-graph (Left picture) shows train: validation: test ratio and Bar-graph (Right picture) shows the count of the respective classes present in Test set (blue), Train set (orange) and Validation set(green) 22

Image represents two columns: Column 1 is the Histogram of the original feature values for the training (blue), validation (green), and test sets (red). Column 2 is the Histograms of the discretized feature values for the training (blue), validation (green), and test (red) sets 23

Scattered plot of Original training data (left-side) and resampled training data (SMOTE) (right-side) points out the benign (blue) and malignant (red) datapoints 24

Scattered plot for the first two standardized features for the resampled training data (left), validation (middle) and test data (right) 25

Heatmap of feature correlation 26

Confusion-matrix for Logistic regression 27

Confusion-matrix for LDA 27

Confusion-matrix for Decision Tree 28

Confusion-matrix for KNN 29

Confusion-matrix for Naïve Bayes 30

Confusion-matrix for SVM 31

# LIST OF TABLES

*Hyper-parameter for Logistic regression 18*

*Hyper-parameter for Linear Discriminant Analysis 18*

*Hyper-parameter for Decision Tree 18*

*Hyper-parameter for K-Nearest Neighbors 19*

*Hyper-parameter for Naïve Bayes 19*

*Hyper-parameter for Support Vector Machine (SVM) 19*

*Standardized features 21*

*Standardized feature 22*

*Table which shows the rescaled features where the mean is zero and standard deviation is one 25*

*Classification-report for logistic regression 27*

*Classification-report for LDA 28*

*Classification-report for Decision Tree 28*

*Classification-report for KNN 29*

*Classification-report for Naïve Bayes 30*

*Classification-report for SVM 30*

*Final evaluation metrices on the test-set 31*

# INTRODUCTION

Breast cancer is still one of the contenders for the most important health issues for females all around the world. Statistics show that more than two million women get diagnosed every year and this emphasizes the need for better ways of diagnosis. Traditional methods like mammograms are broadly used but they have their own limits especially during the detection of accuracy and diagnosis rate. This means that there should be new techniques for early detection so as to improve patient outcomes. In the past, there has been significant advancement in machine learning which has brought great potential in changing how breast cancer is detected and diagnosed. ML models have a capability to go through huge sets of data, recognize patterns and predict results with a very high precision. Due to this they are considered as useful tools in medical diagnosis. My research dives into six ML models - LR, LDA, DT, KNN, NB, and SVM - utilizing the WBC dataset. The goal was to establish the model that is most accurate and reliable in detecting breast cancer. LR had the highest accuracy of 98.13%, closely followed by SVM - 97.9%. Even though it is helpful, the DT model had the least accuracy at 92.48%. These findings indicate the potential of ML in improv-ing the accuracy of breast cancer detection by a large margin. Many groundbreaking studies in this field have been converged with this research.

For instance, [1] utilized various ML algorithms, including Naive Bayes and SVM, to predict breast cancer type, achieving accuracies of 91% and 89%, respectively. Their study underscores the advantage of employing multiple ML models to optimize diagnostic performance. Similarly, [2] presented an ensemble approach combining NB and SVM, resulting in a notable accuracy of 97.3%. This ensemble approach is a testament to the effective-ness of integration of various models to enhance prediction accuracy. Further, studies like by [3] have utilized ML algorithms for BC risk prediction, achieving a highest accuracy of 97.13% with SVM. These findings are corroborated by [4] who proposed the application of novel algorithms like SMO and reported an accuracy rate of 96.19%, demonstrating the evolving landscape of ML techniques in cancer detection. [5] conducted a comparison of various ML algorithms for BC detection and diagnosis, which involved pointing out the plus and minus of each. The understandings that they obtained by the comparison, they made on these models closely aligned with our results and methods during selection along with optimization of ML in addition predictive models abilities beyond health care have been witnessed in different fields recent times also revealed. For instance, various studies have been carried out such as one that focused on using neural networks towards improving accuracy rates when detecting whether a person has got malignant cells or not.

To validate our findings, we will blend these various investigations in this dissertation and indicate how much breast cancer diagnostic is expected by the ML to change. We wish to carry out more research in future where we shall come up with some ML and DL models that bring together various ML and DL methods so that we can increase the rate of accuracy in detection even further; hence creating more trustworthy tools for early diagnosis than ever before. By combining all these contributions into one paper, this shows that ML has potential when it comes to improving detection of BC which will in turn help saving lives by better patient-care.

# OBJECTIVES

The aim of our thesis work are:

1. **Data Pre-processing** - Handling missing values, rebalancing class distribution, and standardizing features for effective model training on the Wisconsin Breast Cancer dataset.
2. **Model Evaluation** - Training six different ML models: LR, LDA, DT, K-Nearest Neighbors, Naïve Bayes and Support Vector Machine. By using these models we evaluate them using metrics like precision, recall and F1 score apart from accuracy.
3. **Model comparison** – Per-formance comparison with various models to establish the most accurate and reliable model for breast cancer detection.

# LITERATURE REVIEW

According to the information provided, breast cancer ranks among the top killer diseases worldwide when it comes to women. Therefore, early discovery is critical for successful treatment and survival. From medical photos and also from other associated data, techniques of ML and DL have shown great impact in helping detect and diagnose BC.

[6] conducted research that proved how ML could be applied to diagnose BC rapidly, inexpensively and accurately through lifestyle factors of women with different basic machine learning models and ensemble methods. It can therefore be inferred that this would en-able detection in resource-limited areas where such facilities are not readily available.

A comprehensive review of ML and DL for detecting breast cancer from medical images and it was found that this is one of those things which can be done with the help of ML and DL models [7]. They highlighted the amount of information contained in them as well as how powerful they are when used together with the two techniques for feature-extraction and analysis towards better detection and diagnosis. Their review would help so much especially doctors who deal with BC patients [8].

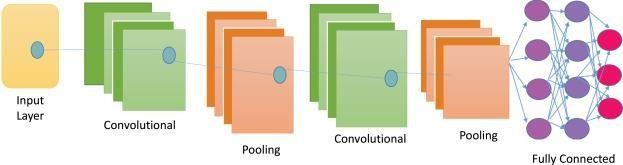


Fig 2.1: An example of Convolutional Neural Network [7]

Alhussan et al. looked at the categorization of BC from digitalized mammograms by using statistical-ML methods and DL techniques. According to them, statistical ML might be seen as a complement to DL in computer-aided breast cancer detection where limited dataset is concerned [9]. A study on DL-network architectures for diagnosing BC. They addressed different aspects of model designing in DL while emphasizing the importance of appropriate choice and preprocessing of input datasets along with hyperparameter tuning towards achieving optimum performance levels [10].

*4 2. Literature Review*

The evidence found in literature indicates that the potentials of ML, DL in the improvement of breast cancer detection and diagnosis are huge. Nevertheless, further detections are needed if we are to exploit these potentials fully and overcome such challenges as those encountered when trying to implement the techniques in clinical settings where patients live. Equally important is continuing with our exploration into statistical ML methods along with deep learning systems or even their combination so as to take advantage of what each has to offer for more accurate and consistent findings.

# MATERIALS AND METHODS

## Datasets

*Datasets* is the collection of data. these datasets may be tabular, numerical, or image data. In this work. we are focused on numerical data as we are working on the detection of BC.

*Data acquisition* is a process of extracting data from a source. Here in this work, we collected data from various datasets.

Dr. William H. Wolberg from University of Wisconsin is the person who started this data set. People can find this data on UCI Machine Learning Repository, but it's also obtainable at Kaggle so anyone from anywhere can-do research and practice with this dataset easily. The main goal of these diagnostic data is to recognize whether a tumor is malignant or not. Each instance contains 569 objects with 30 real-valued features.

These features are related to cell nuclei primarily which include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. Every variable is measured for mean, standard-error and average of the three largest values of each attribute. The target variable in this dataset is diagnosis of breast cancer where 'M' means malignant tumor and 'B' stands for benign tumor. This dataset is used for binary-classification problem which is very important for creating and testing machine learning models designed to identify breast cancer at an early stage with high precision. The data set is complete without any missing values so it can be considered as a near perfect sample set giving good quality info for training and testing predictive models. Various investigations have seen the dataset is utilized to advance the discovery of breast cancer which had major impact where different machine learning methods are compared to one another based on their effectiveness.

**UCI Machine Learning Repository:** Itis an invaluable, well-respected resource with a long history and wide use within the ML community. Datasets such as the WBC Diagnostic dataset, established by esteemed scholars at University of Wisconsin, are frequently mentioned and employed in academic as well as industrial research. The repository guarantees topnotch data quality and comprehensive documentation hence it is considered as a reliable source for benchmark datasets that foster developments in machine learning.

**Kaggle’s** WBC Diagnostic dataset is very significant and credible as it is broadly used for machine learning research and learning tools. It was originally created at University of Wisconsin and has 30 features defining characteristics of cell nuclei as well as 569 cases. With Kaggle’s platform, data scientists can easily analyze the data and compare various models for performance. Many medical data analysis studies cite this particular set; this shows how much the community trust its accuracy or value in determining what type tumors might be present through tests that classify them based off sample findings.

### Pre-Processing Methods

Preprocessing methods are important ways that is used for converting raw data into a format which is suitable for analysis thereby improving the performance of subsequent data mining and ML operations. These techniques are meant to get rid of variations and noises, normalize and scale the data so that it can be modeled. Common methods include splitting the dataset, standardizing values, turning continuous attributes into nominal ones through discretization as well as resampling techniques like SMOTE. Moreover, dimensionality reduction tools such as PCA help in reducing feature space while retaining necessary variance which makes model training quicker and more effective. These steps in pre-processing are necessary if one wants accurate, dependable and easily understandable results from their study on any subject grounded on information collected electronically or otherwise.

### Data Partitioning:

In machine learning, data splitting is a critical stage of the process in which a dataset is divided into distinct subsets for model training and estimation. It is normal to divide the data in three parts: training-set, validation-set, and test set. Training is done using the training set, hyperparameters are modified using the validation set to prevent overfitting the model; testing is performed on the unseen data from the test-set to evaluate its performance. The ratio for this common division is 70-80% for training and 20-30% for testing, but if necessary, an additional split can be made for validation purposes. It is necessary that these models can generalize well on new examples so correct partitioning should be applied. Stratified sampling may be applied in classification tasks so that class distributions across subsets can remain balanced thus ensuring that all categories have sufficient representatives within both training and testing environments.

### Standardization:

Standardization is a pre-processing technique that transforms the characteristics so that they have a mean of zero and a standard deviation of one. In data mining this is essential when working with SVM, KNN, and PCA. This can be mathematically represented as:

z = (x−μ)/σ

In the above equation x stands for the feature value while μ represents its mean value of the feature and σ is the standard deviation. By standardizing the data set, each feature will have equal influence on model training thereby avoiding domination of those with larger scales among them which might lead to biased results especially in linear models like multiple regression analysis. Another advantage is that it speeds up convergence of iterative optimization algorithms such as gradient descent method due to balanced input distribution.

### Discretization:

Discretization is the process of transformation of continuous features to discrete categories or bins. Discretization helps to simplify models and also deals with non-linear relationships between different variables. KBinsDiscretizer in Scikit-learn uses methods like uniform binning which divides the range for each feature by width equally; quantile binning where every bin has approximately same number of samples or k-means clustering which groups similar k-means together then finds best cuts according to mean distance from centres. By doing this we can make any ML algorithm that uses trees more powerful. Most algorithms work better if the input is categorical. It also deals well with outliers because if there are numerous outliers then they will be put into few larger bins where they won’t affect our model too much.

### SMOTE:

The procedure of dealing with class imbalance in datasets through generating synthetic samples for the minority class using KNN is known as SMOTE (Synthetic Minority Over-sampling Technique). The fundamental concept is to construct new cases by linearly interpolating between existing instances of the minority class. The SMOTE algorithm can be considered to involve the following stages:

1. Choose an instance from the minority class at random.
2. Determine its k nearest neighbours among the minority instances.
3. Generate synthetic examples by knitting the chosen instance and its neighbours together along the lines in which are joined previously.

By ensuring that the distribution of classes is balanced, SMOTE prevents classification algorithms from being biased towards the majority class thus improving their performance. It is found to work best when it is used in con-junction with other methods like standardization, feature selection etc.

### Principal Component Analysis

PCA is a dimensionality reduction technique involving transformation of high- dimensional data into a lower-dimensional form while conserving as much variance as possible. It accomplishes this by finding linear combinations of the original features that represent the ‘principal components’. Steps involved to perform PCA:

1. Standardize your dataset;
2. Compute the covariance matrix;
3. Derive the eigenvalues and eigenvectors from the covariance matrix;
4. Sequencing the eigenvectors by eigenvalues in descending order to select the top components.

The formula used is: Z = XW

Where Z represents the matrix of principal components; X stands for standardized data matrix while W denotes eigenvector matrix. Through model simplification, PCA facilitates faster computational performance and sometimes can even lead to better results by removing unnecessary variables (noise) or highly correlated predictors.

## Machine Learning

Selecting the proper algorithm and ensuring the quality of the data are two main elements when it comes to creating effective ML models that can offer accurate predictions or decisions which are dependable [11].

