A

Project Report

on

Prognosis of Breast Cancer using Machine Learning Algorithms

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2022-23

in

Computer Science and Engineering

By

Shivangi Gaur(1900290100144)

Vasu Bansal(1900290100186)

Under the supervision of

Dr. Seema Maitrey

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2023

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature

Name: Vasu Bansal

Roll No.:1900290100186

Date: 29/05/2023

Signature Shivangi

Name: Shivangi Gaur

Roll No.:1900290100144

Date: 29/05/2023

CERTIFICATE

This is to certify that Project Report entitled “Prognosis of Breast Cancer using Machine Learning Algorithms” which is submitted by Shivangi Gaur(1900290100144) and Vasu Bansal(1900290100186) in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

.

Date: 29/05/2023 **Supervisor Name : Dr. Seema Maitrey**

(Associate Professor)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Dr. Seema Maitrey, Department of Computer Science & Engineering, KIET, Ghaziabad, for her constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only her cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially Dr. Dharmendra Kumar, prof. Vipin Deval and Dr. Dilkeshwar Pandey, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Date: 29/05/2023
Signature: Shivangi
Name : Shivangi Gaur
Roll No.: 1900290100144

Date: 29/05/2023
Signature:
Name : Vasu Bansal
Roll No.: 1900290100186

ABSTRACT

Breast cancer is a type of tumor that develops in the breast tissues. It is the most common type of cancer found in women worldwide, and it is one of the leading causesof death in women. This article provides a comparison of machine learning, deep learning, and data mining techniques used for breast cancer prediction. Many researchers have worked on breast cancer diagnosis and prognosis; each technique has a different accuracy rate, which varies depending on the situation, tools, and datasets used. Our primary goal is to compare various existing Machine Learning and Data Mining techniques to identify the most appropriate method for supporting largedatasets with high prediction accuracy.

Our focus is to comparatively analyze different existing Machine Learning and Data Mining techniques in order to find out the most appropriate method that will support the large dataset with good accuracy of prediction.

The main purpose of this project is to analyze all the existing studies of Machine Learning algorithms that are being used for breast cancer prediction and this project provides the all necessary information to the beginners who want to analyze the machine learning algorithms to gain the base of Deep Learning.

The future study will focus that allows users to calculate their risk of developing breast cancer.

TABLE OF CONTENTS

Page No.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML	Machine Learning
ANN	Artificial Neural Network
CNN	Convulational Neural Network
SVM	Support Vector Machine
RF	Random Forest
KNN	K-Nearest Neighbor
DT	Decision Tree
LR	Logistic Regression
NB	Naïve Bayes

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

In the modern day, breast cancer is one of the most deadly and diverse diseases, killing a vast number of women all over the world. In the United States, 281,550 new cases were diagnosed with breast cancer, and 43,600 deaths were reported in the females during 2021. Total number of women dying in 2021 is approximately 963,000, according to the World Health Organization (WHO), Still, the organization predicts that the number could reach 2.9 million globally. According to the World Health Organization, Breast cancer (BC) constitutes the first major cause of women's death.

The indications and manifestations of breast cancer include Breast cancer symptoms can vary among individuals. While some may not experience any symptoms, others may notice a new lump in the breast or underarm. Additional indications of breast cancer may involve enlargement or hardening of a specific breast area, puckering or inflammation of the breast skin, or the appearance of reddened or cancer scaly skin surrounding the breast or nipple, nipple soreness or pulling inwards, discharge from the nipple that is not breast milk, any changes in the size or shape of the breast, and breast pain in any area.

Machine learning [2] and data mining methods are methods are utilized to predict breast cancer, and Identifying the appropriate algorithm for the job is the obstacle to overcome. Breast cancer emerges from malignant tumors that arise when cellular growth becomes unmanageable, resulting in a typical multiplication of adipose tissue.

1.2 THE SIGNS AND SYMPTOMS OF BREAST CANCER

Distinct people will experience different breast cancer symptoms.

- Some individuals have no symptoms whatsoever.
- A new breast or underarm lump (armpit).
- Swelling or thickening of a portion of the breast.

- Dimpling or irritation of the breast skin.
- Redness or flaky skin around the breast or nipple.
- Nipple pulling in or soreness in the nipple region.
- Navel discharge, including blood and substances other than breast milk.
- Any modification to the breast's size or form.
- Breast pain in any location.

1.3 VARIOUS TYPES OF BREAST CANCER

The prediction of breast cancer uses a variety of machine learning and data mining methods. One of the key tasks is to identify the most relevant and suitable algorithm for breast cancer prediction. Malignant tumors, which form when a cell's development spirals out of control, are the cause of breast cancer. Breast cancer is brought on by the abnormal proliferation of numerous fatty and fibrous breast tissues. The tumors were infected with cancer cells differently and it produce various types of breast cancer.

- DCIS, commonly referred to as non-invasive cancer, is a kind of breast cancer that develops when abnormal cells move outside the breast.

- Infiltrative ductal carcinoma (IDC) and invasive ductal carcinoma are the two terms used to describe the second kind. IDC cancer is typically observed in men, and it develops when breast aberrant cells expand throughout all breast tissues.

- The third subtype of breast cancer is known as Mixed Tumors Breast Cancer (MTBC), which is also referred to as invasive mammary breast cancer. Such cancers are brought on by abnormal duct and lobular cells.

- Lobular Breast Cancer (LBC) is the fourth form of cancer and develops inside the lobule. It raises the risk of developing more invasive malignancies.

- Colloid breast cancer, also known as mucinous breast cancer (MBC), is the fifth kind of breast cancer that arises from invasive ductal cells.

- When aberrant tissues surround the duct, it happens. Inflammatory Breast Cancer (IBC) is last type that causes swelling and reddening of breast. It is a fast-growing breast cancer, when the lymph vessels block in break cell, this type of cancer starts to appear.

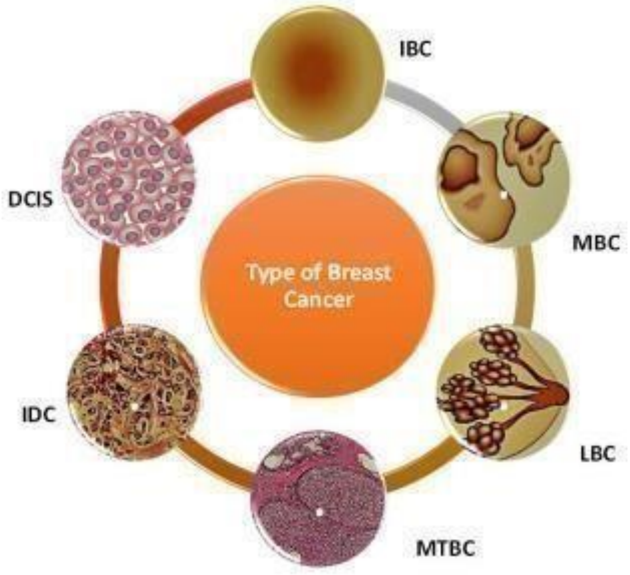


Fig. 1.1 Types of breast cancer

Finding relevant information from a large dataset is a process known as data mining. Data mining functions and techniques can be used to identify any type of disease. For example, machine learning, statistics, databases, fuzzy sets, data warehouses, and neural networks can be used to diagnose and predict the prognosis of various cancer diseases, including prostate cancer, lungs cancer, and leukemia.

The traditional approach to cancer identification is based on "the gold standard" approach, which entails three tests: a clinical examination, radiological imaging, and a pathology examination. While the modern machine learning approaches and algorithms are based on model creation, the traditional method identifies the existence of cancer and is based on regression process.

The purpose of the model is to forecast gives the good, anticipated result in their training and testing stages using previously unknown data. Preprocessing, features selection or extraction, and classification are the three primary methodologies on which machine learning is founded. The main component of machine learning, feature extraction, aids in the diagnosis and prognosis of cancer and may distinguish between benign and malignant tumors.

1.4 PROJECT DESCRIPTION

Our goal is to use machine-learning algorithms to predict and diagnose breast cancer, and to determine which are the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision, and sensitivity.

CHAPTER 2

LITERATURE REVIEW

This section discusses some of the previous work on breast cancer diagnosis done by researchers using various machine learning approaches.

Table 2.1 show some of the related works previously done on breast cancer diagnosis by researchers using different machine learning approaches, deep Learning are discussed.

Table 2.1 Review Papers

Reference	Study	Published Year
Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques	Proposed BCAD Framework, With a 99.12% accuracy, the Multilayer Perceptron Model.	2022
Breast cancer detection using artificial intelligence techniques: A systematic literature review	Review of Techniques based on gene sequencing and MRI imaging on diverse sources of data.	2022
Supervised Classification using Gradient Boosting Machine: Wisconsin Breast Cancer Dataset	The GB based model provides great degree of accuracy in addition to faster classification using lesser computing power compared to deep learning models.	2020

Assessment of deep learning algorithms to predict histopathological diagnosis of breast cancer: first Moroccan prospective study on a private dataset	This proposed a simple and effective method for the classification of HE stained histological BC images in case of very small training data (328 samples).	2022
Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction	In these 5 classification models are used SVM, NN, RF, DT LR. Random forest gives the best results	2018
Predicting factors for survival of breast cancer patients using machine learning techniques	lowest obtained from decision tree (accuracy = 79.8%) and the highest from random forest (accuracy = 82.7%)	2019
Prediction of Breast Cancer Disease using Machine Learning Algorithms	In this work, K-Nearest Neighbor, Support Vector Machines, Logistic Regression, Naive Bayes, Random Forest, and Random Forest classifier demonstrated best outcomes regarding precision and least execution time.	2020

Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis	Researchers can solve the issue of limited available dataset by using some data augmentation techniques. The issue of inequality of positive and negative data should be considered by researchers as it can lead to biasness towards positive or negative prediction .	2020
BREAST CANCER PREDICTION USING MACHINE LEARNING	The accuracy obtained by SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN	2020
Prediction of Breast Cancer Using Supervised Machine Learning Techniques	all the three applied algorithms Support Vector Machine, k Nearest Neighbor and Logistic Regression, SVM gives the highest accuracy of 92.7%	2019
Breast Cancer Prediction Using Machine Learning	KNN has highest accuracy among 4 algorithms, Logistic Regression, Support Vector Machine and Naive Bayes.	2020

An Analysis on Breast Disease Prediction Using Machine Learning Approaches	These are SVM, NB, KNN, RF, DT and LR The SVM classifier attained the uppermost performance with a supreme prediction accuracy of 97.07 percentage	2020
Early Detection of Breast Cancer Using Machine Learning Techniques	Researchers suggest that SVM is the most popular method used for cancer detection applications	2018
Breast Cancer Prediction system	It has been observed KNN classifier yields the highest classification accuracies when used with most predictive variables.	2018
Breast Cancer Detection using Machine Learning Techniques	In this paper we examined different machine learning techniques for breast cancer detection. We performed a comparative analysis of CNN, KNN, SVM, Logistic regression, Naïve Bayes, and Random Forest. It was observed that CNN outperforms the existing methods when it comes to accuracy, precision, and size of the data set.	2021

Breast Cancer Detection with Machine Learning	With this research paper we can see that among Naïve Bayes, Support Vector Machine, Adaboost, Random Forest Classifier, KNN, Decision Tree, XGboost etc. XGboost is the most accurate algorithm for best accurate result for	2022
Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis	In this study, four-layered essential data techniques were proposed with four different machines learning predictive models SVM, LR, KNN, and ensemble classifier. The significant finding demonstrated that the first prediction model (with an SVM polynomial kernel) had acquired the highest accuracy (99.3%).	2022
Breast Cancer Detection with Machine Learning	With this research paper we can see that among Naïve Bayes, Support Vector Machine, Adaboost, Random Forest Classifier, KNN, Decision Tree, XGboost etc. XGboost is the most accurate algorithm for best accurate result for detection of breast cancer with the efficiency of 98.24%	2022

Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis	In this study, four-layered essential data techniques were proposed with four different machines learning predictive models SVM, LR, KNN, and ensemble classifier. The significant finding demonstrated that the first prediction model (with an SVM polynomial kernel) had acquired the highest accuracy (99.3%).	2022
Breast Cancer Prediction using varying Parameters of Machine on using varying Models	The accuracy found by Adam Gradient Learning is highest because it combines benefits of AdaGrad and RMSProp	2020

CHAPTER 3

PROPOSED METHODOLOGY

3.1 METHODOLOGY

Our methodology begins with data acquisition followed by pre-processing, which contains four steps viz: data cleaning, select attributes, set target Role and features extraction. The prepared data is used to build machine learning algorithms that can predict the breast cancer for a new set of measurements. To evaluate the algorithms performances, we show the model new data for which we have labels. This is usually done by splitting the labeled data we have collected into two parts whit Train_test_split method. 75% of the data is used to build our machine learning model, and is called the training data or training set. 25% of the data will be used to access how well the model works and is called test data, test set. After testing the models we compare the obtained results to select the algorithm that provides the high accuracy and identify the most predictive algorithm for the detection of breast cancer.