ML is a type of AI that gives computers the skill to learn without being plainly programmed and improve from experience. ML is based-on the premise that systems can automatically learn and improve from experience; such systems utilize data to create models, which could be used to make predictions or decisions. These models are built using sample data known as training data, without any task-specific program [11].

There are four-main types of machine learning algorithms:

### Supervised learning:

This kind of algorithms are trained on labelled data where every training example comes with an output label. They are commonly used for tasks like classification and regression [11].

### Unsupervised learning:

In these algorithms, no labels are given to the data. The system tries to learn patterns and structures from input data alone. Typical tasks include clustering or association [11].

### Semi-supervised learning:

It is a method of training that utilizes both small amounts of labell-ed in-formation combined with large amounts of un-labelled in-formation during the training process. It can greatly increase accuracy compared to un-supervised methods alone [11].

### Reinforcement learning:

An agent/liason interacts with its own environment by making actions and discovering errors or rewards. Through this trial-and-error process, the agent learns what actions produce the highest rewards [11].

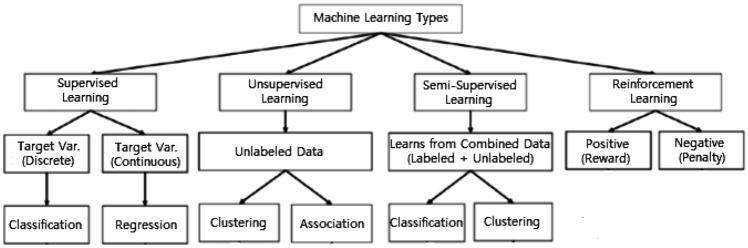


Fig 3.1: Various types of Machine Learning Techniques [11]

The 6 ML algorithms we selected are LR, LDA, Decision Tree, K-Nearest Neighbors (kNN), Naive Bayes and SVM all comes under supervised learning.

Machine Learning 9

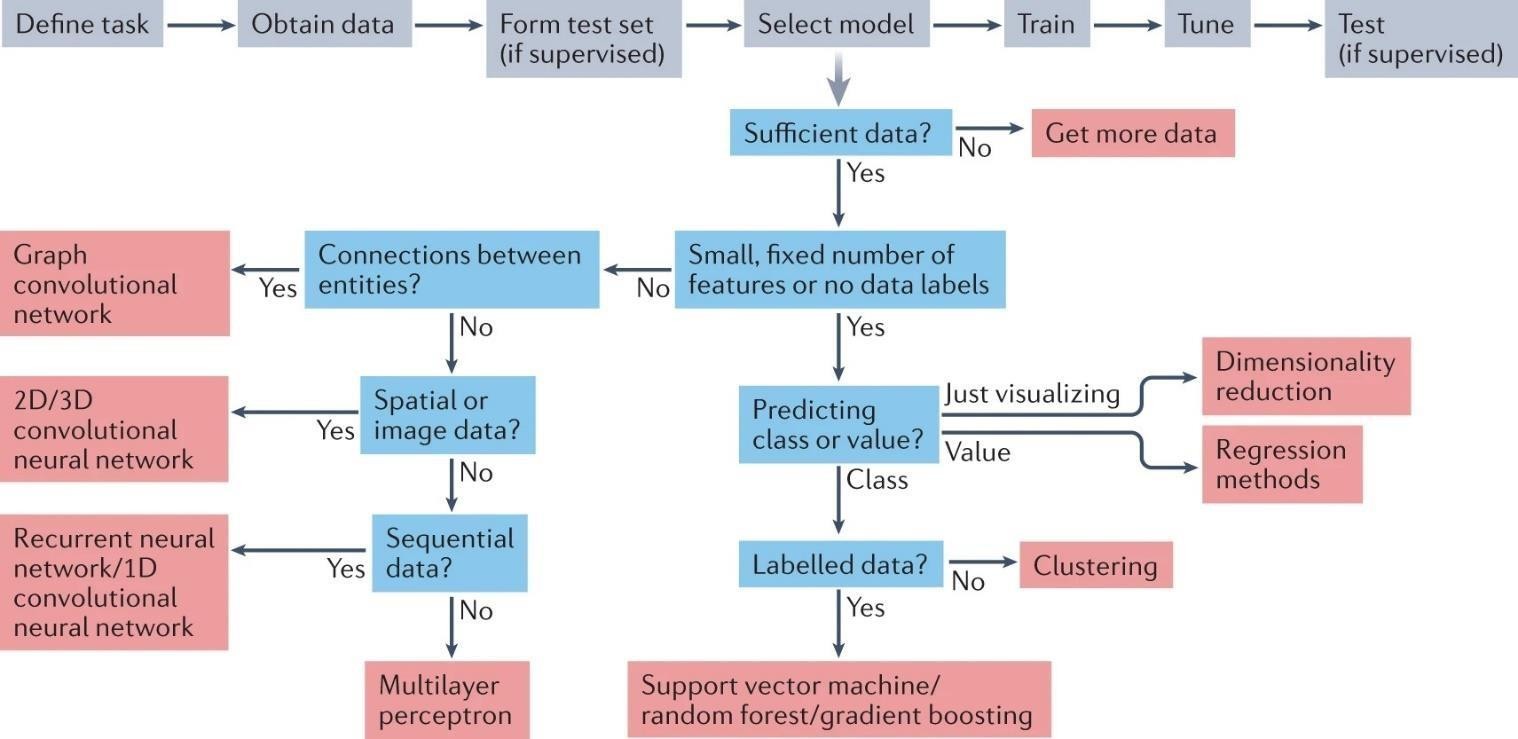


Fig 3.2: Choosing and training a ML method [12]

### MACHINE LEARNING ALGORITHMS

In ML techniques, there are various algorithms depending on the availability and the nature of the dataset.

### Logistic Regression:

A probabilistic statistical model used for classification. It uses a logistic function to calculate probabilities for classes which are estimated by the sigmoid function. Effective when applied to linearly separable datasets, however can overfit high dimensional data. Regularization techniques such as L1 or L2 regularization may be employed to prevent overfitting LR assumes that there is a linear-relationship between dependent and independent variables [11].

### Linear Discriminant Analysis:

A linear-decision boundary classifier that fits class conditional densities to the data and applies Bayes’ rule. LDA projects the data onto a lower dimensional space in order to reduce model complexity as well as computational cost. Usually each class is assumed to have a Gaussian distribution for its density function while sharing same covariance matrix across classes LDA is also related with ANOVA or Regression analysis [11].

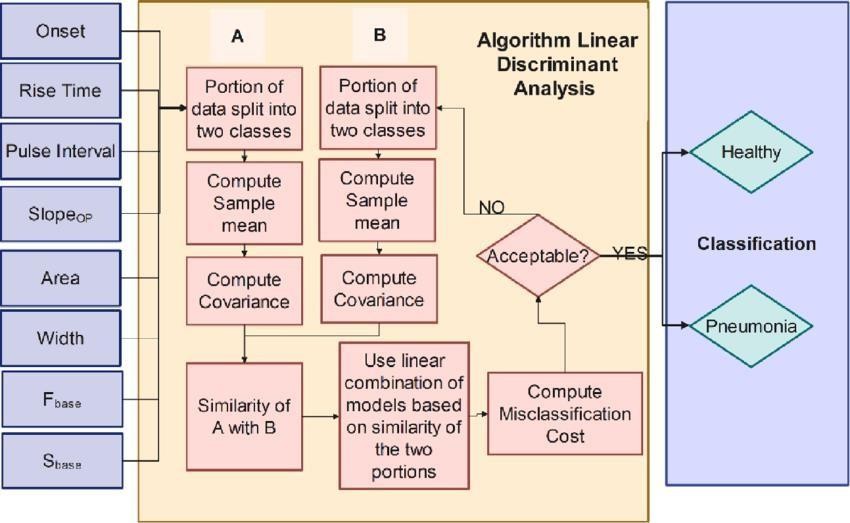


Fig 3.3: Linear Discriminant Analysis Algorithm [13]

### Decision Tree:

It is a rule-based classifiers that rely on IF-THEN rules for class prediction. They are known to be easy to interpret, can handle high-dimensional data, and are accurate. Decision Trees generate rules that can also be used to predict the classes of unseen test cases. Because of its simplicity and speed, this model is frequently used in classification tasks [11].

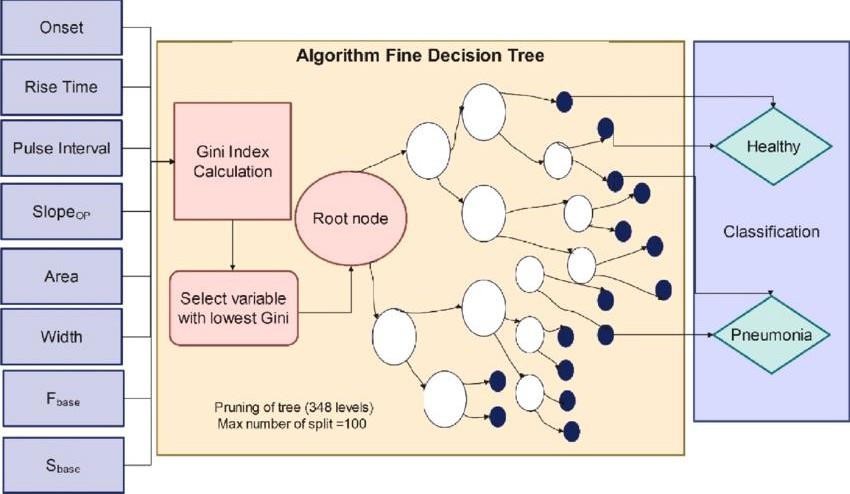


Fig 3.4: Fine Decision Tree Algorithm [13]

### K-Nearest Neighbors (kNN):

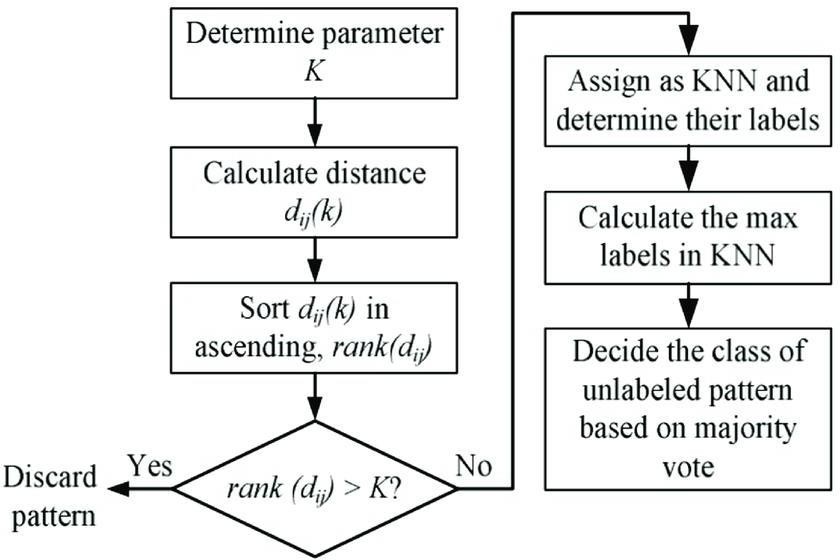


Fig 3.5: K-nearest neighbors algorithm [14]

An instance-based or non-generalizing learning algorithm which is usually referred as a "lazy learner". Rather than creating a general internal model, it stores all training instances. When applied to new data points, classification is based upon the majority vote among their k-nearest neighbors using similarity measures such as Euclidean distance. kNN is tolerant to noisy sets but accuracy heavily relies on quality of sets as well as choice for k [11].

### Naive Bayes:

It is a probabilistic-classifier based on the Bayes theorem. It assumes strong independence among the features, making it simple and effective for large datasets. Naive Bayes is most appropriate for text-classification and spam-filtration, where the independence assumption is frequently valid [11].

### Support Vector Machine (SVM):

It is typically employed for classifying or for regression tasks by creating a hyperplane or hyperplanes in high- or infinite-dimensional space. The greatest hyperplane is the one that maximises the margin between classes, hence minimizing generalisation error for the classifier. SVMs utilize various kernel functions such as linear, polynomial, and radial basis function for dealing with non-linearly separated data. Nevertheless, SVMs tend to struggle with noisy data that contain overlapping classes [11].