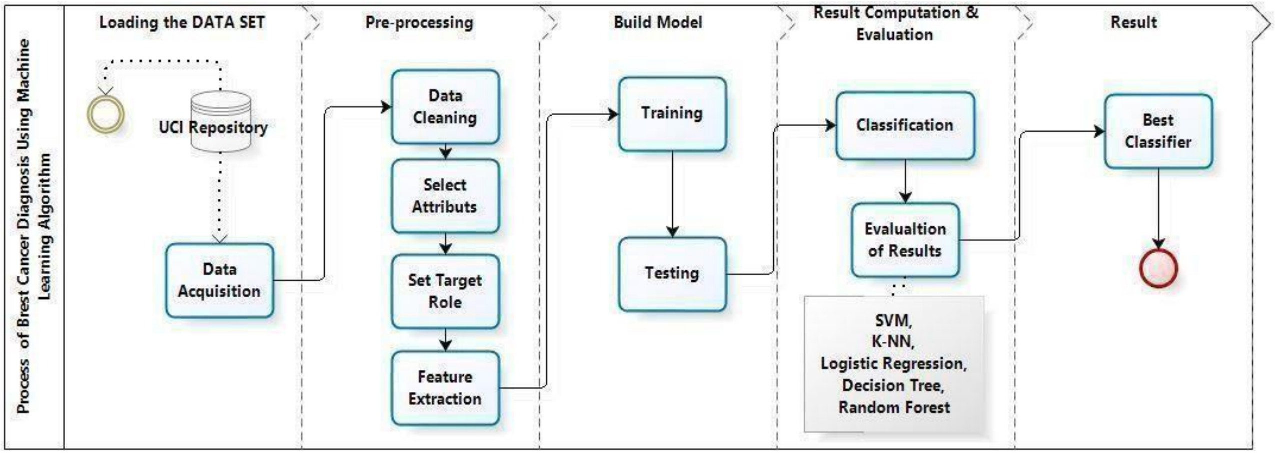


Fig 3.1 Methodology Process

In Fig. 3.2 Based on the findings, the algorithm that predicts the presence of breast cancer with the highest degree of accuracy is chosen. The algorithms that offer the highest level of accuracy and dependability in predicting the existence of breast cancer are identified after models have been examined.

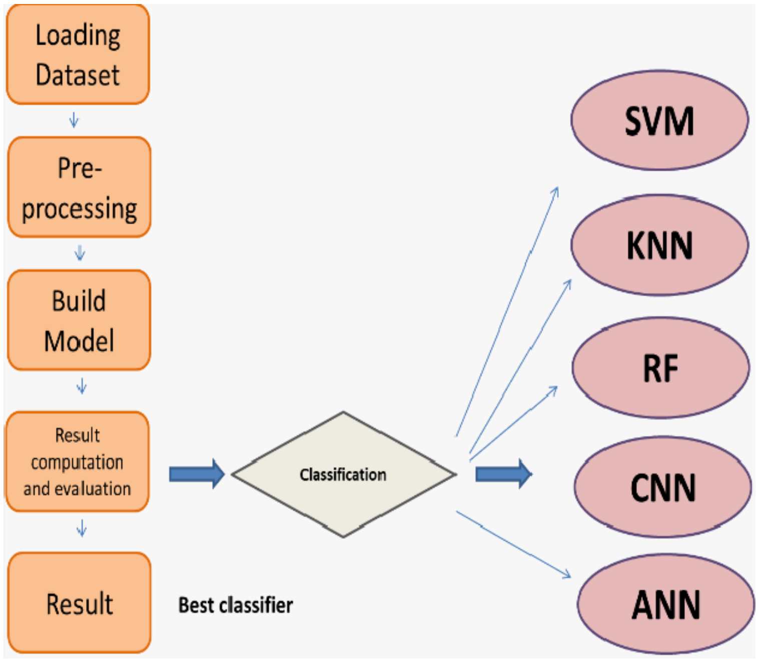


Fig. 3.2 Highest Accuracy ML Algorithms

3.2 Dataset

The most commonly used datasets for breast cancer prediction include the SEER Breast Cancer (SEERBCD) , the Coimbra Breast Cancer (CBC) dataset , the Wisconsin (Prognostic) Breast Cancer (WPBC) dataset , the Wisconsin (Diagnostic) Breast Cancer dataset [26], the Wisconsin Original Breast Cancer (WOBC) dataset , and the Breast Tissue Dataset (BTD) .

The data used to carry out the experiments in the current studies is taken from the Wisconsin Breast Cancer Dataset (WBCD), which is already labelled as malignant and benign. The dataset includes thirty features derived from fine-needle aspiration (FNA) of the breast mass. Typically, cancer datasets take the form of images. The repository consists of feature vectors that describe the image's cell nuclei The characteristics of cell nuclei are made up of ten real-valued features. These characteristics are as follows:

Table 3.1 Properties of dataset

	Features	
Radius mean	Radius SE	Radius worst
Texture mean	Texture SE	Texture worst
Perimeter mean	Perimeter SE	Perimeter worst
Area mean	Area SE	Area worst
Smoothness mean	Smoothness SE	Smoothne worst
Compactness mean	Compactness SE	Compactness worst
Concavity mean	Concavity SE	Concavity worst

3.3 Machine Learning Algorithms

In our project, the predictive analysis of the machine learning algorithms is achieved. The machine learning algorithms applied in our project are:

- **Support Vector Machine (SVM)** is a classifier which divides the datasets into classes to find a maximum marginal hyper plane (MMH) via the nearest data points.

- **Random forests or random decision forests** are an ensemble method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

- **k-Nearest Neighbors (K-NN)** is a supervised classification algorithm. It takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point, which is its nearest neighbors, and has those neighbors vote.

- **Logistic regression** is a very powerful modeling tool, is a generalization of linear regression. Logistic Regression is used to assess the likelihood of a disease or health condition as a function of a risk factor (and covariates). Both simple and multiple logistic regression , assess the association between independent variable(s) (Xi) -- sometimes called exposure or predictor variables — and a dichotomous dependent variable (Y) -- sometimes called the outcome or response variable. It is used primarily for predicting binary or multiclass dependent variables.

- **Decision Tree C4.5** is a predictive modeling tool that can be applied across many areas. It can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions.

CHAPTER 4

RESULTS AND DISCUSSION

Various machine learning such as K-Nearest Neighbor(KNN), Support Vector Machine(SVM), Decision tree, Naïve Bayes Logistic Regression, Random Forest were used for predicting breast cancer on the Wisconsin dataset. The maximum accuracy achieved by ANN, CNN, SVM, Random Forest, KNN algorithms. These Algorithms achieved more than 95% and in order to increase the prediction accuracy, deep learning algorithms such as Convolutional Neural Network (CNN) and Artificial Neural Network (ANN) were implemented.

Table 4.1 Result of ML Techniques

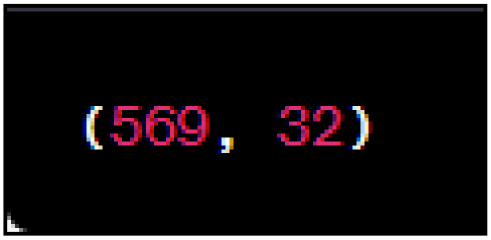
Algorithm	Accuracy	Precision	Sensitivity
ANN	0.99	0.99	0.99
CNN	0.97	0.97	0.98
SVM	0.96	0.98	0.97
Random Forest	0.96	0.98	0.97
KNN	0.95	0.95	0.99
Decision tree	0.95	0.99	0.93

Logistic Regression	0.94	0.96	0.96
Naïve Bayes	0.92	0.93	0.94

So, in our future works, we plan to apply our machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

Output of codes

1. Loading the CSV File: The code will use a library like pandas to read the CSV file into a data structure such as a DataFrame. This is typically done using a function like read_csv().When code loads the dataset from a csv file then output will be:

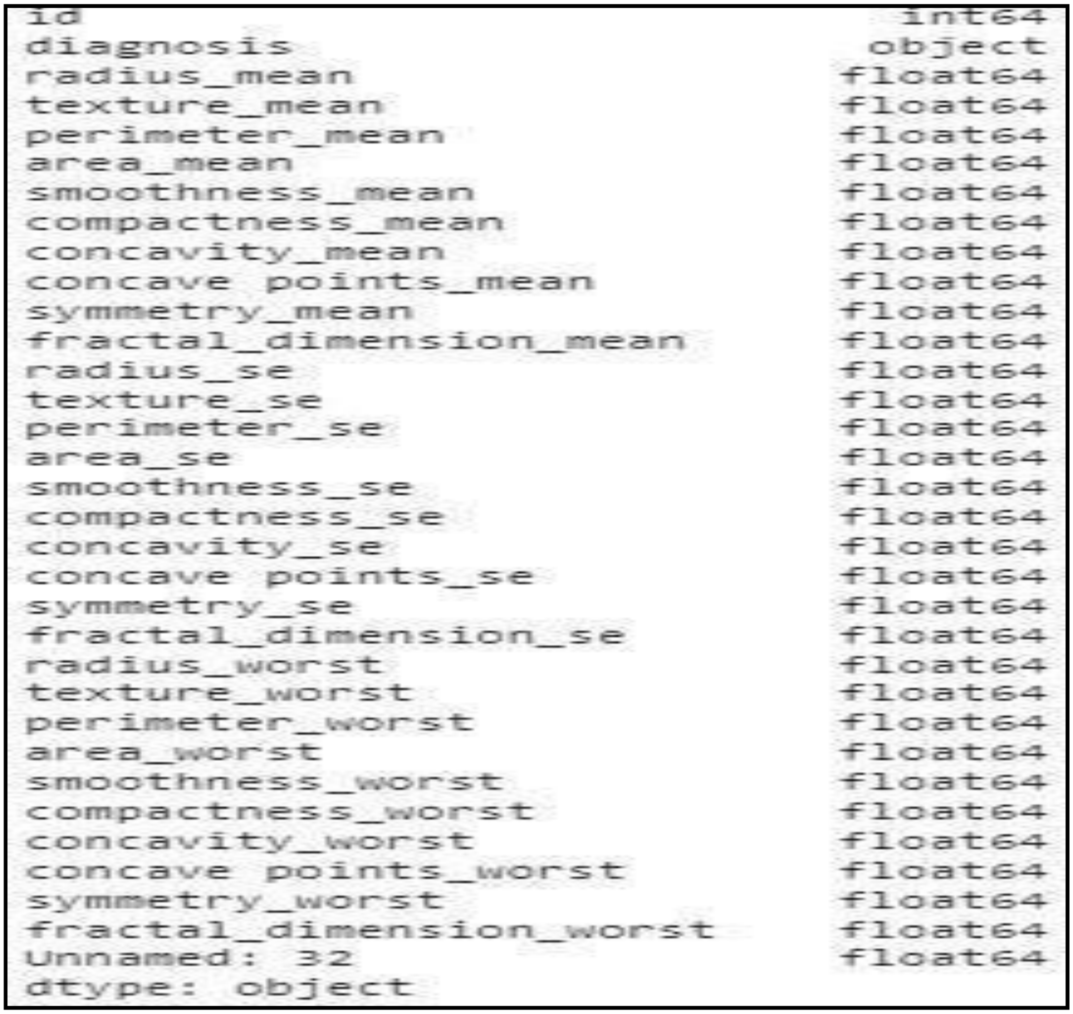


```
(569, 32)
```

Fig. 4.1 Dataset Output

Fig. 4.1 shows that there are 569 rows and 32 columns present in dataset .

2. When we check the datatype of all attributes we will get below output. Fig. 4.2 shows that the datatypes of all the 32 attributes which are present in the dataset

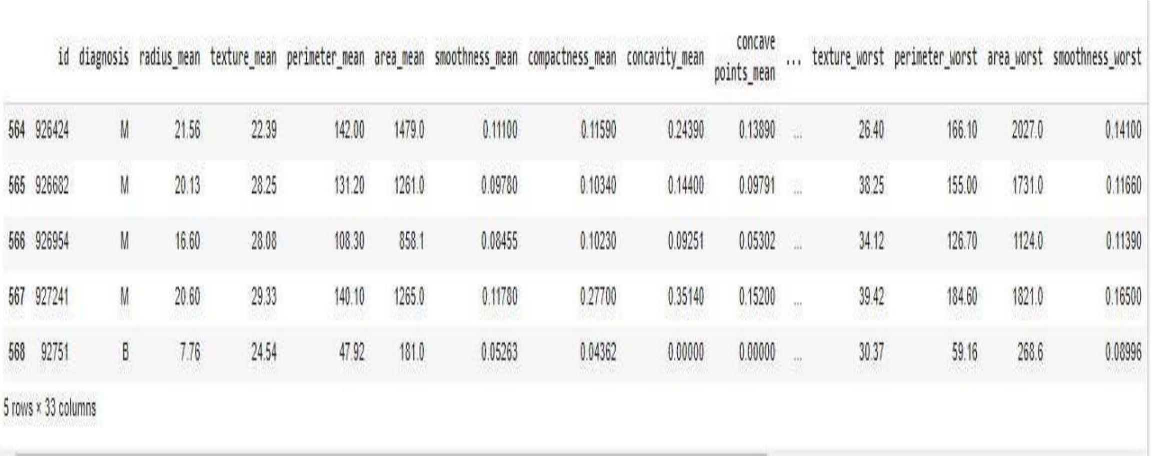


```
id                           int64
diagnosis                   object
radius_mean                float64
texture_mean               float64
perimeter_mean             float64
area_mean                  float64
smoothness_mean            float64
compactness_mean           float64
concavity_mean             float64
concave points_mean        float64
symmetry_mean              float64
fractal_dimension_mean     float64
radius_se                  float64
texture_se                 float64
perimeter_se               float64
area_se                    float64
smoothness_se              float64
compactness_se             float64
concavity_se               float64
concave points_se          float64
symmetry_se                float64
fractal_dimension_se       float64
radius_worst               float64
texture_worst              float64
perimeter_worst            float64
area_worst                 float64
smoothness_worst           float64
compactness_worst          float64
concavity_worst            float64
concave points_worst       float64
symmetry_worst             float64
fractal_dimension_worst    float64
Unnamed: 32                float64
dtype: object
```

Fig. 4.2 Features Output

.

3. When we will execute code to show the first and last rows of dataframe, the below output will be shown:



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_worst	perimeter_worst	area_worst	smoothness_worst
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	...	26.40	166.10	2027.0	0.14100
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	...	38.25	155.00	1731.0	0.11660
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	...	34.12	126.70	1124.0	0.11390
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	...	39.42	184.60	1821.0	0.16500
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	...	30.37	59.16	268.6	0.08996

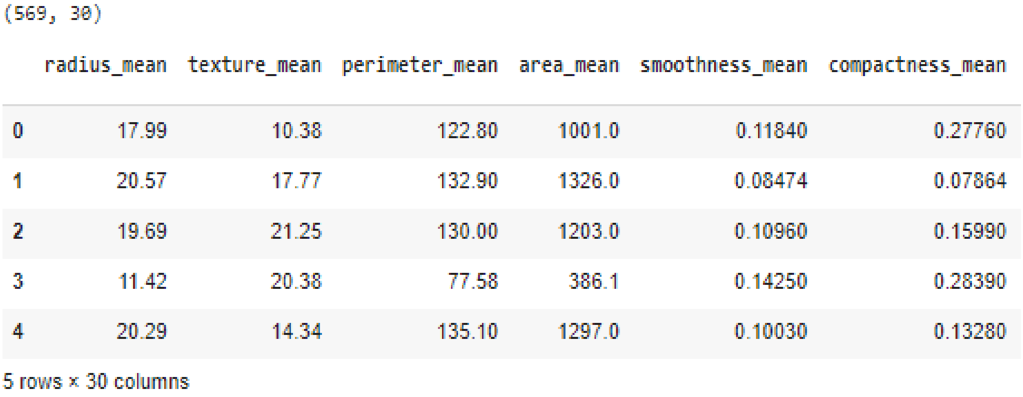
5 rows × 33 columns

Fig. 4.3 Show Data Output

Fig. 4.3 show the first and last rows of a DataFrame in Python, you can use the head() and tail() functions provided by the pandas library.

Here, column1, column2, ..., columnN represent the column names, and value1, value2, ..., valueN represent the corresponding values in each column. The [5 rows x N columns] indicates the size of the displayed DataFrame, showing that it contains 5 rows and N columns.

4. When we load the predicators into the dataframe below output will obtain.



(569, 30)

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	11.42	20.38	77.58	386.1	0.14250	0.28390
4	20.29	14.34	135.10	1297.0	0.10030	0.13280

5 rows × 30 columns

Fig. 4.4 Dataframe Output

Fig. 4.4 show

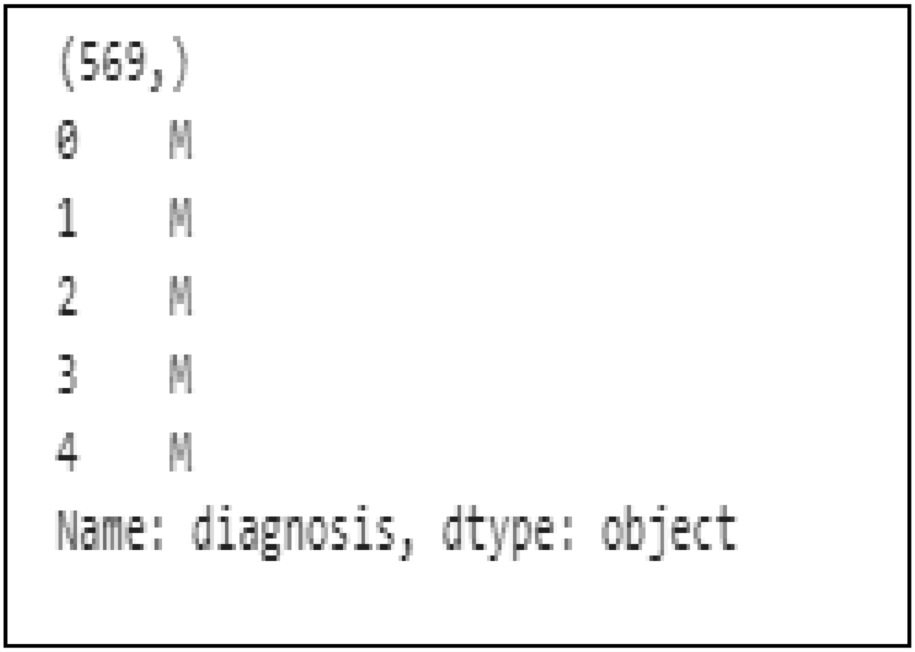
1. The loaded predictors will be organized into a tabular structure called a Data Frame. Each predictor will be represented as a column in the Data Frame, and each row will correspond to a separate sample or instance.

2. The Data Frame will display the predictors in a structured format, showing the values and relevant information. This can include the column names, data types, and the actual values of the predictors.

5. By loading target values in dataframe ‘Y’ we get below output:

Fig 4.5 shows

The 'y.head()' output displays the first few rows of the 'y' Series. Each row is represented by an index number followed by the corresponding value in the 'diagnosis' column.



```
(569,)
0    M
1    M
2    M
3    M
4    M
Name: diagnosis, dtype: object
```

Fig. 4.5 Y Dataframe Output

6. Fig 4.6. shows the correlation matrix of the predictors Dataframe 'X' and the code snippet you provided creates a heatmap using the Seaborn library to visualize the correlation matrix of the predictors DataFrame 'X'.

The output will be a heatmap plot that visually represents the correlation matrix of the 'X' DataFrame. The cells in the heatmap will display color intensities based on the correlation values, allowing you to identify patterns and relationships between the predictors. The color bar on the side of the plot will indicate the correlation values associated with the color mapping.

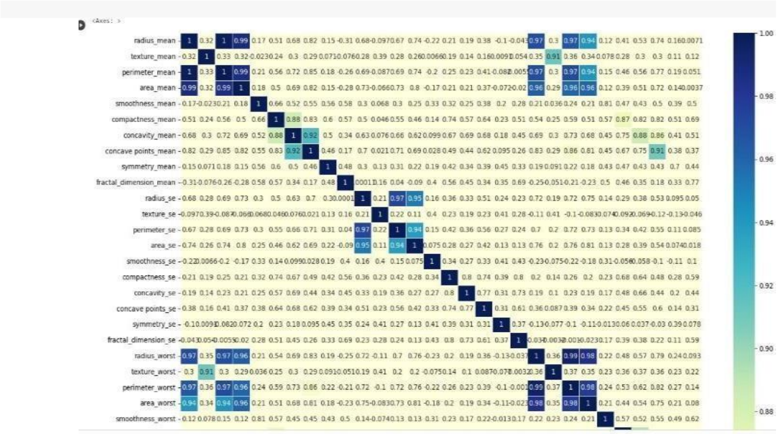


Fig. 4.6 X Dataframe Output

7. Fig 4.7 shows

The output will be a scatter plot with points representing the relationship between "PC1" and "PC2" variables. The points will be colored differently based on the values in the "diagnosis" column, allowing you to visually identify any patterns or separation between the two diagnosis categories. The markers for the points will be displayed as circles ("o") or crosses ("x") depending on the respective values in the "diagnosis" column.

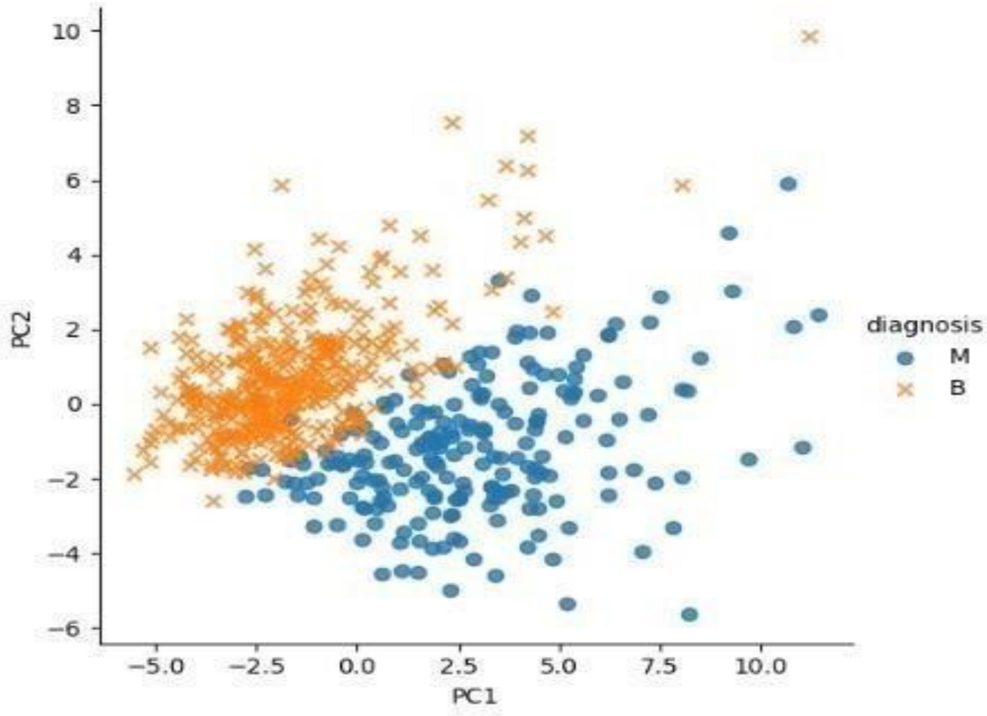
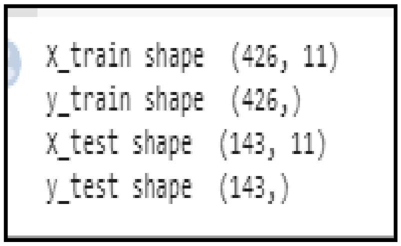


Fig. 4.7 Scatter Plot Output

8. Train-Test data splitting shows the below output:

Fig 4.8

The code snippet you provided splits the predictors DataFrame 'X' and the target values Series 'y' into training and testing sets using the train_test_split function from scikit-learn. It then prints the shape of the resulting train and test sets



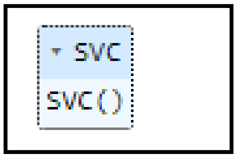
```
X_train shape  (426, 11)
y_train shape  (426,)
X_test shape  (143, 11)
y_test shape  (143,)
```

Fig 4.8 Train Test Output

9. Fig 4.9 shows the model fitting output :

1. svc = SVC(): This line creates an instance of the SVC classifier. By default, it uses the radial basis function (RBF) kernel.

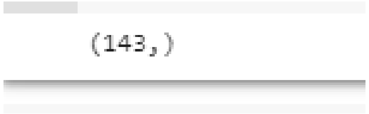
2. svc.fit(X_train, y_train): This line fits the SVC model to the training data. The fit() method takes the training predictors X_train and the corresponding target values y_train as input and trains the model on this data.



```
▾ SVC
SVC()
```

Fig. 4.9 Model Fiting Output

10. Fig.4.10 shows predicted value code output and The specific value of n_test will depend on the number of instances in the testing set (X_test) and the shape of the predicted target values. It represents the number of instances for which the SVC model has provided predictions.



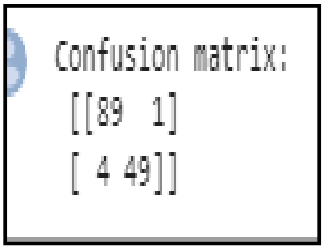
```
(143,)
```

Fig. 4.10 Predicated Value Output

11. Confusion matrix output:

Fig 4.11 shows

The confusion matrix is a 2x2 matrix that represents the counts of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). The elements of the confusion matrix quantify the performance of the classification model, indicating the number of correct and incorrect predictions.



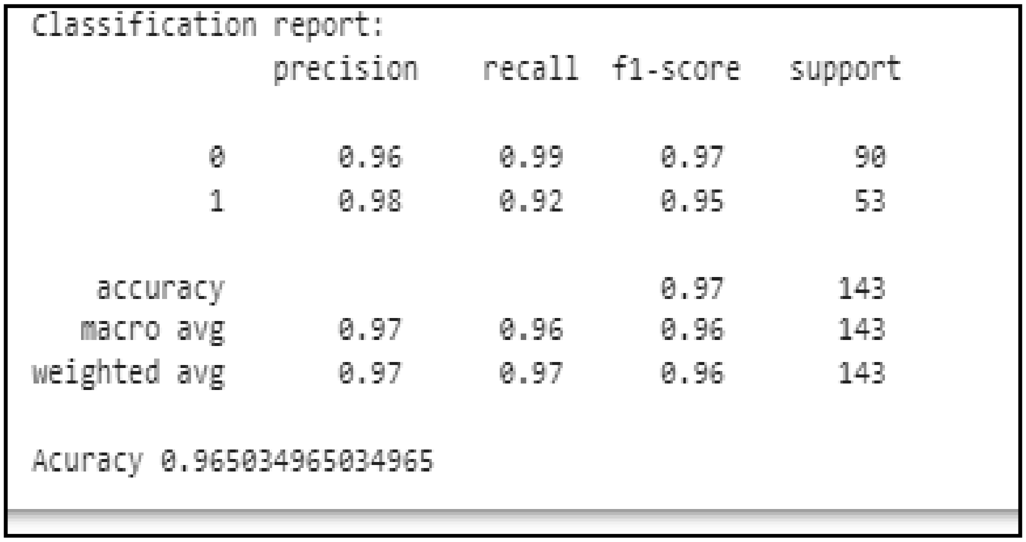
```
Confusion matrix:
 [[89  1]
 [ 4 49]]
```

Fig. 4.11 Confusion matrix Output

12. Final output is Classification Report:

Fig 4.12

The output will display the classification report and the accuracy score, providing a comprehensive evaluation of the SVC model's performance.



```
Classification report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        90
           1       0.98      0.92      0.95        53

    accuracy                           0.97       143
   macro avg       0.97      0.96      0.96       143
weighted avg       0.97      0.97      0.96       143

Acuracy 0.965034965034965
```

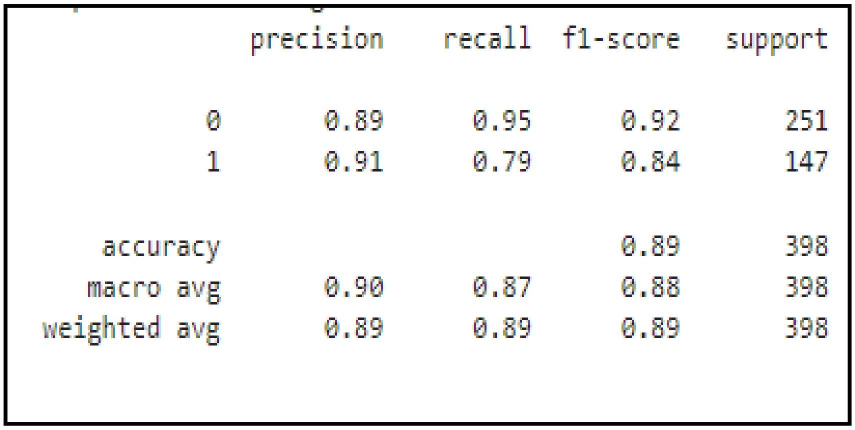
Fig. 4.12 Classification Report

Code Output of Logistic Regression

1. Training data output of logistic Regression algorithm:

Fig 4.13 shows

We use the score() method, passing the predictor variables (test_data[predictor_var]) and the true outcome values (test_data[outcome_var]) to get the accuracy score (accuracy).