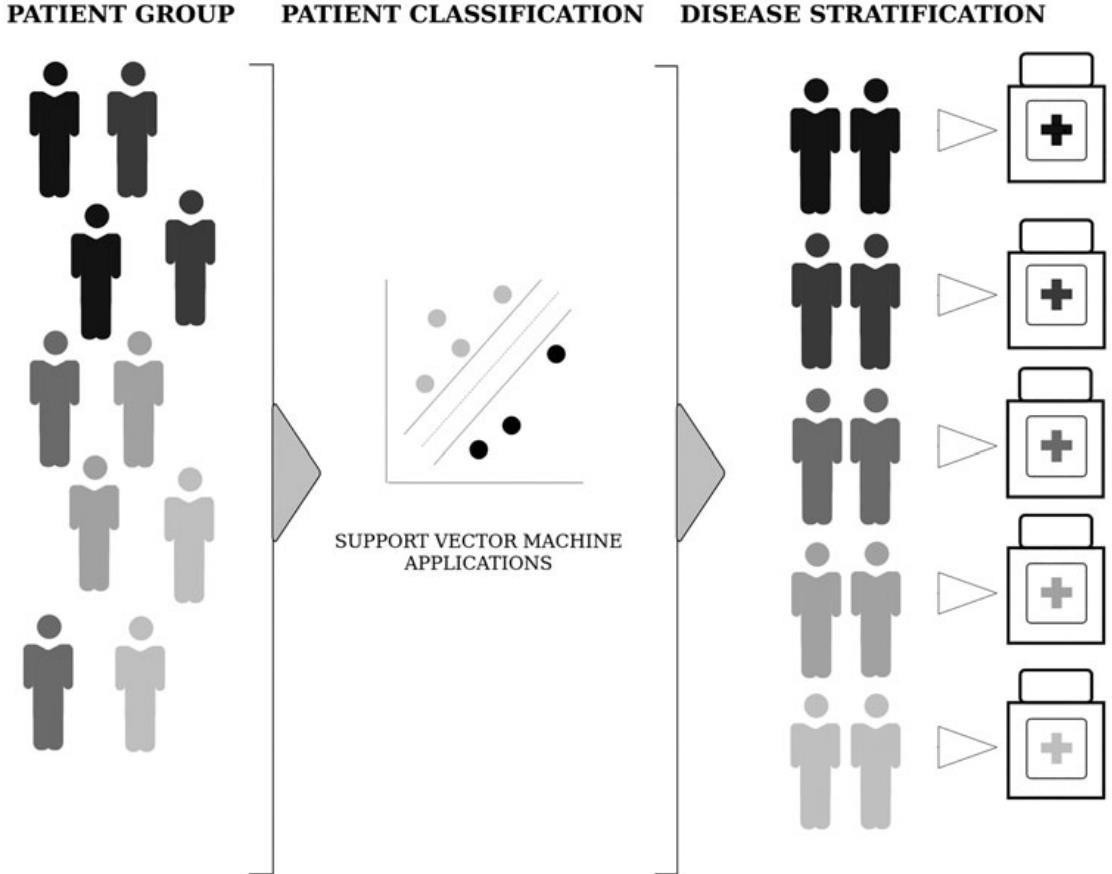


Fig 3.6 A conceptual description of disease stratification using SVM [15]

# EVALUATION CRITERIA

To validate trained machine learning models, they should be evaluated against a validation dataset. All the evaluation-metrics for classification-algorithms are based on the *Confusion Matrix* (truth matrix). This prediction gives the accuracy scores and actual values. The predicted model is compared with the actual values and added to the confusion-matrix as True-positives, True-Negatives, False-Positives, and False-Negatives, as shown in figure 3.7. In the evaluation criteria, there are various evaluation methods such as precision, recall, accuracy, and F1 score.

The ratio between the True positives and all the positives in the evaluation is known as the *Precision* 3.1. Recall is an evaluation criterion where the identification of True positives is done 3.2. The ratio of the sum of all correct predictions and the sum of predictions infers the *accuracy* 3.3. F1 score provides one score that takes both the precision value and the recall value into to single numerical value. 3.4 [16].

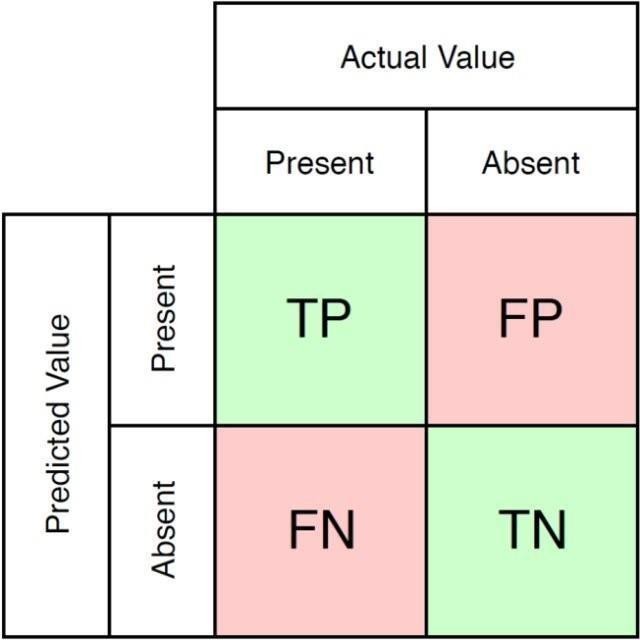


Fig. 3.7: Schematic representation of Confusion Matrix.[17]

*TP*

*Precision* = *TP* +*FP*

(3.1)

*TP*

*Recall* = *TP* + *FN* (3.2)

*TP* +*TN*

*Accuracy* = *TP*+ *FP* + *TN* + *FN* (3.3)

*F* 1 = 2 ∗ *Precision* ∗ *Recall*

*Precision* +*Recall*

(3.4)

## The ML Models for The Detection of Breast Cancer

Based on the above-discussed methods, for this thesis work, where we are concerned with detecting the tumor into specific types and comparing with the six different models for higher accuracy scores. As we know SVM and Naive Bayes algorithms are quite old, and also, they work on binary data, for the machine learning model we choose them as well as other 4 commonly used models which are most popular, and also through our literature review we can see that they are preferred models for classification in the machine learning method. As our data is labeled, it is supervised learning.

## Methodology Proposed Method

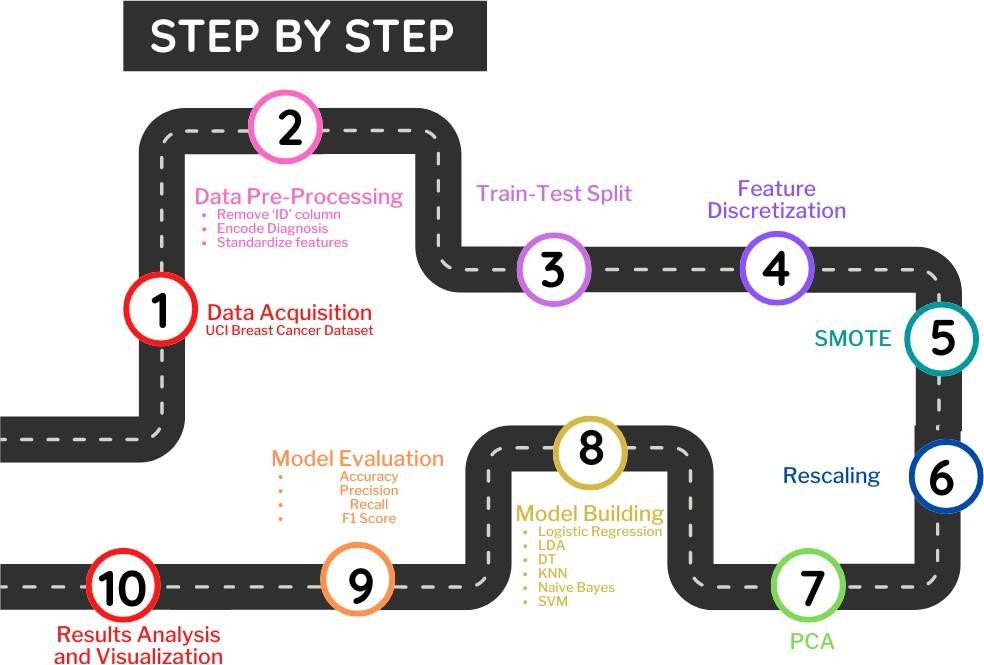


Fig. 3.8: Schematic representation of proposed method

As mentioned in earlier, the data needs to be pre-processed before subjecting it to the model training. The methodology for pre-processing and model training has been discussed in this section.

### Data Acquisition:

This report uses the Breast Cancer Wisconsin (Diagnostic) dataset that was retrieved from the UCI Machine Learning Repository. In the ML community, this particular dataset is very well known and contains 569 instances with 30 features each derived from the digitized images of the fine needle aspirate (FNA) of breast mass. Each instance is classified as either benign or malignant making it suitable for binary classification problems. These include attributes such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The features are computed from a digitized image of FNA on breast mass and describe various characteristics about cell nuclei present in the picture. The reason in choosing this dataset includes its applicability in real-world medical diagnosis, good curation and equal representation of positive and negative cases which provides strong grounds for building ML models for breast cancer diagnosis.

### Data Pre-processing:

This phase ensures the quality and suitability of the dataset for ML algorithms. Initially, the 'ID' column does not contribute anywhere so it is removed. The 'Diagnosis' column, originally containing categorical values 'M' for malignant and 'B' for benign, was encoded numerically (1 for malignant, 0 for benign) to facilitate mathematical operations in the models.

Feature standardization was applied to normalize the scale of all features. This step is important because the original features have varying scales and units, which could bias certain algorithms that are sensitive to the magnitude of input values. Here transformation of each feature to have zero mean and unit variance is done. This confirms that all the features contribute equally to the model training process and prevents features with larger scales from dominating the learning process.

Additionally, performance of an exploratory data analysis to identify any potential outliers or anomalies in the dataset is made. While no significant outliers were detected that warranted removal, this step is important for understanding the distribution of our data.

Lastly, conducted a correlation analysis among the features to identify any highly correlated variables. Although we didn't remove any features based on correlation in this study, understanding these relationships is valuable for feature selection in future iterations in the model or for interpreting the status of different features in the diagnosis process.

### Train-Test\_Split:

The dataset was broken into training set, validation set and testing set. We allocated 70% of the data for training, 15% for validation and 15% for testing, a common split ratio in machine learning practices. This division allows us to train our models on a substantial portion of the data while reserving a substantial amount for unbiased evaluation. The split was stratified based on the target variable (diagnosis) to certify that both sets maintain the same proportion of malignant to benign cases as the original dataset. This stratification is crucial in maintaining the representativeness of both sets, especially important in medical diagnostics where class-balance can hugely impact model performance. A random state of 42 is set to ensure reproducibility of the split, allowing for the comparisons across different runs and model iterations.

### Feature Discretization:

Feature discretization was implemented, transforming continuous-features to discrete ones. We chose to use 10 bins with a uniform strategy, meaning each bin has approximately the same number of samples. This process was applied to both the training and testing sets, with the discretizer fitted only on the training-data to prevent data leakage. Discretization can help in reducing the impact of small variations and potential noise in the continuous data, potentially improving the performance of certain algorithms. It can also help in identifying non-linear relationships that might not be apparent in continuous data. However, it is essential to to keep in mind that discretization can lead to loss of information. We have considered this condition, concluding that for our specific case of breast cancer diagnosis, the potential benefits of improved model interpret-ability and reduced sensitivity to small variations outweighed the costs of information loss.

*3. Materials and methods 17*

### SMOTE (Synthetic Minority Over-sampling Technique):

To point the issues in the class-imbalance in our dataset, we decided to use a technique SMOTE. This is like a magic trick for creating synthetic samples of the minority class out of thin air. We applied SMOTE only to our training data so that our test data remained untouched and pure. The working of SMOTE is pretty clever: it collects examples that are close together in the feature space, then a line is drawn between them, and finally, it gives a new sample somewhere along that line. This technique helps balance out the dataset by promoting the number of minority class samples. By balancing the classes, we hope to improve our model's ability to learn from both classes, which will enhance its overall performance and reduce any bias towards the majority-class. This step certifies that our models are just as good at identifying benign cases as they are at catching malignant ones.

### Rescaling:

After applying SMOTE to address class imbalance, we performed rescaling on the resampled training data and the test data. We utilize StandardScaler to standardize the features. This step was crucial to ensure all features were on a similar scale, preventing features with greater magnitudes from ruling the analysis.

The StandardScaler was first fit on the resampled training data and then used to transform both the training and test sets. This process centered the feature distributions around zero with unit variance, calculated as (x - mean) / standard deviation. Rescaling was essential for our subsequent PCA and model training steps, as many ML algorithms are sensitive to the scale of input features. By standardizing, we ensured that all features contributed equally in the analysis and improved the numerical stability of our models.

### Principal Component Analysis (PCA):

We applied PCA to reduce the dimensionality of our data-set while retentive the most important information. Using sklearn's PCA module, we set n\_components=10, aim-ing to seizure the maximum significant variance in the data with fewer dimensions.

PCA was performed on the rescaled data, transforming our original 30 features into 10 principal components. This transformation not only reduced computational complexity but also mitigated potential overfitting by eliminating less informative features.

We analyzed the explained variance ratio to understand the standing of each principal component. The first component accounted for 42.22% of the total variance, with subsequent components contributing progressively less. We visualized the cumulative explained variance and created pairplots of the first 10 principal components to gain visions inside the data's structure in the reduced dimensional space. This PCA step was crucial for enhancing our model's efficiency and interpretability.

## Model training

Before subjecting the data to a model, the data was broken into training-set, test-set, and validation-set. The data was split into the 70, 15, and 15 ratios, respectively. Upon the data split, initially, the test dataset and extracted feature were trained using the built model.