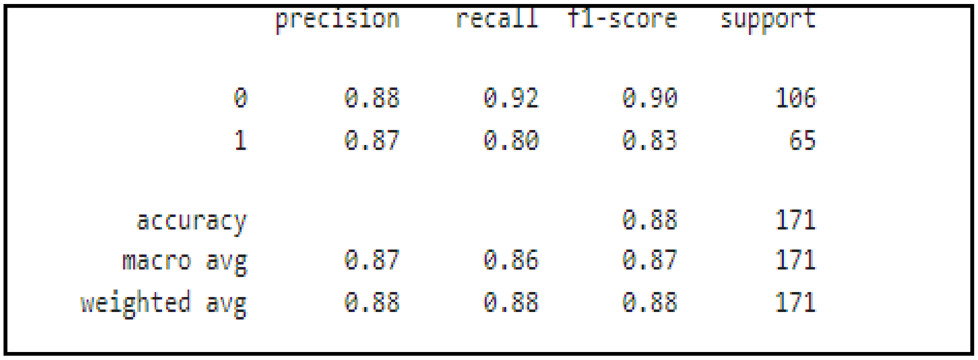
```
              precision    recall  f1-score   support

           0       0.89      0.95      0.92       251
           1       0.91      0.79      0.84       147

    accuracy                           0.89       398
   macro avg       0.90      0.87      0.88       398
weighted avg       0.89      0.89      0.89       398
```

Fig. 4.13 Training Data Output of Regression algorithm

2.Report for testing data for logistic regression



```
              precision    recall  f1-score   support

           0       0.88      0.92      0.90       106
           1       0.87      0.80      0.83        65

    accuracy                           0.88       171
   macro avg       0.87      0.86      0.87       171
weighted avg       0.88      0.88      0.88       171
```

Fig. 4.14 Testing data for logistic regression

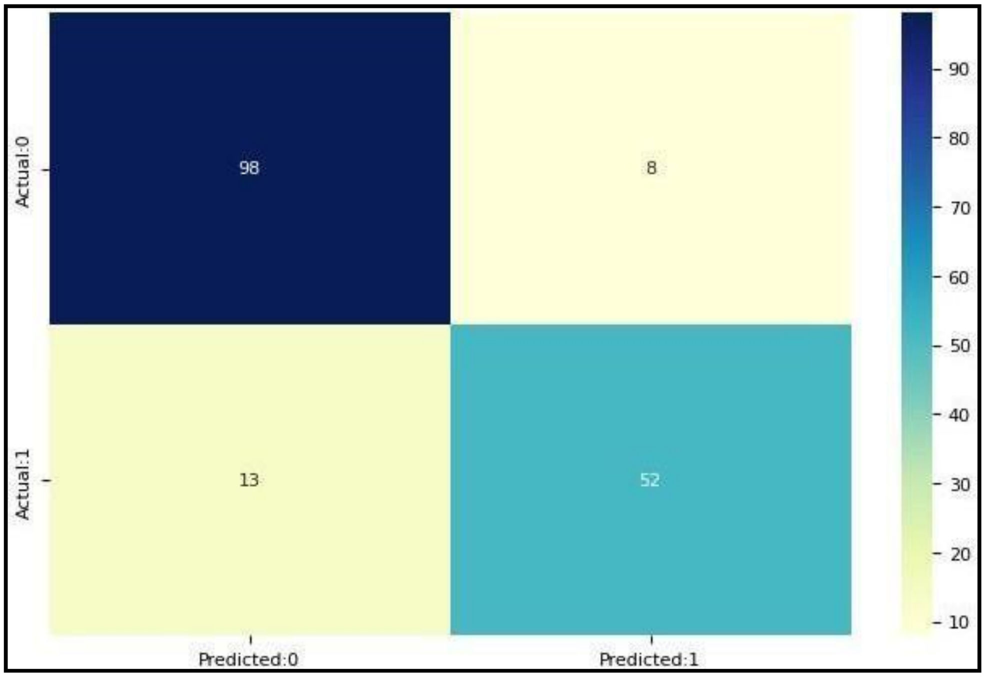


Fig. 4.15 Graph Output on Logistic Regression

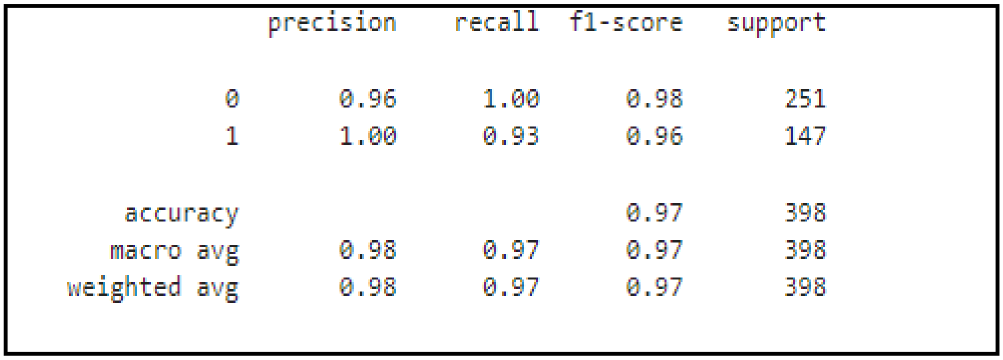
Code Output of Decision Tree

1. Training data report for Decision Tree Algorithm:

Fig 4.16 shows

In the code above, we import the DecisionTreeClassifier class from scikit-learn and define the predictor variables (predictor_var) and the outcome variable (outcome_var). Then, we create an instance of the decision tree classifier (model) using DecisionTreeClassifier().

Next, we train the model by calling the fit() method on the training data (train_data). We pass the predictor variables (train_data[predictor_var]) and the outcome variable (train_data[outcome_var]) to the fit() method.



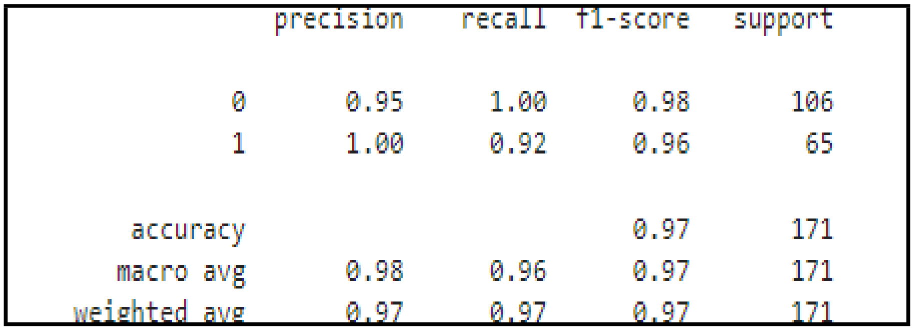
```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       251
           1       1.00      0.93      0.96       147

    accuracy                           0.97       398
   macro avg       0.98      0.97      0.97       398
weighted avg       0.98      0.97      0.97       398
```

Fig. 4.16 Training data report for Decision Tree Algorithm

Testing data report for decision tree:



```
              precision    recall  f1-score   support

           0       0.95      1.00      0.98       106
           1       1.00      0.92      0.96        65

    accuracy                           0.97       171
   macro avg       0.98      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171
```

Fig. 4.17 Testing data report for decision tree

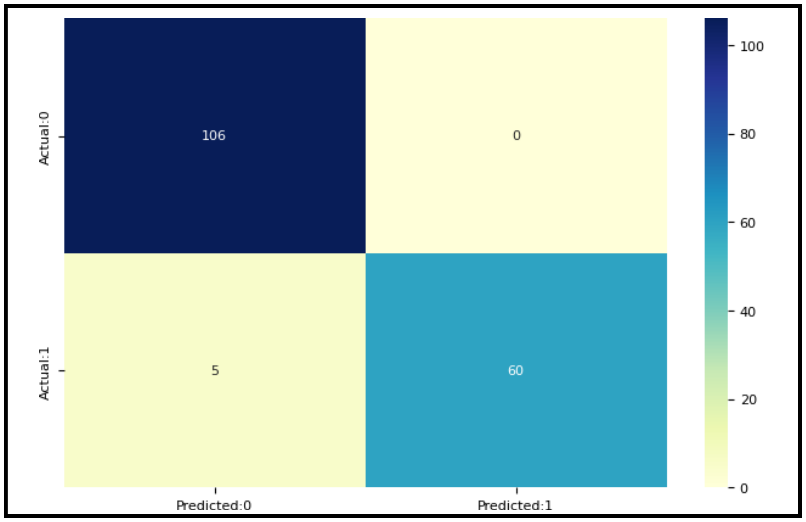


Fig. 4.18 Graph Output on Decision tree

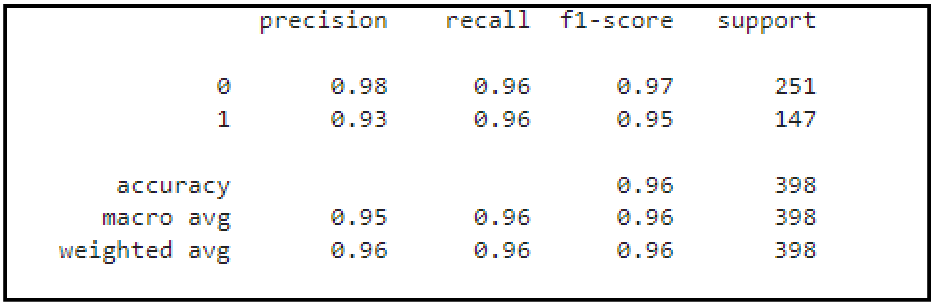
2. Report of training data for Random Forest algorithm:

Fig. 4.19 shows

1.from sklearn.ensemble import RandomForestClassifier: This line imports the RandomForestClassifier class from the sklearn.ensemble module, which is used for training a Random Forest classifier.

2.predictor_var = features_mean: The predictor_var is set to features_mean, which presumably contains the predictor variables for the Random Forest classifier. However, the specific definition and content of features_mean are not provided in the code snippet.

model = RandomForestClassifier(n_estimators=100, min_samples_split=25, max_depth=7, max_features=2): This line creates an instance of the Random Forest classifier. The n_estimators parameter sets the number of trees in the forest (100 in this case).



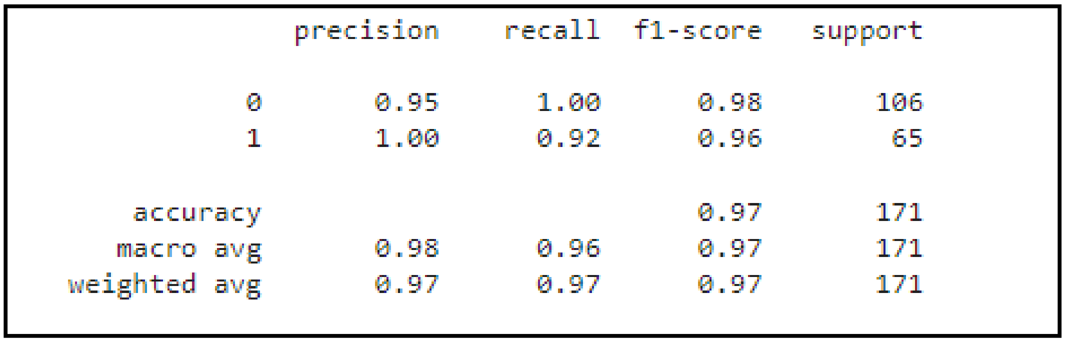
```
              precision    recall  f1-score   support

           0       0.98      0.96      0.97       251
           1       0.93      0.96      0.95       147

    accuracy                           0.96       398
   macro avg       0.95      0.96      0.96       398
weighted avg       0.96      0.96      0.96       398
```

Fig. 4.19 Report of training data for Random Forest algorithm

3. Report of testing data for Random Forest:



```
              precision    recall  f1-score   support

           0       0.95      1.00      0.98       106
           1       1.00      0.92      0.96        65

    accuracy                           0.97       171
   macro avg       0.98      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171
```

Fig. 4.20 Report of testing data for Random Forest

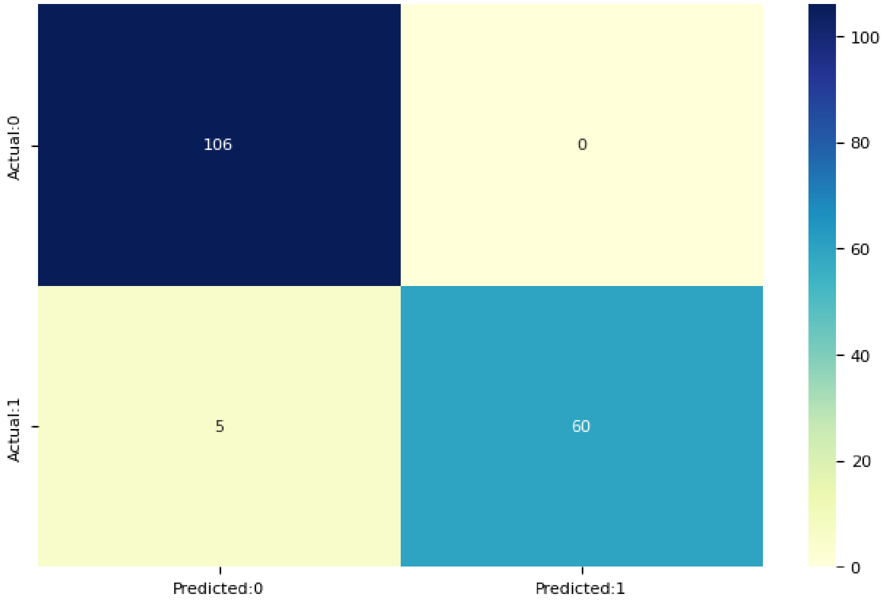


Fig. 4.21 Graph Output on Random Forest

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Prognosis of breast cancer using machine learning algorithms has shown promising results in accurately predicting the progression of the disease and the likelihood of survival. Through the use of various ML algorithms such as Random Forest, Logistic Regression, and Support Vector Machines, etc. researchers have been able to develop models that can accurately predict the survival rate and the risk of recurrence in breast cancer patients.