The hyperparameters for:

### Logistic Regression:

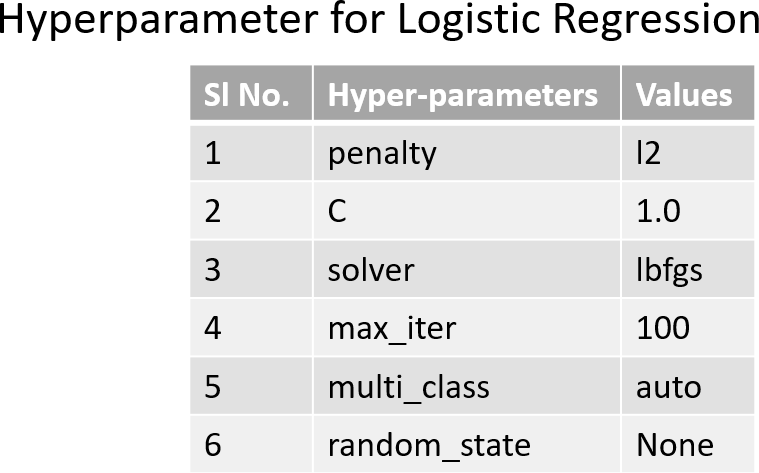


Table 3.1: Hyper-parameter for Logistic regression

### Linear Discriminant Analysis (LDA):

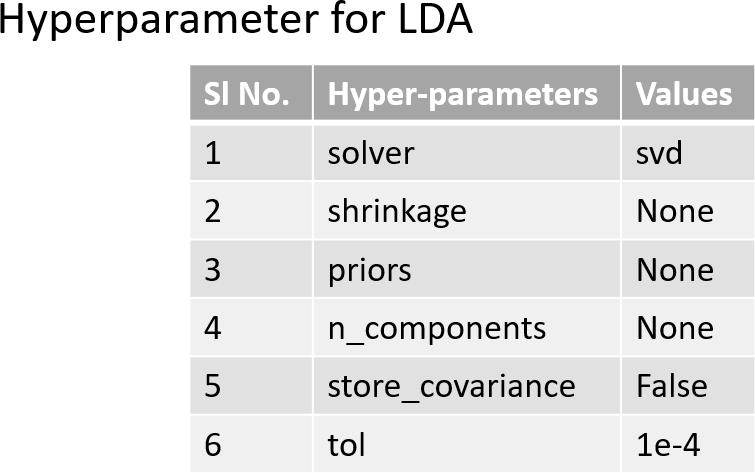


Table 3.2: Hyper-parameter for Linear Discriminant Analysis

### Decision Tree:

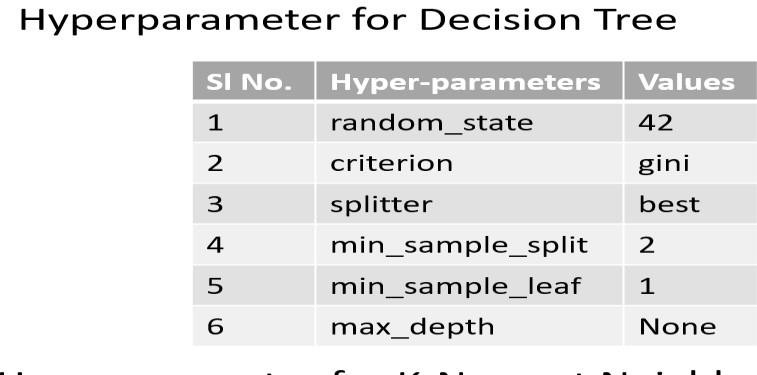


Table 3.3: Hyper-parameter for Decision Tree

### K-Nearest Neighbors (KNN):

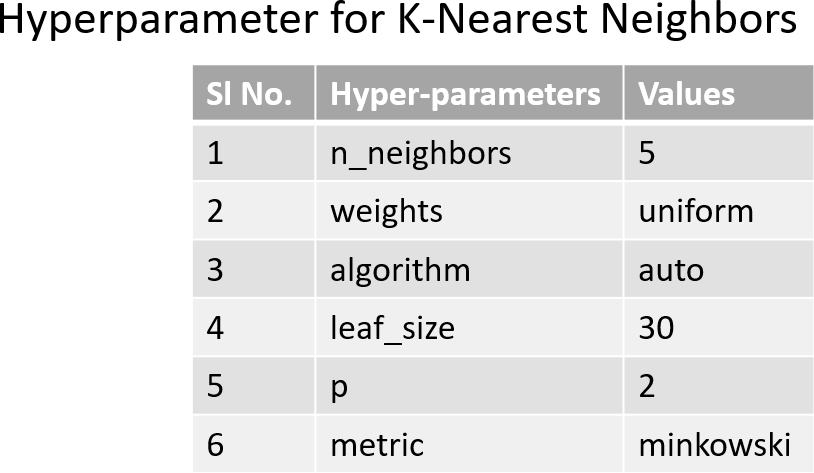


Table 3.4: Hyper-parameter for K-Nearest Neighbors

### Naïve Bayes:

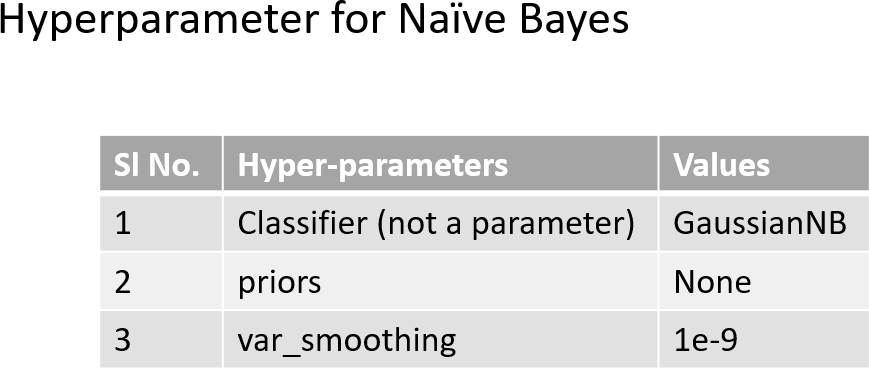


Table 3.5: Hyper-parameter for Naïve Bayes

### Support Vector Machine (SVM):

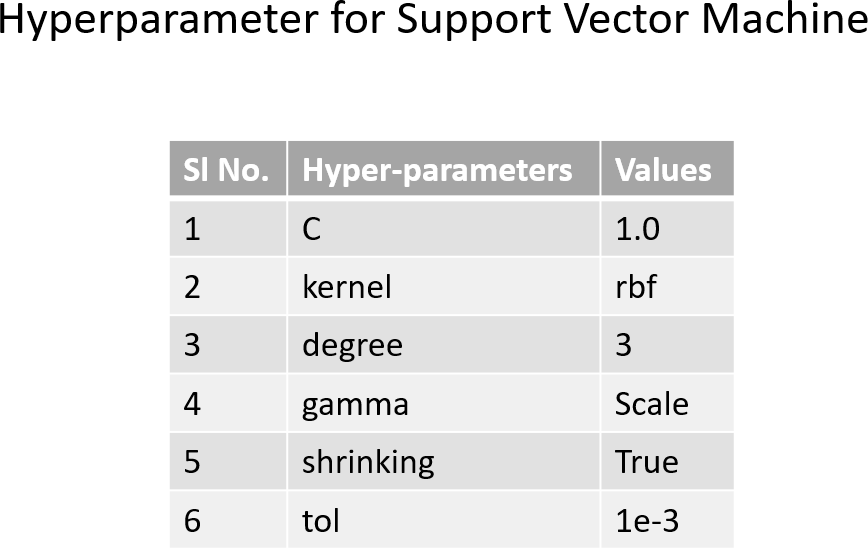


Table 3.6: Hyper-parameter for SVM

*20 4. Results and discussion*

# RESULTS AND DISCUSSION

Here, the results are interpreted and is discussed. The datasets are acquired, pre-processed, and subjected into to model training. Once the model was trained, they were evaluated for precision, sensitivity, accuracy and F1 score. The results of each step from data pre-processing till metrics are shown here.

## Data Pre-Processing:

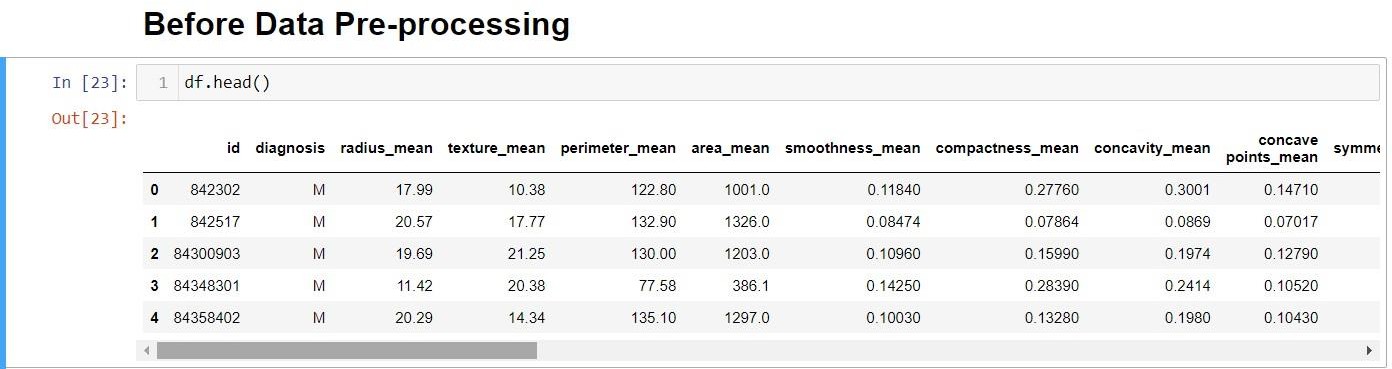


Fig. 4.1: Representation of Dataset before Pre-processing

This is how the dataset looks when visualized in python. There are features (columns) which will play an important role in our research. We are removing the column named ‘id’ since it has no impact here and we are embedding the ‘diagnosis’ column from ‘M’ and ‘B’ to ‘1’ and ‘0’ (binary classification).

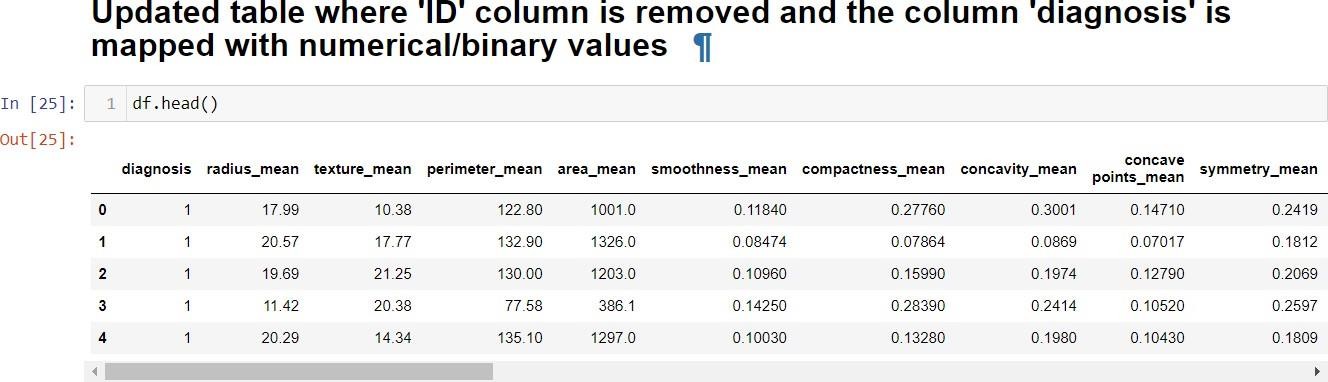


Fig. 4.2: Representation of Dataset after encoding and removal of column ‘ID’

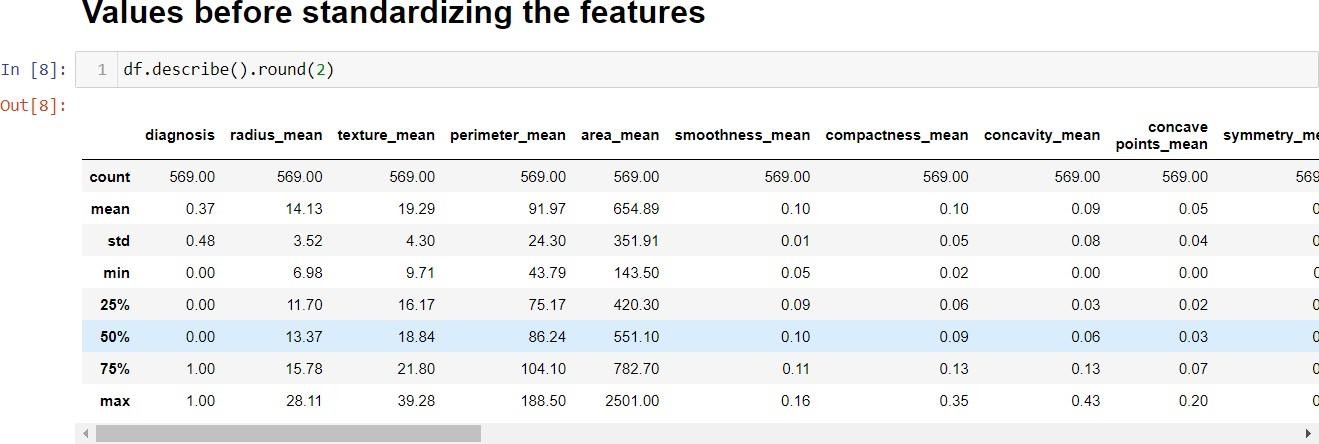


Fig. 4.3: Dataset before feature standardization

After standardizing the features, the mean and standard deviation of each features will be standardized to 0 and 1 respectively. The Fig 4.3 clearly shows the original mean and standard deviation before standardizing. The table 4.1 shows the updated mean and standard deviation after application of feature standardization.