Based on the results of the study, the best machine learning algorithm for predicting breast cancer prognosis may depend on several factors, such as the size of the dataset, the number of variables used, and the specific type of breast cancer being studied.

It's essential to note that while ML-based prognosis models can improve the accuracy of diagnosis and treatment recommendations, they should be used in conjunction with clinical expertise and patient involvement. ML algorithms cannot replace the importance of personalized care, communication between patients and healthcare providers, and shared decision-making.

In conclusion, we looked at several studies that discovered that deep learning algorithms, machine learning approaches, and some proposed algorithms provide improved accuracy for detecting breast cancer on various datasets. However, machine learning algorithms like SVM, KNN, RF, ANN, and CNN provide the highest level of accuracy. However, further research is needed to validate the effectiveness of these models and to identify the most appropriate algorithm for specific breast cancer types and patient populations.

5.2 FUTURE SCOPE

The future scope of research in prognosis of breast cancer using machine learning algorithms is vast and promising. Some potential areas for future research include:

- **Development of personalized treatment plans**: As the accuracy of prognosis models improves, it may be possible to develop personalized treatment plans based on a patient's specific risk factors and medical history. This could lead to more effective treatments and improved patient outcomes.
- **Integration of multimodal data**: Integrating data from various sources, including medical images, genetic profiles, and clinical records, may provide more comprehensive information for developing accurate prognosis models.
- **Transfer learning**: Transfer learning, which involves leveraging knowledge learned from one dataset to improve the accuracy of predictions on a different dataset, could help overcome the challenge of limited data availability in some settings.
- **Development of real-time prediction tools**: Real-time prediction tools that can continuously monitor patients' health status and adjust treatment plans accordingly may improve patient outcomes and reduce the risk of recurrence.
- **Validation of models in diverse populations**: To ensure the effectiveness of prognosis models, it's important to validate them in diverse patient populations, including those with different demographic backgrounds and medical histories.

In Future, models which based on these techniques will developed and try to develop a software which is accessible by user itself for Breast Cancer Prediction.

REFERENCES

[1] ‘WHO | Breast cancer’,WHO,http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[2] “UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.”

[3] S. Aamir, A. Rahim, Z. Aamir, S. F. Abbasi, M. S. Khan, M. Alhaisoni, M. A. Khan, K. Khan, and J. Ahmad “Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques,” August 16,**2022**.

[4] A.B. Nassif, M.A. Talib, Q. Nasir, Y.Afadar,and O.Elgendy“Breast cancer detection using artificial intelligence techniques: A systematic literature review,” March 5,**2022**. **[5]** H.E. Agouri, M. Azizi, H. E. Attar, M. E. Khannoussi, A. Ibrahimi, R. Kabbaj, H. Kadiri, S. BekarSabein, S. EchCharif, C. Mounjid, B. E. Khannoussi, “Assessment of deep learning algorithms to predict histopathological diagnosis of breast cancer: first Moroccan prospective study on a private dataset,” Feb 19,2022.

[6] M.Mangukiya, A.Vaghani, M.Savani, “Breast Cancer Detection with Machine Learning,” Feb 2022.

[7] A. Rasool, C.Bunterngchit, L.Tiejian, Md. R. Islam, Q. Qu, and Q. Jiang, “Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis” March 9,2022.

[8] H. Zhang, H. Liu, L. Ma, J. Liu, D. Hu “Ultrasound Image Features under Deep Learning in Breast Conservation Surgery for Breast Cancer” Sept 17,2021

[9] S. Bhise, S. Bepari, S. Gadekar, D. Kale, A. S. Gaur, Dr. S. Aswale, “Breast Cancer Detection using Machine Learning Techniques,” July 7,2021.

[10] S. Aryal, B. Paudel, “Supervised Classification using Gradient Boosting Machine: Wisconsin Breast Cancer Dataset,” June 2020.

[11] M. Srivenkatesh, “Prediction of Breast Cancer 0Disease using Machine Learning Algorithms,” Feb 4,2020.

[12] N. Fatima, L. Liu, S. Hong, H. Ahmed, “Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis,” Aug 14,2020.

[13] R. Rawal, “BREAST CANCER PREDICTION USING MACHINE LEARNING,” May 2020 Gaurav Singh, “Breast Cancer Prediction using Machine Learning,” July 30,2020.

[14] M.Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M.Sazzadur Rahman, Imran Mahmud, R. Akter, “An Analysis on Breast Disease Prediction Using Machine Learning Approaches,” Feb 2020.

[15] P. Gupta, S. Garg, “Breast Cancer Prediction using varying Parameters of Machine Learning Models,” June 4,2022.

[16] M. D. Ganggayah, N.A Talib, Y. C. Har, P. Lio, S. K. Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” March 22, 2019.

[17] Ch. Shravya, K. Pravalika, Shaik Subhani, “Prediction of Breast Cancer Using Supervised Machine Learning Techniques,” April 2019.

[18] Yixuan Li, Zixuan Chen, “Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction,” Oct 18,2018. **[19]** Madhu Kumari, Vijendra Singh, “Breast Cancer Prediction System,” June 8,2018.

[20] H. Asri, H. Mossanif, H. Al Moatassime, T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” May 12,2016.

[21] K. Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios
a. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Nov15,2014.

APPENDIX 1

Code of SVM Algorithm

Certainly! The code snippet, "import pandas as pd," is a common way to import the Python library Pandas and alias it as "pd" for convenience. Pandas is a powerful library used for data manipulation and analysis. To perform data analysis operations on breast cancer data using Pandas, you would typically follow these steps:

Import the required libraries:

```
#for loading data and for performing data analysis operations on it import pandas as pd
import numpy as np

#for data visualization import seaborn as sns
import matplotlib.pyplot as plt

#for PCA (feature engineering)

from sklearn.decomposition import PCA


#for data scaling

from sklearn.preprocessing import StandardScaler


#for splitting dataset

from sklearn.model_selection import train_test_split


#for fitting SVM model

from sklearn.svm import SVC


#for displaying evaluation metrics from sklearn.metrics import classification_report from
sklearn.metrics import confusion_matrix

#for file operations
import os
```

```
print("All required libraries loaded!")
```

The code loads a dataset from a CSV file.

The read_csv() function in Pandas is used to read a CSV file and create a DataFrame from it. The file path "/content/sample_data/breast-cancer.csv" is the location of the CSV file you want to load. You may need to modify this path based on the actual location of your file.The **shape** attribute of a DataFrame returns a tuple representing the dimensions of the DataFrame, i.e., the number of rows and columns. In this case, **df.shape** will give you the number of rows and columns in the DataFrame **df**.

The output will be a tuple with two values: the number of rows and the number of columns in the DataFrame.

For example, if the dataset has 569 rows and 32 columns, the output will be:

The figure **Fig. 4.1** in result and discussion shows the output of this code.

The **dtypes** attribute in Pandas is used to check the data types of all the attributes (columns) loaded into a DataFrame. It returns a Series object that contains the data types of each column.

To check the data types of all attributes in the DataFrame **df**, you can simply use the **dtypes** attribute as follows:

```
#check the data types of all the attributes loaded into the dataframe

df.dtypes
```

The output will be a Series object where the index represents the column names, and the values represent the data types of each column.

The output is shown in fig. 4.2.

To see the first few rows of the loaded data in the DataFrame **df**, you can use the **head()** function. Similarly, to see the last few rows, you can use the **tail()** function. Here's how you can use these functions:

```
#see first few rows of the data loaded

df.head()

#see last few rows of the data loaded

df.tail()
```

The **Fig. 4.3** shows the output of this code.

Loading the predictors into a DataFrame named 'X' by selecting specific columns from the original DataFrame 'df'. The columns 'id', 'diagnosis', and 'Unnamed:32' are excluded from 'X'.

```
#loading the predictors into dataframe 'X'

#NOTE: we are not choosing columns - 'id', 'diagnosis', 'Unnamed:32'X = df.iloc[:,2:32]

print(X.shape)

X.head()
```

The code uses the **iloc** indexer in Pandas to select specific columns from the DataFrame 'df'. The **iloc** indexer allows for integer-based indexing and slicing.

:, 2:32 specifies that you want to select all rows (denoted by **:**) and columns from index 2 to 31 (32 is excluded). This effectively excludes columns with index 0 ('id'), index 1 ('diagnosis'), and index 32 ('Unnamed:32').

Output is shown in **Fig. 4.4** in Result and discussion section.

By running the provided code, you would obtain the shape of 'X' and see the first few rows of the predictors loaded into the DataFrame 'X'.

```
#loading target values into dataframe 'y'

y = df.diagnosis

print(y.shape)

y.head()
```

The output of the above code is shown in **Fig 4.5.**

```
plt.figure(figsize=(18, 12))

sns.heatmap(X.corr(), vmin=0.85, vmax=1, annot=True, cmap='YlGnBu', linewidths=.5)
```

It creates a heatmap using the Seaborn library to visualize the correlation matrix of the predictors DataFrame 'X'.

The output of above code is shown in **Fig. 4.6** in Result section.

The code uses Seaborn's lmplot() function to create a scatter plot to visualize the data from the DataFrame Xy.

```
#visualize data

sns.lmplot(x = "PC1",y = "PC2", hue="diagnosis", data=Xy, fit_reg=False,
markers=["o", "x"])

plt.show()
```

By calling plt.show(), the scatter plot will be displayed in a separate window or inline within a Jupyter Notebook, allowing you to visualize the data and analyze any patterns or relationships between the variables.

The output of above code is shown in **Fig. 4.7.**

```
X=(Xy.iloc[:,0:11]).values
```

```
#75:25 train:test data splitting

X_train,  X_test,  y_train,  y_test  =  train_test_split(X,  y,  test_size=0.25,
random_state=0)

print("X_train shape ",X_train.shape)

print("y_train shape ",y_train.shape)

print("X_test shape ",X_test.shape)

print("y_test shape ",y_test.shape
```

The output of above code is shown in **Fig. 4.8.**

```
#model fitting

svc = SVC()

svc.fit(X_train, y_train)
```

Fig. 4.9 shows the model fitting output.

```
#predict values

y_pred_svc =svc.predict(X_test)

y_pred_svc.shape
```

output is shown in **Fig. 4.10.**

```
#print confusion matrix

cm = confusion_matrix(y_test, y_pred_svc)

print("Confusion matrix:\n",cm)
```

Output is shown in Fig. **4.11**

To print the classification report and accuracy score, you would need to import the classification_report and accuracy_score functions from scikit-learn. Here's how you can use these functions:

In this code, it is assumed that you have already made predictions y_pred_svc using the SVC model on the test data X_test and stored the ground truth labels in y_test.

```
#print classification report

from sklearn.metrics import accuracy_score

creport = classification_report(y_test, y_pred_svc)

print("Classification report:\n",creport)

print("Acuracy",accuracy_score(y_test,y_pred_svc
```

The classification report is shown in **Fig.4.12.**

Code of Logistic Regression Algorithm

# Logistic Regression

```
predictor_var=['radius_mean','perimeter_mean','area_mean','compactness_mean','con
cave points_mean'] outcome_var = 'diagnosis' model=LogisticRegression()
classification_model(model,train_data,predictor_var,outcome_var, test_data)
```

predictor_var is a list that contains the names of the predictor variables/features that you want to use for classification. In this case, it includes 'radius_mean', 'perimeter_mean', 'area_mean', 'compactness_mean', and 'concave points_mean'. These are the features that will be used to predict the outcome variable 'diagnosis'.

outcome_var is a string that represents the name of the outcome variable/target variable you want to predict. In this case, it is 'diagnosis'.

Accuracy for training data: 89.196%

Report for Training Data:

Fig. 4.13 shows the training data output.

Accuracy for test data: 87.719%

Report for Test Data:

Fig. 4.14 shows the output of testing data.

Code of Decision Tree Algorithm

```
# Decision Tree

predictor_var = ['radius_mean','perimeter_mean','area_mean','compac
tness_mean','concave points_mean']
model = DecisionTreeClassifier()
classification_model(model,train_data,predictor_var,outcome_var)        # Overfitted redu
```

Accuracy : 100.000%

```
predictor_var = ['radius_mean'] model =
DecisionTreeClassifier()
classification_model(model,train_data,predictor_var,outcome_va
r, test_data)
```

Accuracy for training data: 97.487%

Training data report shown in **Fig. 4.16.**

Accuracy for test data: 97.076%

Report for Test Data is shown in **Fig.4.17.**

Code of Random Forest Algorithm

```
# Random Forest classifier predictor_var = features_mean

model = RandomForestClassifier(n_estimators=100,min_samples_split=25, max_depth=7,
max_features=2)
classification_model(model, train_data,predictor_var,outcome_var, test_data)
```

Assuming the classification_model() function is defined elsewhere and follows a similar structure as previously discussed, it will train the Random Forest classifier on the training data (train_data) using the specified predictor variables (predictor_var) and the outcome variable (outcome_var). Then, it will evaluate the model's performance on the test data (test_data).

Accuracy for training data: 95.980%

Report for Training Data:

Report is shown in **Fig. 4.19.**

Accuracy for test data: 97.076%

Report for testing data is shown in **Fig. 4.20.**

APPENDIX 2

OUTCOME

The outcome of the project, a review paper on Prediction of Cancer in Breast Using Machine Learning Techniques: A Comparative Study of Classification Algorithms, has provided a comprehensive analysis of the advancements, challenges, and future prospects in this field.

In this annexure, I have included the research paper titled "Prediction of Cancer in Breast Using Machine Learning Techniques: A Comparative Study of Classification Algorithms" and the proof of submission for The 14th International Conference on Computing, Communication and Networking Technologies (ICCCNT) organized by IIT Delhi.

Conference Details

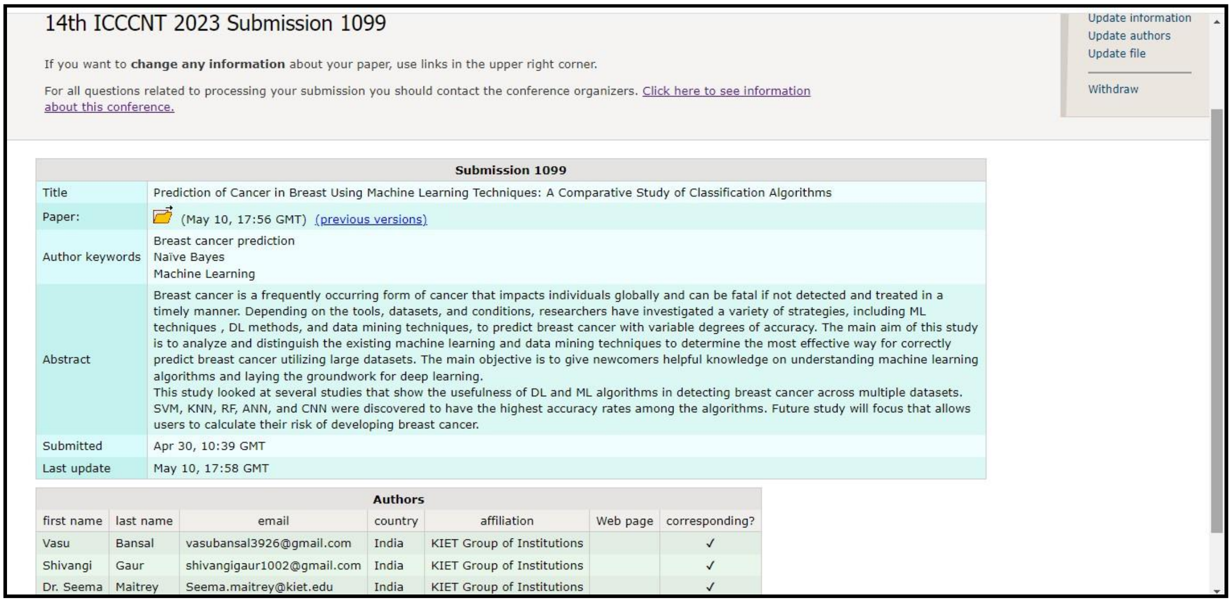
Paper Title: Prediction of Cancer in Breast Using Machine Learning Techniques: A Comparative Study of Classification Algorithms

Name of Scopus Conference: The 14th International Conference on Computing,

Communication and Networking Technologies (ICCCNT) By IIT Delhi (IEEE Xplore)

Date of Submission: 10th May, 2023

Proof of Submission:



14th ICCCNT 2023 Submission 1099

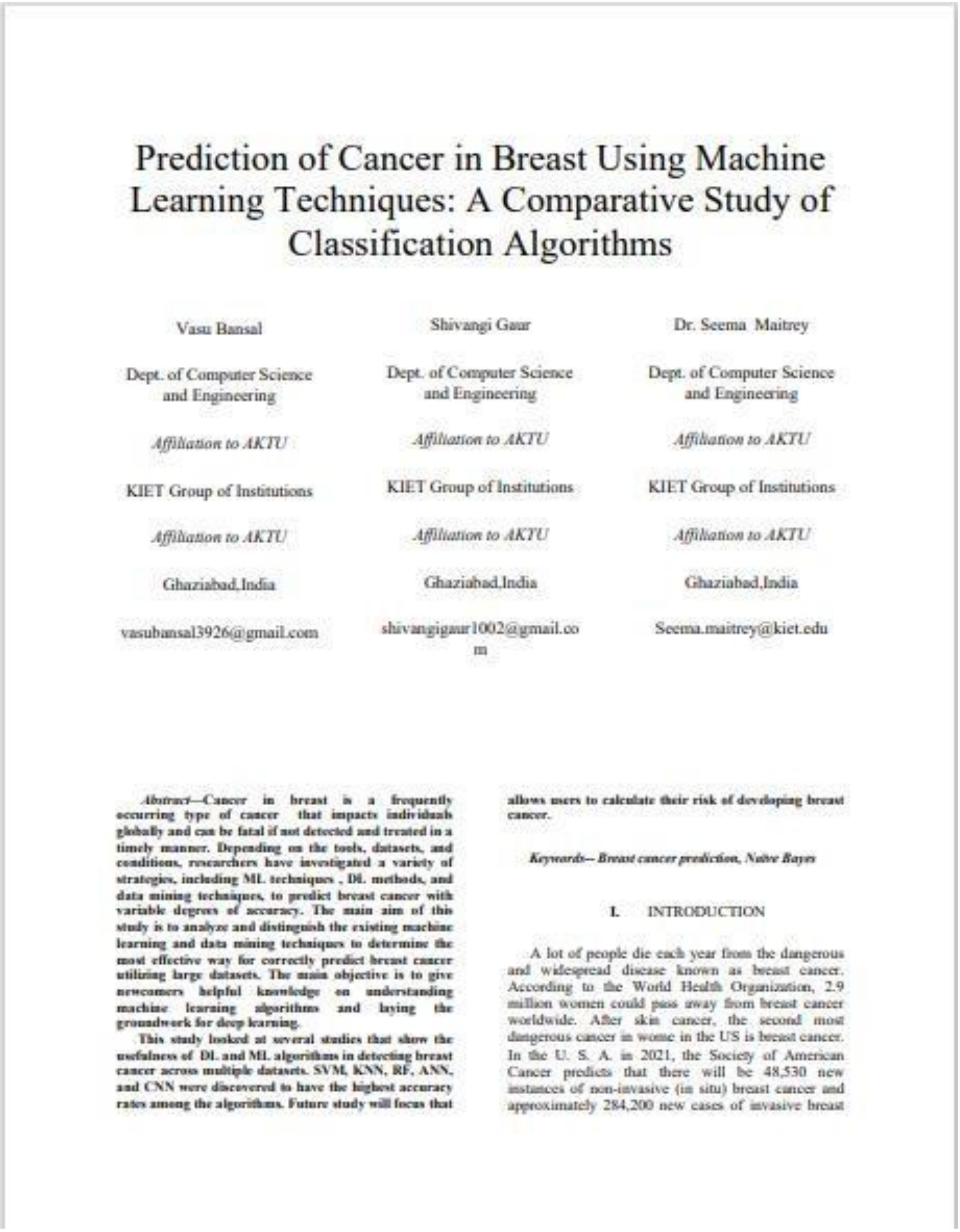
If you want to **change any information** about your paper, use links in the upper right corner.

For all questions related to processing your submission you should contact the conference organizers. Click here to see information about this conference.

Submission 1099	
Title	Prediction of Cancer in Breast Using Machine Learning Techniques: A Comparative Study of Classification Algorithms
Paper:	(May 10, 17:56 GMT) (previous versions)
Author keywords	Breast cancer prediction Naïve Bayes Machine Learning
Abstract	Breast cancer is a frequently occurring form of cancer that impacts individuals globally and can be fatal if not detected and treated in a timely manner. Depending on the tools, datasets, and conditions, researchers have investigated a variety of strategies, including ML techniques , DL methods, and data mining techniques, to predict breast cancer with variable degrees of accuracy. The main aim of this study is to analyze and distinguish the existing machine learning and data mining techniques to determine the most effective way for correctly predict breast cancer utilizing large datasets. The main objective is to give newcomers helpful knowledge on understanding machine learning algorithms and laying the groundwork for deep learning. This study looked at several studies that show the usefulness of DL and ML algorithms in detecting breast cancer across multiple datasets. SVM, KNN, RF, ANN, and CNN were discovered to have the highest accuracy rates among the algorithms. Future study will focus that allows users to calculate their risk of developing breast cancer.
Submitted	Apr 30, 10:39 GMT
Last update	May 10, 17:58 GMT

Authors						
first name	last name	email	country	affiliation	Web page	corresponding?
Vasu	Bansal	vasubansal3926@gmail.com	India	KIET Group of Institutions		✓
Shivangi	Gaur	shivangigaur1002@gmail.com	India	KIET Group of Institutions		✓
Dr. Seema	Maitrey	Seema.maitrey@kiet.edu	India	KIET Group of Institutions		✓

FINAL SUBMITTED PAPER



Prediction of Cancer in Breast Using Machine Learning Techniques: A Comparative Study of Classification Algorithms

Vasu Bansal
Dept. of Computer Science and Engineering
Affiliation to AKTU
KIET Group of Institutions
Affiliation to AKTU
Ghaziabad,India
vasubansal3926@gmail.com

Shivangi Gaur
Dept. of Computer Science and Engineering
Affiliation to AKTU
KIET Group of Institutions
Affiliation to AKTU
Ghaziabad,India
shivangigaur1002@gmail.com

Dr. Seema Maitrey
Dept. of Computer Science and Engineering
Affiliation to AKTU
KIET Group of Institutions
Affiliation to AKTU
Ghaziabad,India
Seema.maitrey@kiet.edu

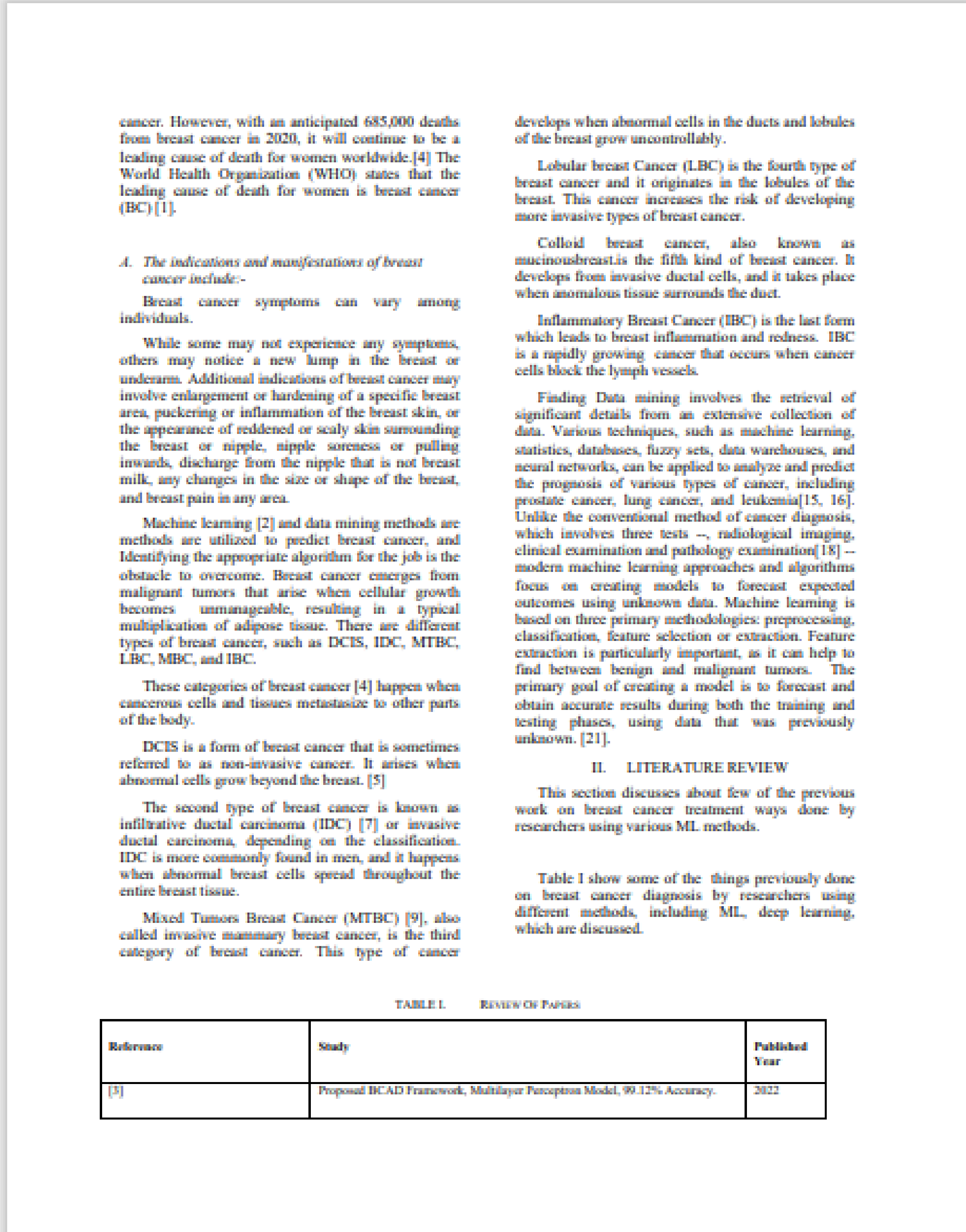
***Abstract*—Cancer in breast is a frequently occurring type of cancer that impacts individuals globally and can be fatal if not detected and treated in a timely manner. Depending on the tools, datasets, and conditions, researchers have investigated a variety of strategies, including ML techniques , DL methods, and data mining techniques, to predict breast cancer with variable degrees of accuracy. The main aim of this study is to analyze and distinguish the existing machine learning and data mining techniques to determine the most effective way for correctly predict breast cancer utilizing large datasets. The main objective is to give newcomers helpful knowledge on understanding machine learning algorithms and laying the groundwork for deep learning.**

This study looked at several studies that show the usefulness of DL and ML algorithms in detecting breast cancer across multiple datasets. SVM, KNN, RF, ANN, and CNN were discovered to have the highest accuracy rates among the algorithms. Future study will focus that allows users to calculate their risk of developing breast cancer.