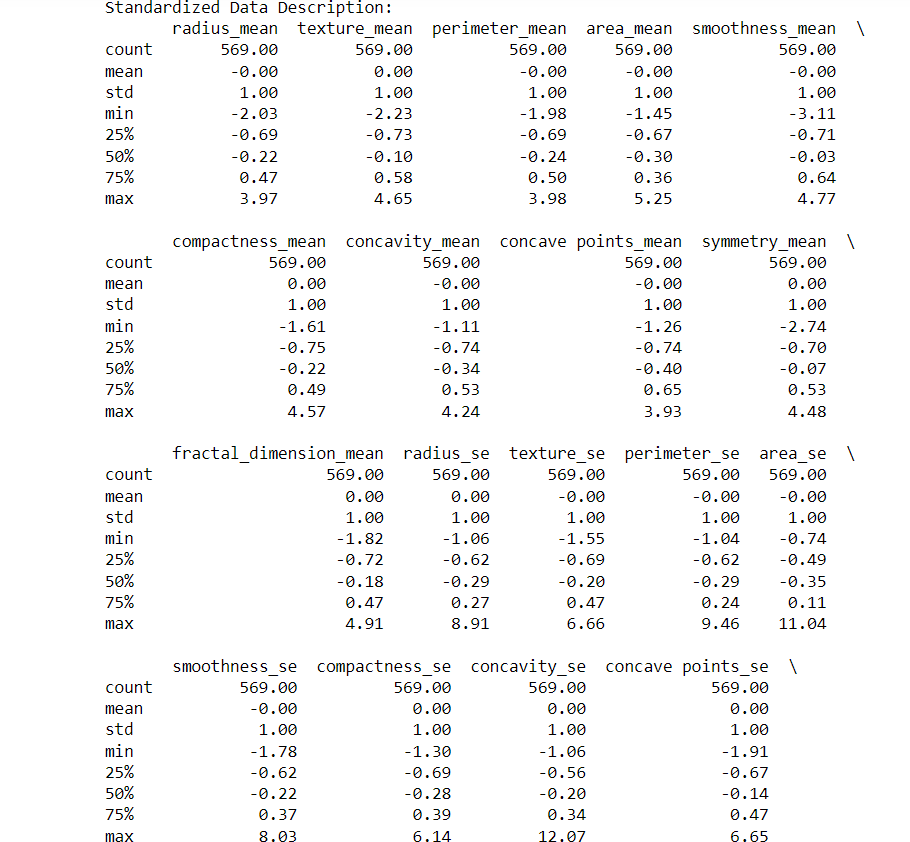


Table 4.1: Standardized features

## Train-test Split:

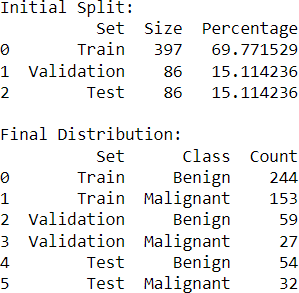


Table 4.2: Standardized feature

Here, we are partitioning our data into 3 parts: Training set, validation set and test set. There are 569 instances. Out of which, we are splitting them in the ratio of 70:15:15 (train: validation: test). The table 4.2 shows a proper description of each sets and the type of cancer class present in each set. Fig 4.3 gives a visual representation of the partitioned data.

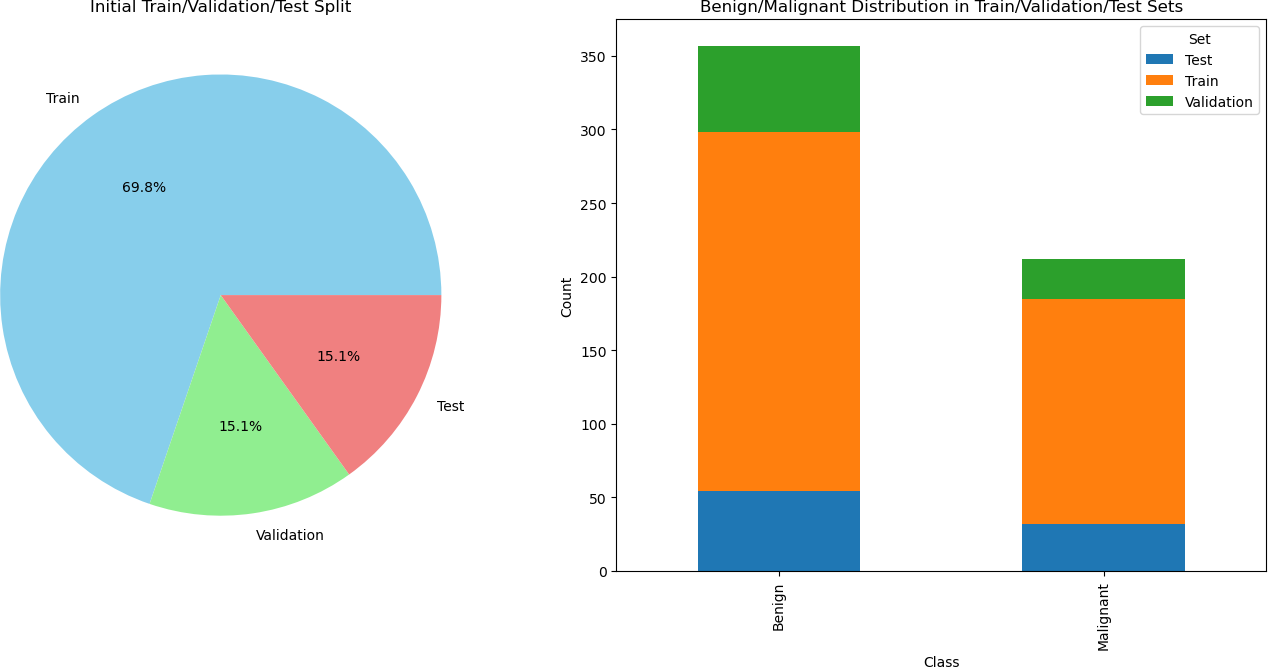


Fig. 4.3: Pie-graph (Left picture) shows train: validation: test ratio and Bar-graph (Right picture) shows the count of the respective classes present in Test set (blue), Train set (orange) and Validation set(green)

## Feature Discretization:

The data is discretized meaning, the continuous features are now transformed to discrete features. This process was applied to reduce the noise and data leakage. We chose to use 10 bins with a uniform strategy, meaning each bin has approximately the same number of samples. This process was applied to both the training and testing sets, with the discretizer fitted only on the training-data to prevent data leakage.

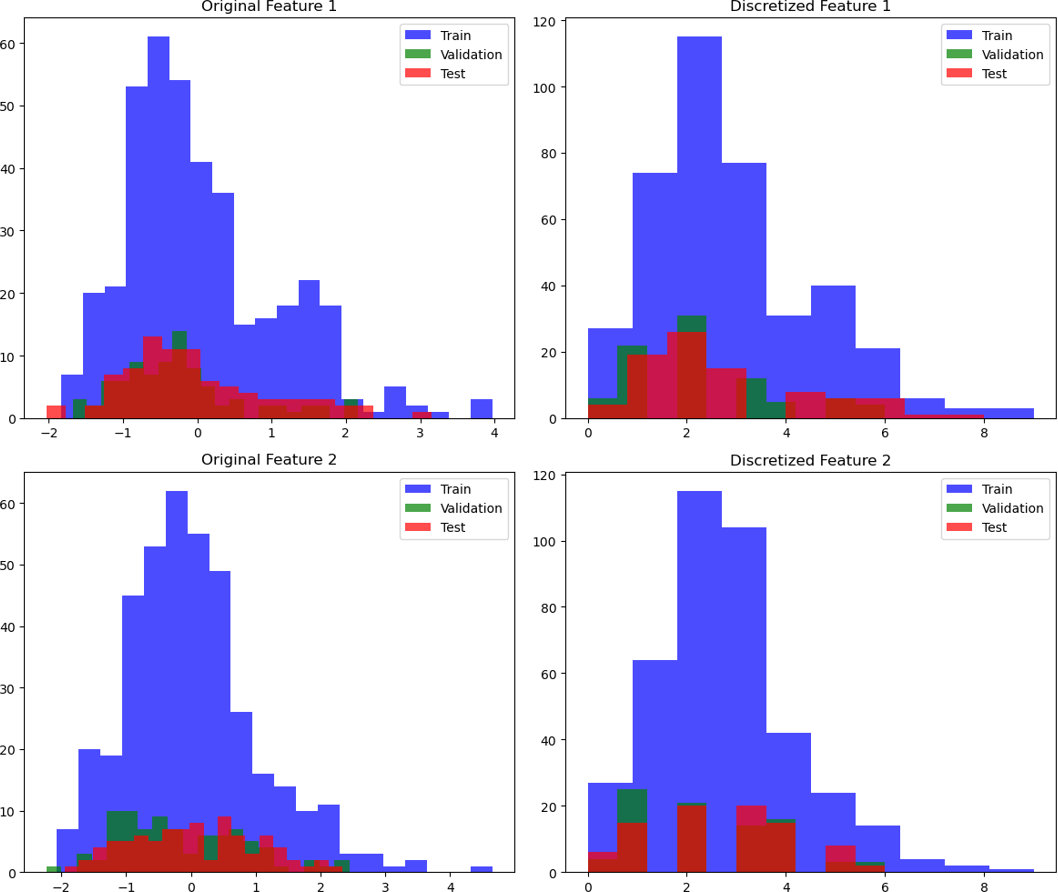


Fig. 4.4: Image represents two columns: Column 1 is the Histogram of the original feature values for the training (blue), validation (green), and test sets (red). Column 2 is the Histograms of the discretized feature values for the training (blue), validation (green), and test (red) sets

# SMOTE:

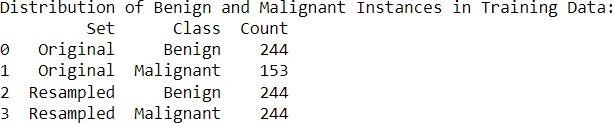


Table 4.3: Distribution of Benign and Malignant in training data

Initially, the complete dataset comprising 569 instances was made into training-set, validation-set, and test-set. Following an 85/15 split, where 85% of data was kept for training and validation, the training data itself was further divided. Specifically, 70% of the original 569 instances, amounting to approximately 398 instances, were allocated for training purposes. This subdivision was vital for implementing SMOTE effectively, as it allowed for detailed model training while maintaining a separate validation subset for fine-tuning and assessing model performance. The remainder of the data, including the validation set of about 86 instances and a distinct test set of approximately 85 instances, ensured a comprehensive evaluation framework for validating the efficacy of the SMOTE technique in handling imbalanced datasets.