Keywords*— *Breast cancer prediction, Naïve Bayes

I. INTRODUCTION

A lot of people die each year from the dangerous and widespread disease known as breast cancer. According to the World Health Organization, 2.9 million women could pass away from breast cancer worldwide. After skin cancer, the second most dangerous cancer in wome in the US is breast cancer. In the U. S. A. in 2021, the Society of American Cancer predicts that there will be 48,530 new instances of non-invasive (in situ) breast cancer and approximately 284,200 new cases of invasive breast



cancer. However, with an anticipated 685,000 deaths from breast cancer in 2020, it will continue to be a leading cause of death for women worldwide.[4] The World Health Organization (WHO) states that the leading cause of death for women is breast cancer (BC) [1].

A. The indications and manifestations of breast cancer include:-

Breast cancer symptoms can vary among individuals.

While some may not experience any symptoms, others may notice a new lump in the breast or underarm. Additional indications of breast cancer may involve enlargement or hardening of a specific breast area, puckering or inflammation of the breast skin, or the appearance of reddened or scaly skin surrounding the breast or nipple, nipple soreness or pulling inwards, discharge from the nipple that is not breast milk, any changes in the size or shape of the breast, and breast pain in any area.

Machine learning [2] and data mining methods are methods are utilized to predict breast cancer, and Identifying the appropriate algorithm for the job is the obstacle to overcome. Breast cancer emerges from malignant tumors that arise when cellular growth becomes unmanageable, resulting in a typical multiplication of adipose tissue. There are different types of breast cancer, such as DCIS, IDC, MTBC, LBC, MBC, and IBC.

These categories of breast cancer [4] happen when cancerous cells and tissues metastasize to other parts of the body.

DCIS is a form of breast cancer that is sometimes referred to as non-invasive cancer. It arises when abnormal cells grow beyond the breast. [5]

The second type of breast cancer is known as infiltrative ductal carcinoma (IDC) [7] or invasive ductal carcinoma, depending on the classification. IDC is more commonly found in men, and it happens when abnormal breast cells spread throughout the entire breast tissue.

Mixed Tumors Breast Cancer (MTBC) [9], also called invasive mammary breast cancer, is the third category of breast cancer. This type of cancer develops when abnormal cells in the ducts and lobules of the breast grow uncontrollably.

Lobular breast Cancer (LBC) is the fourth type of breast cancer and it originates in the lobules of the breast. This cancer increases the risk of developing more invasive types of breast cancer.

Colloid breast cancer, also known as mucinousbreast.is the fifth kind of breast cancer. It develops from invasive ductal cells, and it takes place when anomalous tissue surrounds the duct.

Inflammatory Breast Cancer (IBC) is the last form which leads to breast inflammation and redness. IBC is a rapidly growing cancer that occurs when cancer cells block the lymph vessels.

Finding Data mining involves the retrieval of significant details from an extensive collection of data. Various techniques, such as machine learning, statistics, databases, fuzzy sets, data warehouses, and neural networks, can be applied to analyze and predict the prognosis of various types of cancer, including prostate cancer, lung cancer, and leukemia[15, 16]. Unlike the conventional method of cancer diagnosis, which involves three tests --, radiological imaging, clinical examination and pathology examination[18] -- modern machine learning approaches and algorithms focus on creating models to forecast expected outcomes using unknown data. Machine learning is based on three primary methodologies: preprocessing, classification, feature selection or extraction. Feature extraction is particularly important, as it can help to find between benign and malignant tumors. The primary goal of creating a model is to forecast and obtain accurate results during both the training and testing phases, using data that was previously unknown. [21].

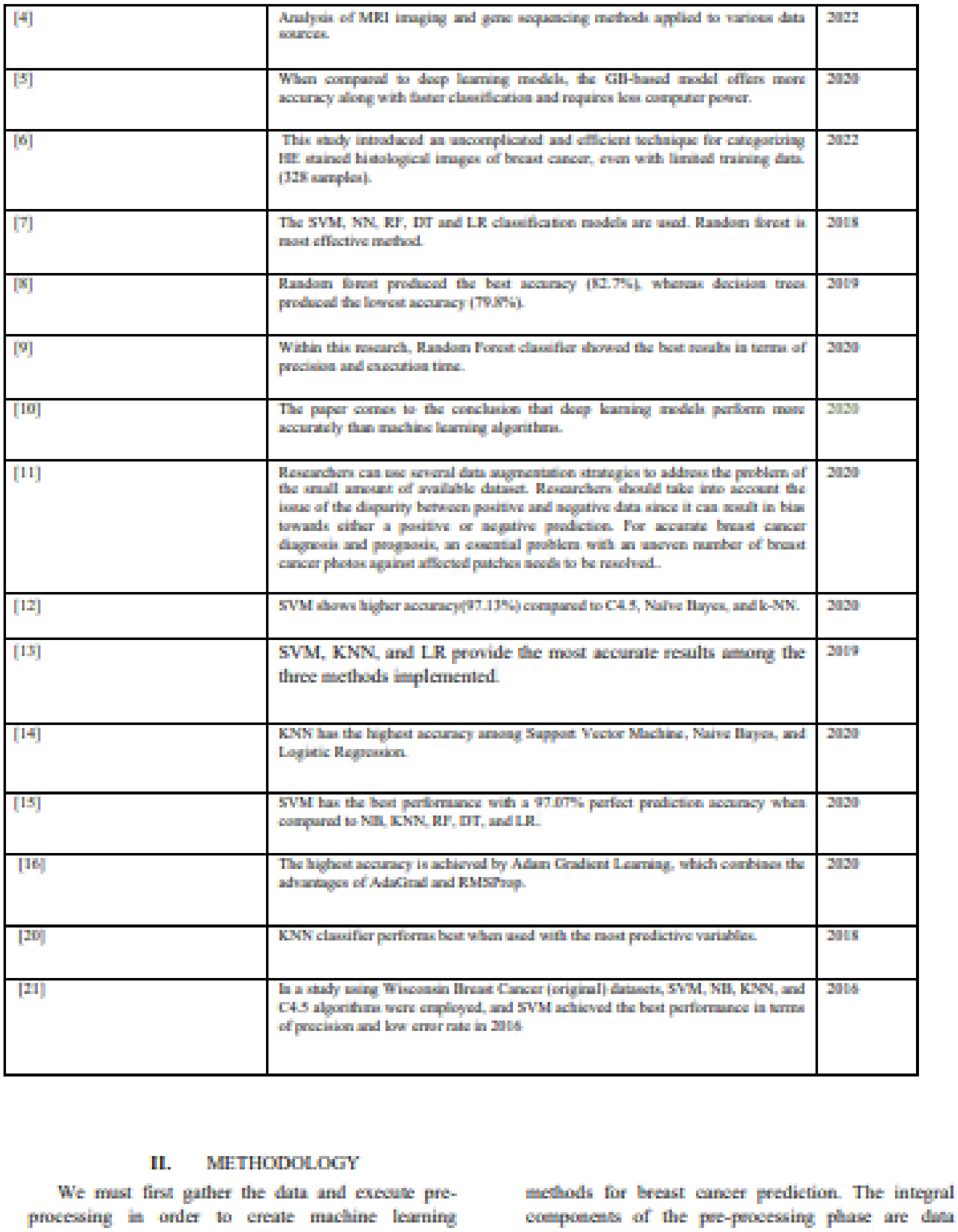
II. LITERATURE REVIEW

This section discusses about few of the previous work on breast cancer treatment ways done by researchers using various ML methods.

Table I show some of the things previously done on breast cancer diagnosis by researchers using different methods, including ML, deep learning, which are discussed.

TABLE I. REVIEW OF PAPERS

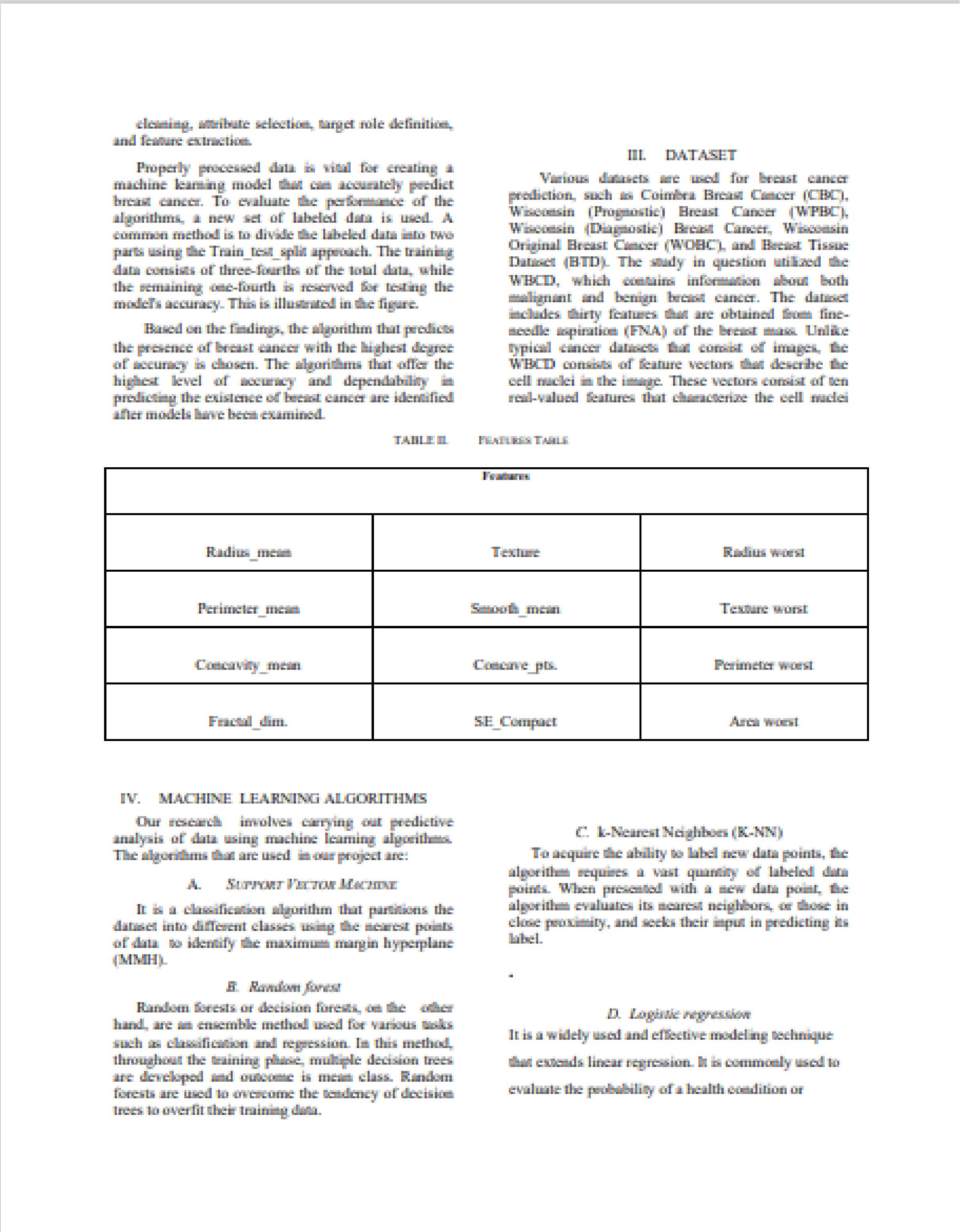
Reference	**Study**	**Published Year**
[3]	Proposed BCAD Framework, Multilayer Perceptron Model, 99.12% Accuracy.	2022



[4]	Analysis of MRI imaging and gene sequencing methods applied to various data sources.	2022
[5]	When compared to deep learning models, the GB-based model offers more accuracy along with faster classification and requires less computer power.	2020
[6]	This study introduced an uncomplicated and efficient technique for categorizing HE stained histological images of breast cancer, even with limited training data. (328 samples).	2022
[7]	The SVM, NN, RF, DT and LR classification models are used. Random forest is most effective method.	2018
[8]	Random forest produced the best accuracy (82.7%), whereas decision trees produced the lowest accuracy (79.8%).	2019
[9]	Within this research, Random Forest classifier showed the best results in terms of precision and execution time.	2020
[10]	The paper comes to the conclusion that deep learning models perform more accurately than machine learning algorithms.	2020
[11]	Researchers can use several data augmentation strategies to address the problem of the small amount of available dataset. Researchers should take into account the issue of the disparity between positive and negative data since it can result in bias towards either a positive or negative prediction. For accurate breast cancer diagnosis and prognosis, an essential problem with an uneven number of breast cancer photos against affected patches needs to be resolved..	2020
[12]	SVM shows higher accuracy(97.13%) compared to C4.5, Naïve Bayes, and k-NN.	2020
[13]	SVM, KNN, and LR provide the most accurate results among the three methods implemented.	2019
[14]	KNN has the highest accuracy among Support Vector Machine, Naive Bayes, and Logistic Regression.	2020
[15]	SVM has the best performance with a 97.07% perfect prediction accuracy when compared to NB, KNN, RF, DT, and LR.	2020
[16]	The highest accuracy is achieved by Adam Gradient Learning, which combines the advantages of AdaGrad and RMSProp.	2020
[20]	KNN classifier performs best when used with the most predictive variables.	2018
[21]	In a study using Wisconsin Breast Cancer (original) datasets, SVM, NB, KNN, and C4.5 algorithms were employed, and SVM achieved the best performance in terms of precision and low error rate in 2016	2016

II. METHODOLOGY

We must first gather the data and execute pre-processing in order to create machine learning methods for breast cancer prediction. The integral components of the pre-processing phase are data



cleaning, attribute selection, target role definition, and feature extraction.

Properly processed data is vital for creating a machine learning model that can accurately predict breast cancer. To evaluate the performance of the algorithms, a new set of labeled data is used. A common method is to divide the labeled data into two parts using the Train_test_split approach. The training data consists of three-fourths of the total data, while the remaining one-fourth is reserved for testing the model's accuracy. This is illustrated in the figure.

Based on the findings, the algorithm that predicts the presence of breast cancer with the highest degree of accuracy is chosen. The algorithms that offer the highest level of accuracy and dependability in predicting the existence of breast cancer are identified after models have been examined.

III. DATASET

Various datasets are used for breast cancer prediction, such as Coimbra Breast Cancer (CBC), Wisconsin (Prognostic) Breast Cancer (WPBC), Wisconsin (Diagnostic) Breast Cancer, Wisconsin Original Breast Cancer (WOBC), and Breast Tissue Dataset (BTD). The study in question utilized the WBCD, which contains information about both malignant and benign breast cancer. The dataset includes thirty features that are obtained from fine-needle aspiration (FNA) of the breast mass. Unlike typical cancer datasets that consist of images, the WBCD consists of feature vectors that describe the cell nuclei in the image. These vectors consist of ten real-valued features that characterize the cell nuclei

TABLE II. FEATURES TABLE

Features		
Radius_mean	Texture	Radius worst
Perimeter_mean	Smooth_mean	Texture worst
Concavity_mean	Concave_pts.	Perimeter worst
Fractal_dim.	SE_Compact	Area worst

IV. MACHINE LEARNING ALGORITHMS

Our research involves carrying out predictive analysis of data using machine learning algorithms. The algorithms that are used in our project are:

A. *SUPPORT VECTOR MACHINE*

It is a classification algorithm that partitions the dataset into different classes using the nearest points of data to identify the maximum margin hyperplane (MMH).

B. Random forest

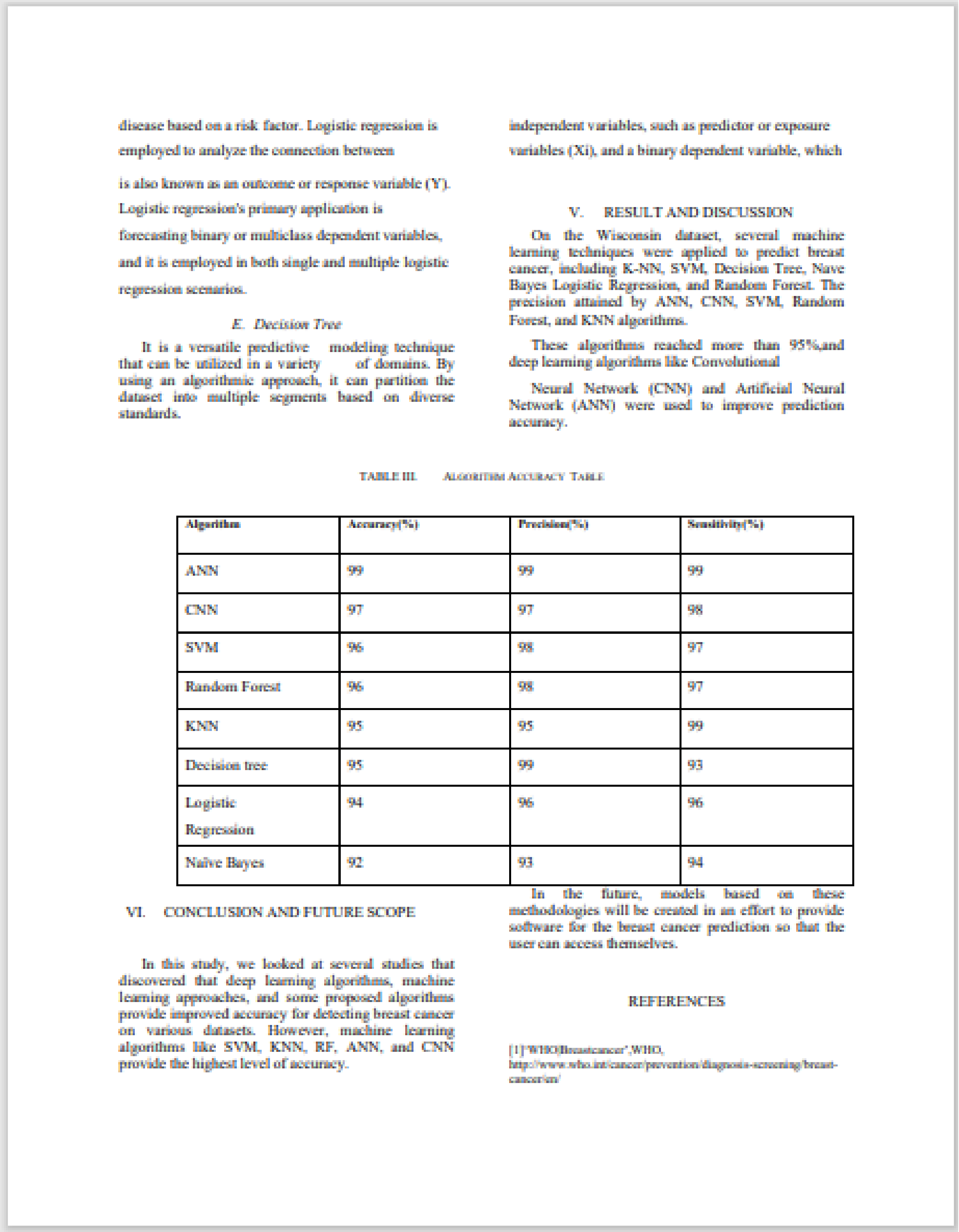
Random forests or decision forests, on the other hand, are an ensemble method used for various tasks such as classification and regression. In this method, throughout the training phase, multiple decision trees are developed and outcome is mean class. Random forests are used to overcome the tendency of decision trees to overfit their training data.

C. k-Nearest Neighbors (K-NN)

To acquire the ability to label new data points, the algorithm requires a vast quantity of labeled data points. When presented with a new data point, the algorithm evaluates its nearest neighbors, or those in close proximity, and seeks their input in predicting its label.

D. Logistic regression

It is a widely used and effective modeling technique that extends linear regression. It is commonly used to evaluate the probability of a health condition or



disease based on a risk factor. Logistic regression is employed to analyze the connection between

is also known as an outcome or response variable (Y). Logistic regression's primary application is forecasting binary or multiclass dependent variables, and it is employed in both single and multiple logistic regression scenarios.

E. Decision Tree

It is a versatile predictive modeling technique that can be utilized in a variety of domains. By using an algorithmic approach, it can partition the dataset into multiple segments based on diverse standards.

independent variables, such as predictor or exposure variables (Xi), and a binary dependent variable, which

V. RESULT AND DISCUSSION

On the Wisconsin dataset, several machine learning techniques were applied to predict breast cancer, including K-NN, SVM, Decision Tree, Nave Bayes Logistic Regression, and Random Forest. The precision attained by ANN, CNN, SVM, Random Forest, and KNN algorithms.

These algorithms reached more than 95%,and deep learning algorithms like Convolutional

Neural Network (CNN) and Artificial Neural Network (ANN) were used to improve prediction accuracy.

TABLE III. ALGORITHM ACCURACY TABLE

Algorithm	**Accuracy(%)**	**Precision(%)**	**Sensitivity(%)**
ANN	99	99	99
CNN	97	97	98
SVM	96	98	97
Random Forest	96	98	97
KNN	95	95	99
Decision tree	95	99	93
Logistic Regression	94	96	96
Naïve Bayes	92	93	94

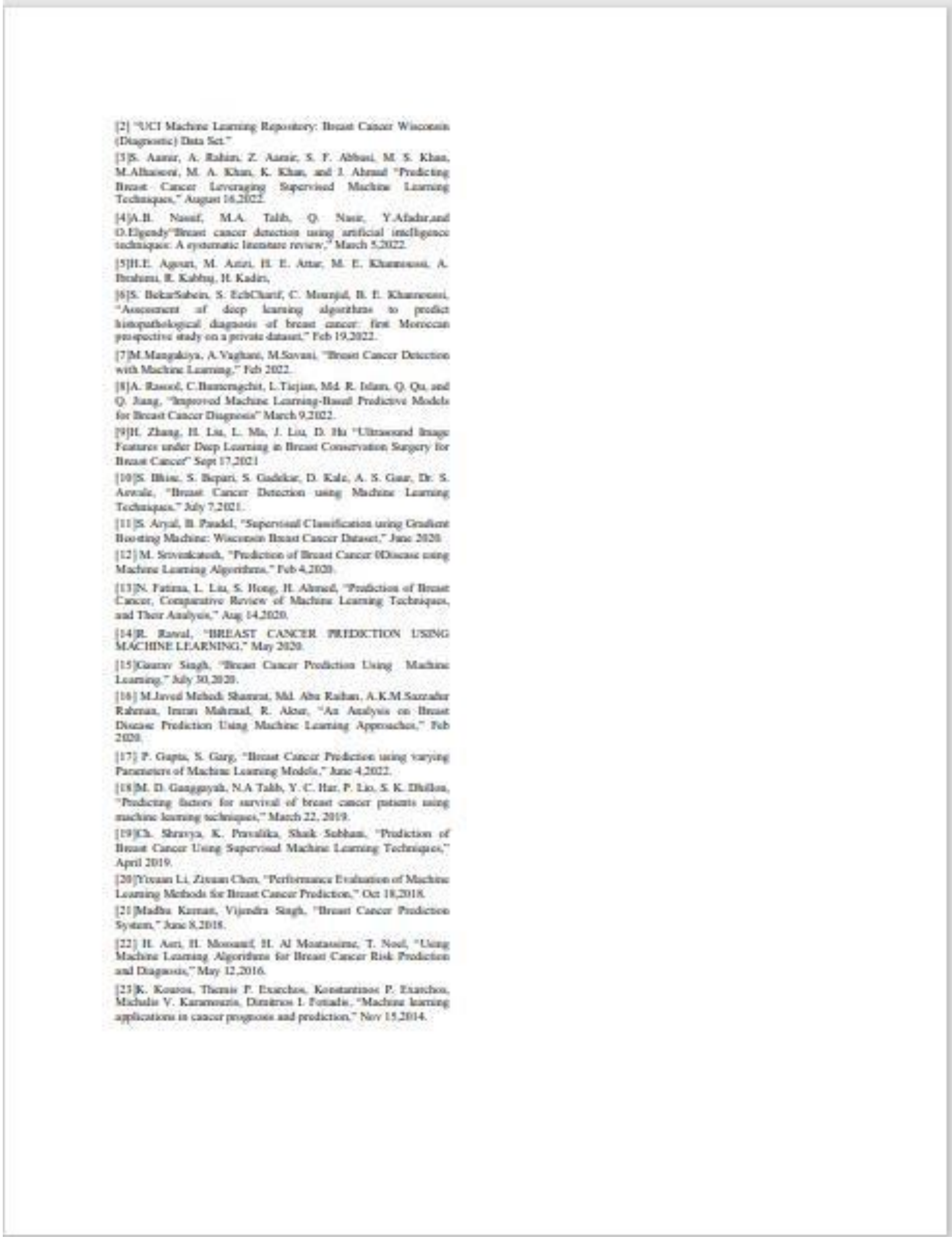
VI. CONCLUSION AND FUTURE SCOPE

In this study, we looked at several studies that discovered that deep learning algorithms, machine learning approaches, and some proposed algorithms provide improved accuracy for detecting breast cancer on various datasets. However, machine learning algorithms like SVM, KNN, RF, ANN, and CNN provide the highest level of accuracy.

In the future, models based on these methodologies will be created in an effort to provide software for the breast cancer prediction so that the user can access themselves.

REFERENCES

[1]'WHO|Breastcancer',WHO, http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/



[2] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set."

[3]S. Aamir, A. Rahim, Z. Aamir, S. F. Abbasi, M. S. Khan, M.Alhaisoni, M. A. Khan, K. Khan, and J. Ahmad "Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques," August 16,2022.

[4]A.B. Nassif, M.A. Talib, Q. Nasir, Y.Afadar,and O.Elgendy"Breast cancer detection using artificial intelligence techniques: A systematic literature review," March 5,2022.

[5]H.E. Agouri, M. Azizi, H. E. Attar, M. E. Khannoussi, A. Ibrahimi, R. Kabbaj, H. Kadiri,

[6]S. BekarSabein, S. EchCharif, C. Mounjid, B. E. Khannoussi, "Assessment of deep learning algorithms to predict histopathological diagnosis of breast cancer: first Moroccan prospective study on a private dataset," Feb 19,2022.

[7]M.Mangukiya, A.Vaghani, M.Savani, "Breast Cancer Detection with Machine Learning," Feb 2022.

[8]A. Rasool, C.Bunterngchit, L.Tiejian, Md. R. Islam, Q. Qu, and Q. Jiang, "Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis" March 9,2022.

[9]H. Zhang, H. Liu, L. Ma, J. Liu, D. Hu "Ultrasound Image Features under Deep Learning in Breast Conservation Surgery for Breast Cancer" Sept 17,2021

[10]S. Bhise, S. Bepari, S. Gadekar, D. Kale, A. S. Gaur, Dr. S. Aswale, "Breast Cancer Detection using Machine Learning Techniques," July 7,2021.

[11]S. Aryal, B. Paudel, "Supervised Classification using Gradient Boosting Machine: Wisconsin Breast Cancer Dataset," June 2020.

[12] M. Srivenkatesh, "Prediction of Breast Cancer 0Disease using Machine Learning Algorithms," Feb 4,2020.

[13]N. Fatima, L. Liu, S. Hong, H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," Aug 14,2020.

[14]R. Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING," May 2020.

[15]Gaurav Singh, "Breast Cancer Prediction Using Machine Learning," July 30,2020.

[16] M.Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M.Sazzadur Rahman, Imran Mahmud, R. Akter, "An Analysis on Breast Disease Prediction Using Machine Learning Approaches," Feb 2020.

[17] P. Gupta, S. Garg, "Breast Cancer Prediction using varying Parameters of Machine Learning Models," June 4,2022.

[18]M. D. Ganggayah, N.A Taib, Y. C. Har, P. Lio, S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," March 22, 2019.

[19]Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques," April 2019.

[20]Yixuan Li, Zixuan Chen, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction," Oct 18,2018.

[21]Madhu Kumari, Vijendra Singh, "Breast Cancer Prediction System," June 8,2018.

[22] H. Asri, H. Mousannif, H. Al Moatassime, T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," May 12,2016.

[23]K. Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Nov 15,2014.