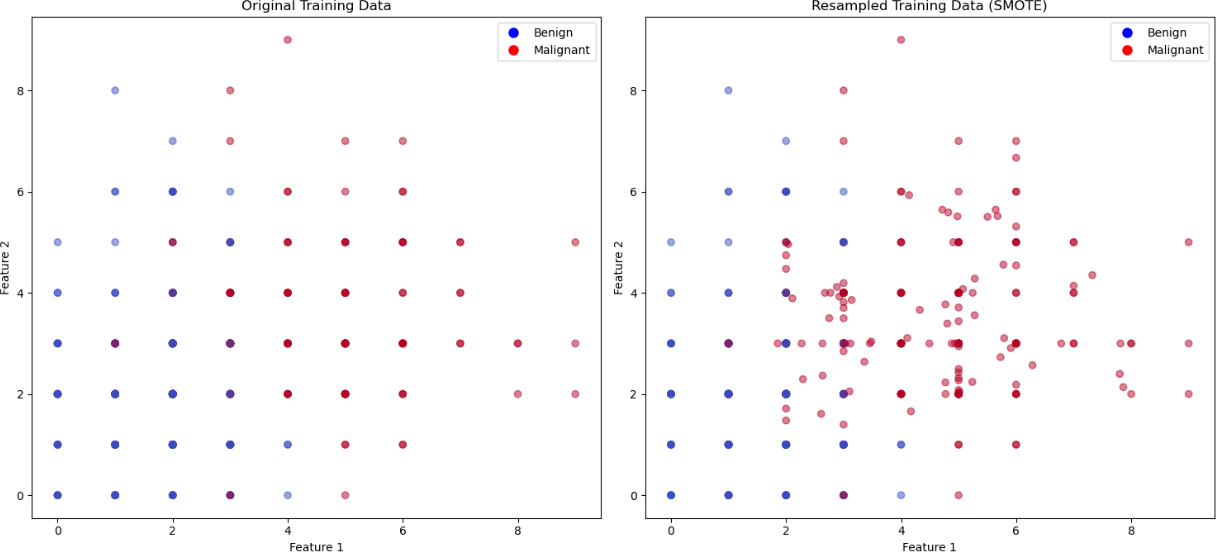


Fig 4.5: Scattered plot of Original training data (left-side) and resampled training data (SMOTE) (right-side) points out the benign (blue) and malignant (red) datapoints

## Feature Rescaling

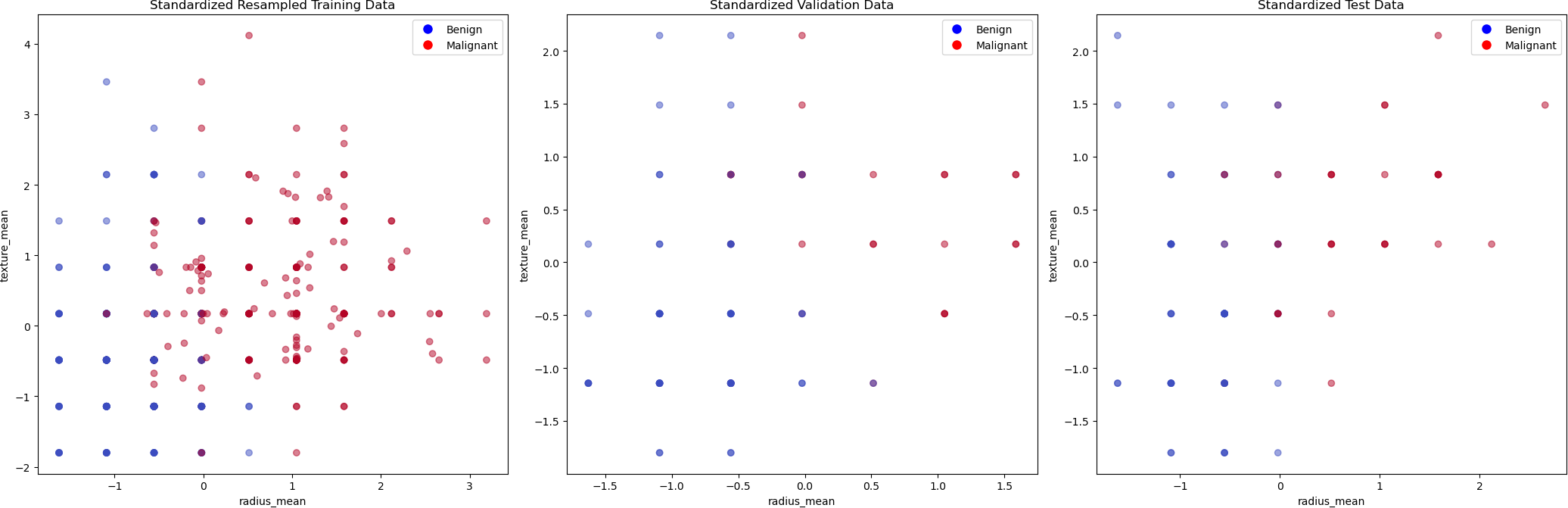


Fig 4.6: Scattered plot for the first two standardized features for the resampled training data (left), validation (middle) and test data (right)

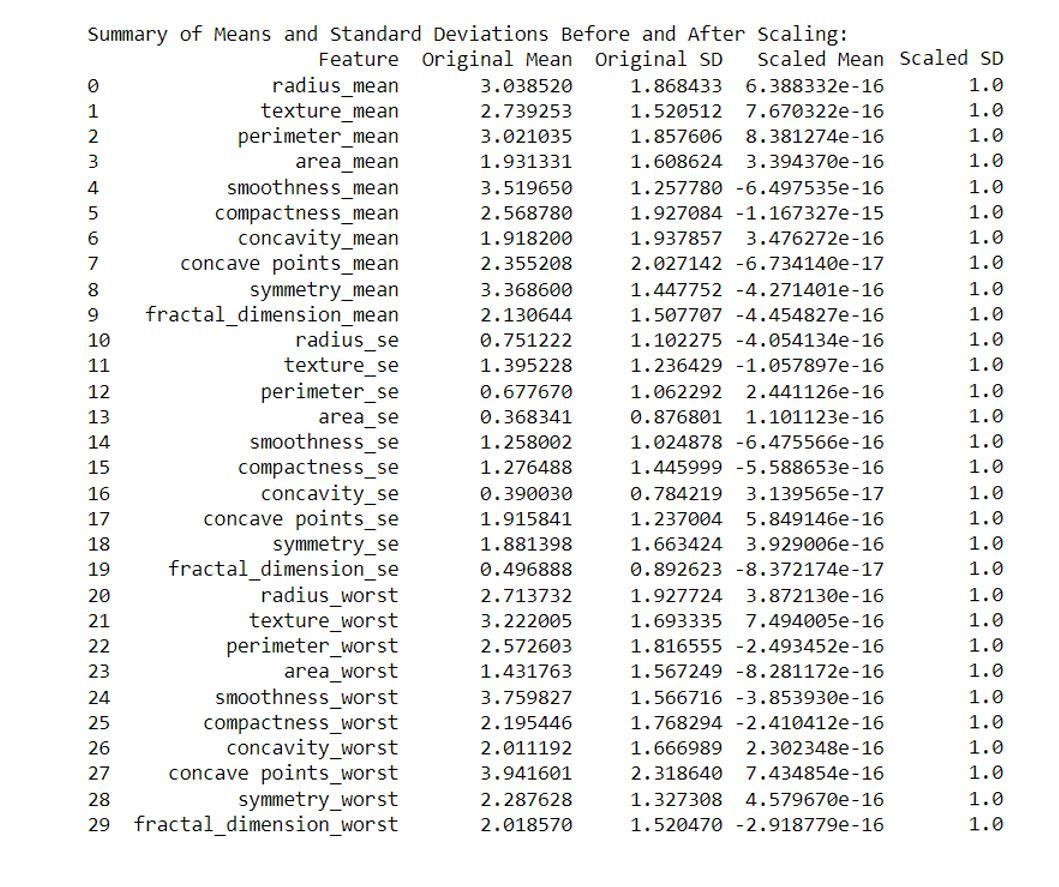


Table 4.4: Table which shows the rescaled features where the mean is zero and standard deviation is one

## Model Training and Evaluation:

Logistic Regression astounded everybody by being correct 97.67% of the time with a precision, recall, and F1 score of 0.9688 everything. This shows that it was generally excellent while maintaining a balance between the two.

Linear Discriminant Analysis had an accuracy rate of 96.51%, its precision being perfect at 1.0000; however, the recall dropped slightly to 90.62% hence the F1 score standing at 0.9508. This tells us that although good at reducing false alarms, it failed to capture all the actual ones.

Decision Tree had 93.02% accuracy, where its recall amounted to 87.50% with precision being 93.33% resulting into a F1 score of 90.32. While this model performed relatively well, there were more false-negatives than in others.

KNN had an accuracy rate of 0.9535 with a precision and recall of both at 0.9375, thus giving an F1 score of 0.9375 consistently. The precision and recall are well balanced in this model.

Naive Bayes achieved 0.9186 accuracy, while the precision was 0.9032 and recall 0.8750 which led to F1 score being 0.8889; the model worked effectively except it had high false-positive rate with slightly lower precision.

Support Vector-Machine (SVM) displayed 0.9651 accuracy where its precision stood at 0.9677 and recall 0.9375, hence an F1 score of 0.9524. This shows that SVM model has good recall although the precision is very high indicating reliable performance.

Before going into the evaluation part. Let us see the features correlations in the dataset in the form of heatmap.

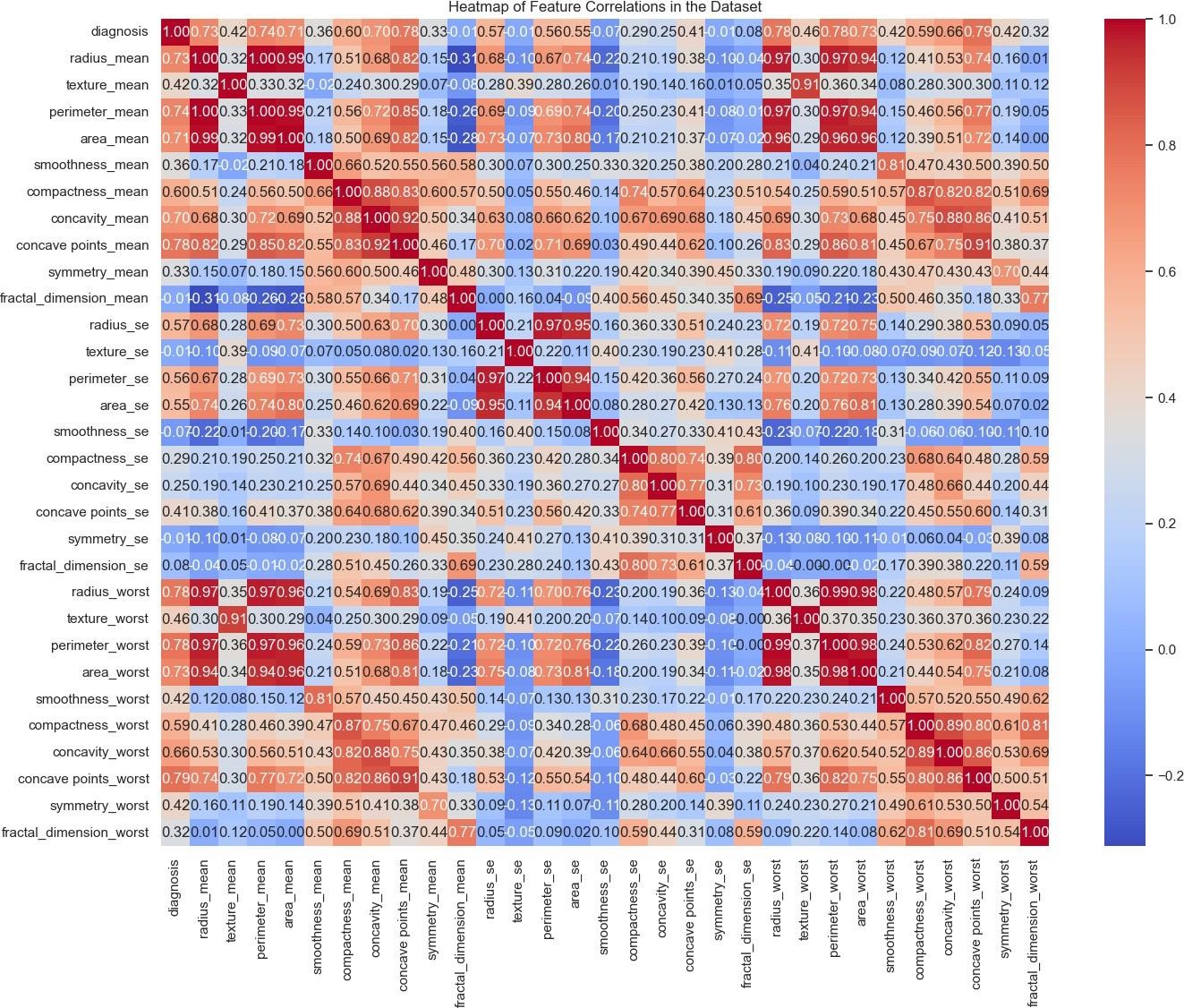


Fig 4.7: Heatmap of feature correlation

### Evaluation:

**Logistic regression:**

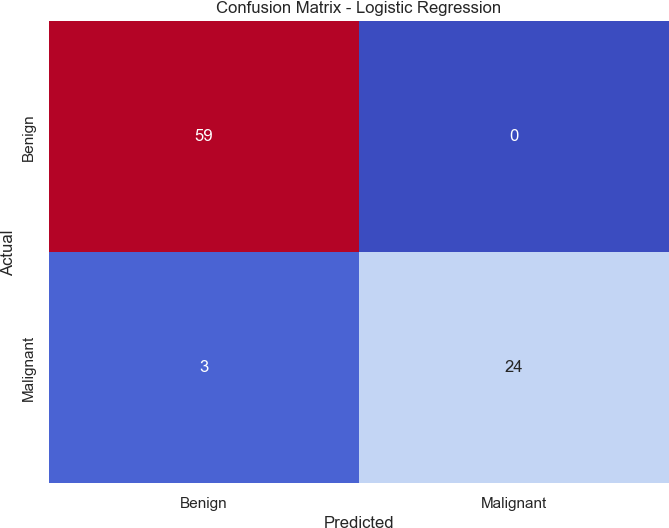


Fig 4.8: Confusion-matrix for Logistic regression

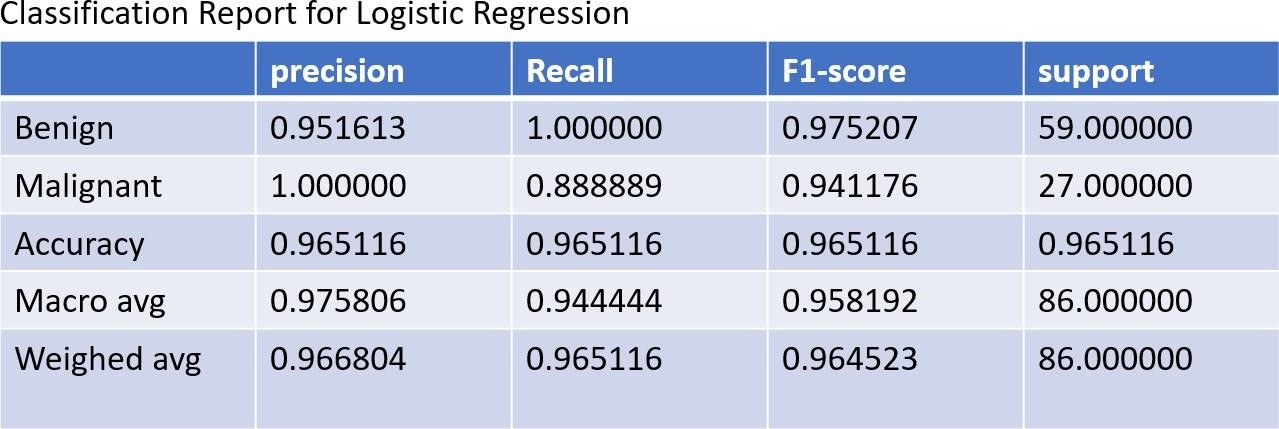


Table 4.5: Classification-report for logistic regression

### Linear Discriminant Analysis:

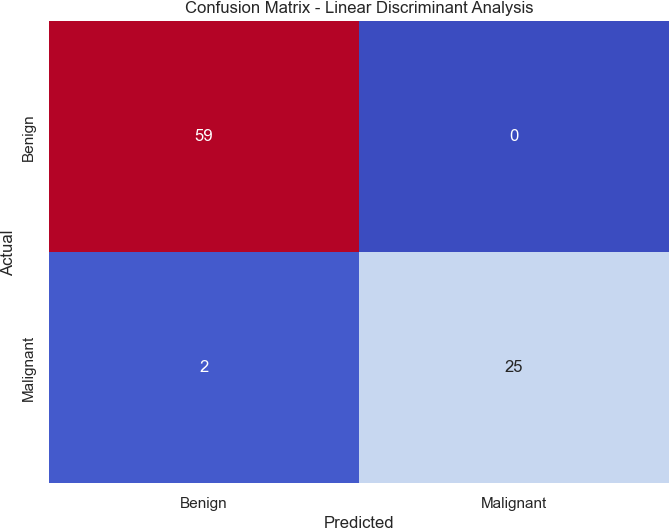


Fig 4.9: Confusion-matrix for LDA

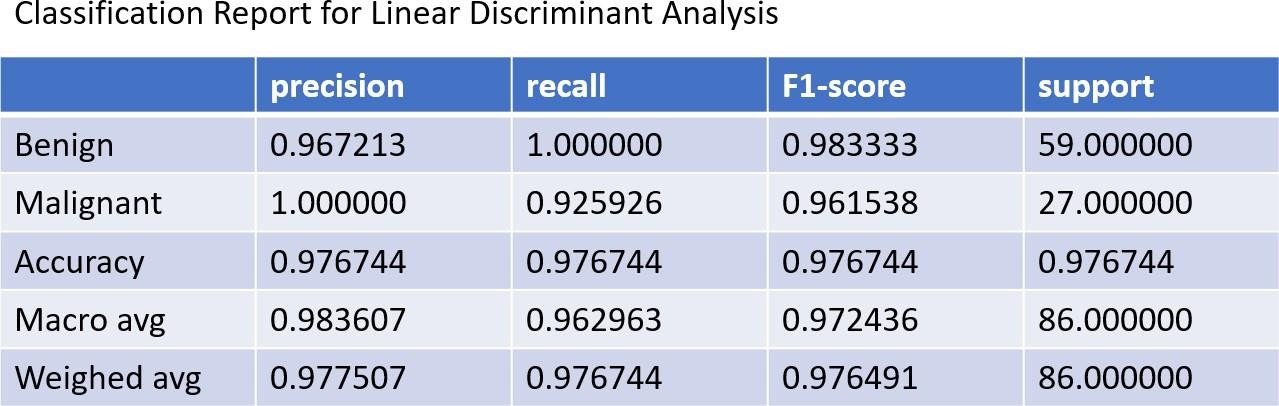


Table 4.6: Classification-report for LDA

### Decision Tree:

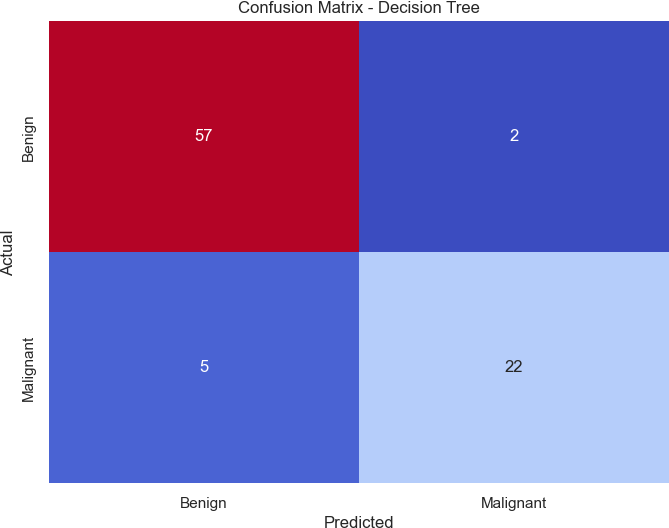


Fig 4.10: Confusion-matrix for Decision Tree

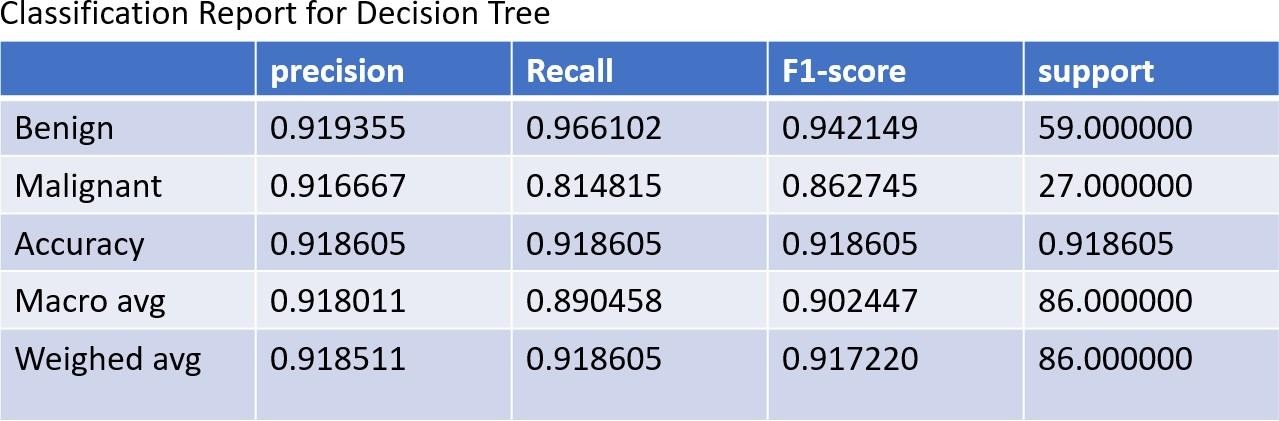


Table 4.7: Classification-report for Decision Tree

### K-Nearest Neighbours:

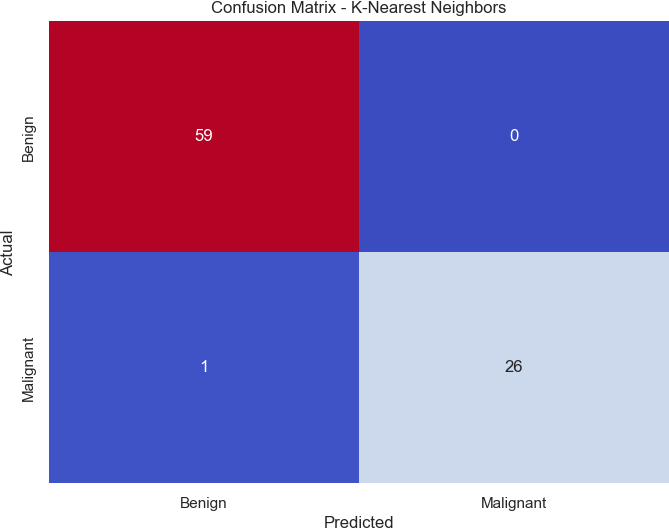


Fig 4.11: Confusion-matrix for KNN

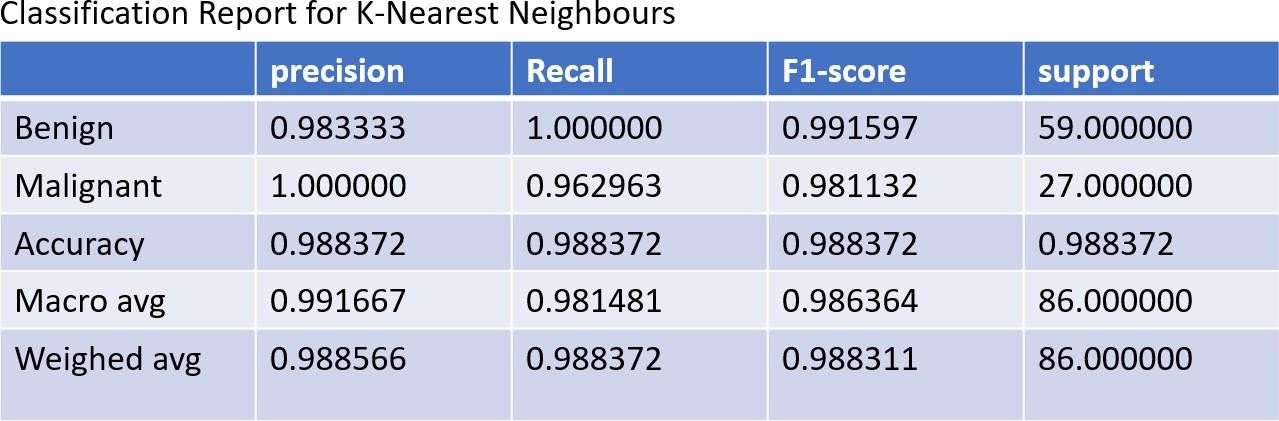


Table 4.8: Classification-report for KNN

### Naïve Bayes:

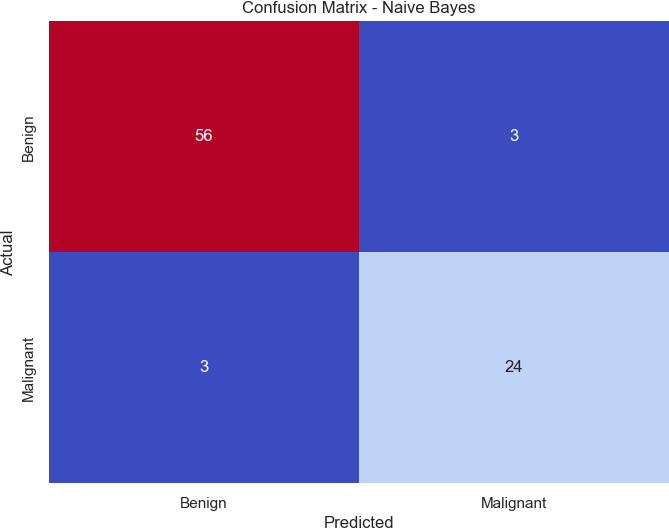


Fig 4.12: Confusion-matrix for Naïve Bayes

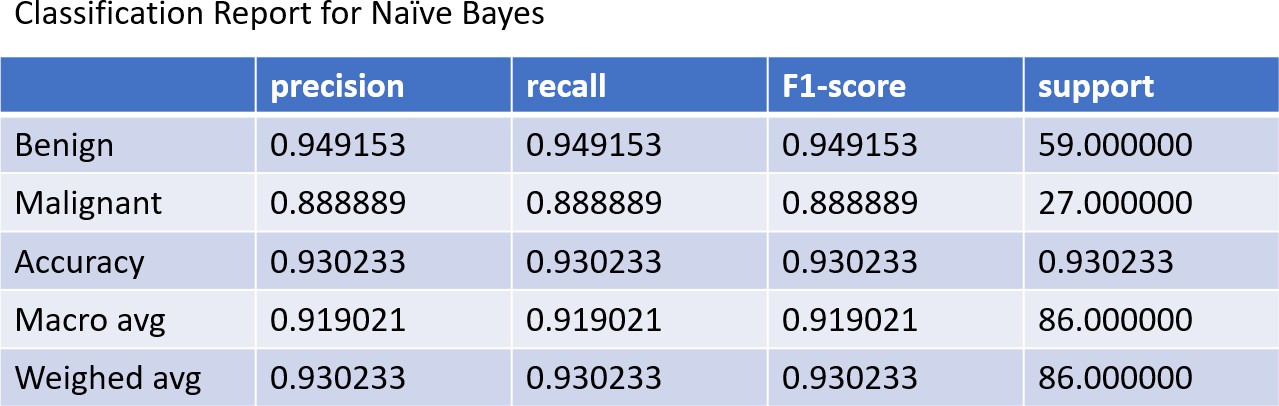


Table 4.9: Classification-report for Naïve Bayes

### Support Vector Machine:

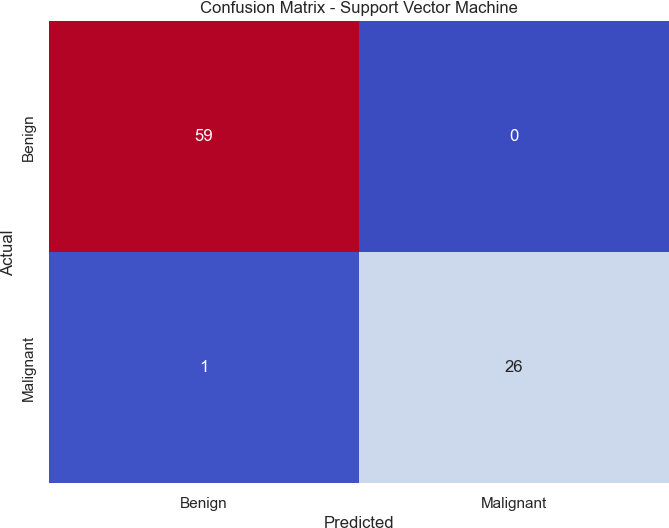


Fig 4.13: Confusion-matrix for SVM

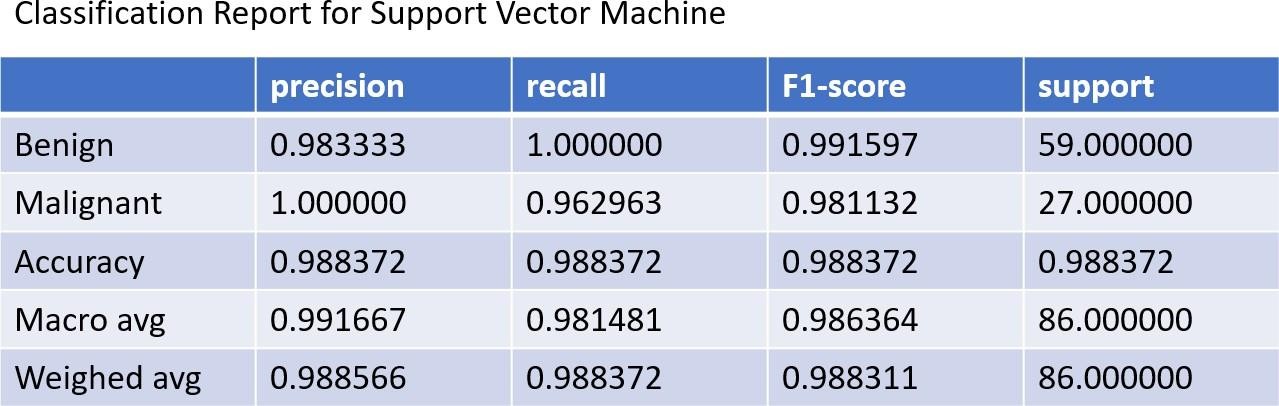


Table 4.10: Classification-report for SVM

### Final Evaluation on Test Set:

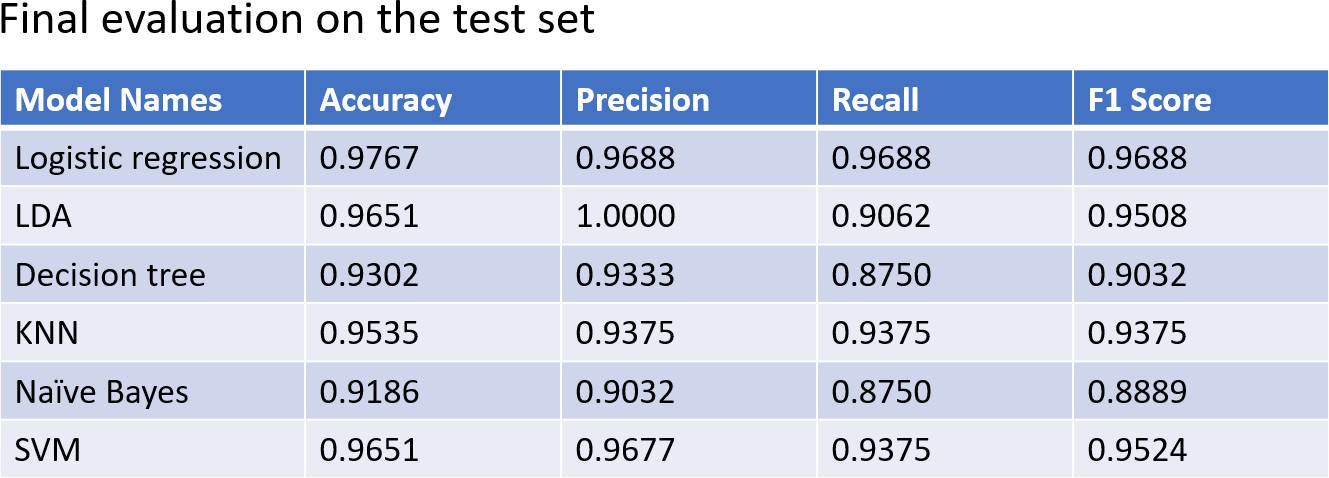


Table 4.11: Final evaluation metrices on the test set

This research presents a comprehensive examination of six machine learning (ML) models for breast cancer detection: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Decision Tree (DT), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM). Each model was evaluated using precision, recall, accuracy, and F1 score metrics. Initially, the dataset was visualized and cleaned by removing the 'ID' column and encoding the 'diagnosis' column from categorical to binary values. Feature standardization was applied to normalize the dataset, ensuring all features had a mean of zero and a standard deviation of one. This was crucial for eliminating biases due to varying scales of features. The dataset was then split into training (70%), validation (15%), and testing (15%) sets, maintaining the same proportion of malignant to benign cases, ensuring representativeness and minimizing performance impact due to class imbalance

This process involved transforming continuous features into discrete ones using 10 bins with a uniform strategy. This step helped reduce noise and potential data leakage, simplifying the models and improving their ability to handle non-linear relationships and outliers. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied to create synthetic samples of the minority class. This improved the model’s ability to learn from both classes and reduced bias towards the majority class. The training data was rescaled post-SMOTE to ensure all features were on a similar scale. Principal Component Analysis (PCA) was then used to reduce dimensionality, transforming the original 30 features into 10 principal components, which explained approximately 85% of the variance. This step enhanced computational efficiency and mitigated overfitting by removing less informative features.

The models were trained and assessed. LR achieved the highest accuracy of 97.67%, with precision, recall, and F1 scores of 0.9688. LDA followed with 96.51% accuracy but had a slightly lower recall, affecting its F1 score. KNN showed balanced precision and recall with 95.35% accuracy. NB and DT had relatively lower accuracy rates of 91.86% and 93.02%, respectively, with NB exhibiting a higher false-positive rate. SVM demonstrated high precision and recall, achieving 96.51% accuracy. A heatmap of feature correlations provided insights into the interrelationships among features. The final evaluation confirmed that LR, KNN, and SVM models maintained high performance (recall) on the test set, emphasizing their potential for improving breast cancer detection accuracy. LR achieved 97.67% accuracy on the test set, highlighting its robustness

*5. Conclusion\_and\_Future\_Prospectives*

# CONCLUSION AND FUTURE PROSPECTIVES

So, to sum it up, the research has successfully shown the great ability of machine learning algorithms to improve breast cancer detection. The thorough examination of six different ML models LR, LDA, DT, KNN, NB, and SVM on the dataset has elicited hopeful results that may potentially affect early cancer diagnosis and patient care. Our results indicate that the use of ML models such as LR, KNN, and SVM leads to highly accurate, precise, and recall in the detection of breast cancer. The high level of performance of LR, which has 97.67% accuracy on the test set, points to the potential of such methods to be efficient tools in the medical diagnostic process. The excellent performances of KNN and SVM, in both validation and testing, additionally substantiate the soundness of the ML approaches in healthcare application.

The excellent performance of our preprocessing techniques, which include standardization, discretization, and the use of SMOTE for class balancing, shows how important proper data treatment is for making the ML models work effectively. Besides it, the use of PCA has been a good example of how this statistical method can be used to reduce the dimensionality and distill the medical data being an important part with minimal power loss. These are also, optimistic results, however, the formation of a new vision on machine learning as a supplementation to rather than substitution for the traditional diagnostic methods is the base for the further identification of the risks and the like. The far-reaching potential of these algorithms actually results from the fact is that they can work together with healthcare professionals, providing another layer of knowledge to the whole process of diagnosis.

Henceforth, many headways will be made in research. More fidelity could be obtained by using ensemble methods alone possibly. Moreover, employing some of the test datasets from various societies during the model deployment can help the researcher validate the robustness of their procedures in real-world applications. On general level, our research outputs a possible way in which a natural Canberra inheritance could be developed. As they become more and more complex, these innovations are beginning to offer us the potential of better disease detection, and reducing costs, and ultimately advancing patient health. Essentially, this research's objective is more than only data and algorithms. It is about using technology in order to rescue people and provide them with hope that they are confronting the situation of the breast, thus, getting closer step by step to a future when this disease no more will be such a significant hazard to women's health throughout the world.

*References 33*

**REFERENCES**

[1] Silva J, Lezama OB, Varela N, Borrero LA. Integration of data mining classification techniques and ensemble learning for predicting the type of breast cancer recurrence. InGreen, Pervasive, and Cloud Computing: 14th International Conference, GPC 2019, Uberlândia, Brazil, May 26–28, 2019, Proceedings 14 2019 (pp. 18-30). Springer International Publishing.

[2] Hazra A, Mandal SK, Gupta A. Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms. International Journal of Computer Applications. 2016 Jul;145(2):39-45.

[3] Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016 Jan 1;83:1064-9.

[4] Chaurasia DV, Pal S. A novel approach for breast cancer detection using data mining techniques. International journal of innovative research in computer and communication engineering (An ISO 3297: 2007 Certified Organization) Vol. 2017 Jun 29;2.

[5] Bazazeh D, Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In2016 5th international conference on electronic devices, systems and applications (ICEDSA) 2016 Dec 6 (pp. 1-4). IEEE.

[6] Shanbehzadeh M, Kazemi-Arpanahi H, Ghalibaf MB, Orooji A. Performance evaluation of machine learning for breast cancer diagnosis: A case study. Informatics in Medicine Unlocked. 2022 Jan 1;31:101009.

[7] Houssein EH, Emam MM, Ali AA, Suganthan PN. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. Expert Systems with Applications. 2021 Apr 1;167:114161.

[8] Kadhim RR, Kamil MY. Comparison of machine learning models for breast cancer diagnosis. IAES International Journal of Artificial Intelligence. 2023 Mar 1;12(1):415.

[9] Jafari A. Machine-learning methods in detecting breast cancer and related therapeutic issues: a review. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2024 Jan 13:1-1.

[10] Mridha MF, Hamid MA, Monowar MM, Keya AJ, Ohi AQ, Islam MR, Kim JM. A comprehensive survey on deep-learning-based breast cancer diagnosis. Cancers. 2021 Dec 4;13(23):6116.

[11] Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN computer science. 2021 May;2(3):160.

[12] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nature reviews Molecular cell biology. 2022 Jan;23(1):40-55.

*References 34*

[13] Tashakkori R, Mozdgir A, Karimi A, BozorgzadehVostaKolaei S. The prediction of NICU admission and identifying influential factors in four different categories leveraging machine learning approaches. Biomedical Signal Processing and Control. 2024 Apr 1;90:105844.

[14] Soomro AA, Mokhtar AA, Salilew WM, Abdul Karim ZA, Abbasi A, Lashari N, Jameel SM. Machine Learning Approach to Predict the Performance of a Stratified Thermal Energy Storage Tank at a District Cooling Plant Using Sensor Data. Sensors. 2022 Oct 10;22(19):7687

[15] Ozer ME, Sarica PO, Arga KY. New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines. Omics: a journal of integrative biology. 2020 May 1;24(5):241-6.

[16] Suresh S. *Modeling pre-tactical Air Traffic Flow Management decision processes with Maschine Learning/Deep Learning approaches* (Doctoral dissertation, TU Hamburg).

[17] Hulleman WG, Vos RA. Modeling abiotic niches of crops and wild ancestors using deep learning: A generalized approach. BioRxiv. 2019 Oct 31:82634

