JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Implementation of Machine Learning in Early Detection of Breast Cancer

[1]Namrata Kataki, [2]Sanjana Dutta Choudhury, [3]Hrishikesh Gogoi, [4]Swaraj Bora, [5]Pratyosh Pawan Bora

[1] Assistant Professor, [2,3,4,5]U.G Scholar

Department of ECE, Girijananda Chowdhury Institute of Management & Technology,
Guwahati, India

Abstract : Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. The objective of this paper is to describe a predictive model built using Machine Learning algorithms for early detection of breast cancer in order to improve the prognosis and chances of survival through timely clinical treatment to patients. Most influential machine learning algorithms have been used for comparison in terms of accuracy, sensitivity, specificity and precision. Data mining approaches, for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients' health, improving healthcare value and quality and in making real time decision to save people's lives. Therefore, the correct diagnosis of breast cancer and classification of patients into malignant and benign groups is the center of a lot of research. The early diagnosis of breast cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments.
.

IndexTerms - Breast Cancer, Machine Learning, Classification, Malignant, Benign

I. Introduction

According to the world health organization (WHO) Breast cancer is the most frequent cancer among women, impacting 2.1 million women each year and also causes the greatest number of cancer related deaths among women. In 2018, it is estimated that 627,000 women died from breast cancer – that is approximately 15% of all cancer deaths among women. While breast cancer rates are higher among women in more developed regions, rates are increasing in nearly every region globally. In order to improve breast cancer outcomes and survival, early detection is critical. There are two early detection strategies for breast cancer: early diagnosis and screening. Limited resource settings with weak health systems where the majority of women are diagnosed in late stages should prioritize early diagnosis programs based on awareness of early signs and symptoms and prompt referral to diagnosis and treatment. Early diagnosis strategies focus on providing timely access to cancer treatment by reducing barriers to care and/or improving access to effective diagnosis services. The goal is to increase the proportion of breast cancers identified at an early stage, allowing for more effective treatment to be used and reducing the risks of death from breast cancer. Since early detection of cancer is the key to effective treatment of breast cancer we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data.

II. Some risk factors for breast cancer

Most cases of breast cancer cannot be linked to a specific cause. However following are some of the known risk factors for breast cancer.

1) **Age.** The chance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50.
2) **Personal history of breast cancer**. A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast.
3) **Family history of breast cancer**. A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk.
4) **Genetic factors.** Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.
5) **Childbearing and menstrual history.** The older a woman is when she has her first child, the greater her risk of breast cancer. Some other risk factors of breast cancer are –

a) Women who menstruate for the first time at an early age (before 12)
b) Women who go through menopause late (after age 55)
c) Women who've never had children.

III. ROLE OF MACHINE LEARNING IN DETECTION OF BREAST CANCER

A mammogram is an x-ray picture of the breast. It can be used to check for breast cancer in women who have no signs or symptoms of the disease. It can also be used if there is any lump or other sign of breast cancer. Screening mammography can help in reducing the number of deaths from breast cancer among women ages 40 to 70. But it can also have drawbacks. Mammograms can sometimes find something that looks abnormal but isn't cancer. This leads to further testing and can cause anxiety and harassment of patients. Sometimes mammograms can miss cancer when it is there. It also exposes the patient to radiation. Now while it is difficult to figure out for physicians by seeing only images of x-ray that weather the tumor is toxic or not training a machine learning model according to the identification of the tumor can be of great help.

IV. PROPOSED SYSTEM

This analysis aims to define the most helpful features in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant based upon tumour features i.e., its radius, area, smoothness, texture, perimeter. To achieve that we have used machine learning classification methods to fit a function that can predict the discrete class of new input. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, and then it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector.
Ten real-valued features are computed for each cell nucleus:

(1) Radius (mean of distances from center to points on the perimeter), (2) Texture (standard deviation of gray-scale values), (3) Perimeter, (4) Area, (5) Smoothness (local variation in radius lengths), (6) Concavity (severity of concave portions of the contour), (7) Compactness (perimeter² / area — 1.0), (8) Concave points (number of concave portions of the contour), (9) Symmetry, (10) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

V. DATA MINING AND MACHINE LEARNING

The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate. In this project we have used the following machine learning algorithm-

1. Decision tree algorithms
2. Logistic regression
3. Random Forest Classifier

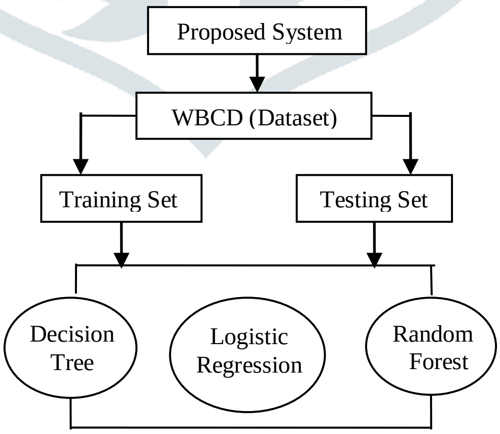

Fig 1: Block diagram of the proposed system

.

VI. . DATA EXPLORATION AND PROCESSING

. We will first go with importing the necessary libraries and import our dataset. We can examine the data set using the pandas.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	patient_id	patient_id2	patient_id3	patient_id4	patient_id5	patient_id6	patient_id7	patient_id8	patient_id9	patient_id10	patient_id11	patient_id12	radius_se	texture_se	perimeter_se	area_se	smoothnes
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.0
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.0
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.0
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.0
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.0
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.0
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.0
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.0
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.0
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.0
16	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.0
17	84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.0
18	848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.0
19	84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.0
20	849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.0
21	8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.0
22	8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.0
23	8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.0
24	8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.0
25	851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.0
26	852552	M	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8068	0.9017	5.455	102.6	0.0
27	852631	M	17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.046	0.976	7.276	111.4	0.0
28	852763	M	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.2545	0.9832	2.11	21.05	0.0

breast_cancer

Fig 2: A part of used data set

We can observe that the data set contain 569 rows and 33 columns. 'Diagnosis' is the column which we are going to predict, which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant). Each row represents a patient and 33 features on the 569 patients. The last column Unnamed 32 has unnamed values so we need to remove that column with empty values (Fig 3). So, we count the number of empty columns and drop the columns with empty values (fig 4).

```
id                          0
diagnosis                   0
radius_mean                 0
texture_mean                0
perimeter_mean              0
area_mean                   0
smoothness_mean             0
compactness_mean            0
concavity_mean              0
concave points_mean         0
symmetry_mean               0
fractal_dimension_mean      0
radius_se                   0
texture_se                  0
perimeter_se                0
area_se                     0
smoothness_se               0
compactness_se              0
concavity_se                0
concave points_se           0
symmetry_se                 0
fractal_dimension_se        0
radius_worst                0
texture_worst               0
perimeter_worst             0
area_worst                  0
smoothness_worst            0
compactness_worst           0
concavity_worst             0
concave points_worst        0
symmetry_worst              0
fractal_dimension_worst     0
Unnamed: 32               569
dtype: int64
```

Fig 3: Columns with empty values

```
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
dtype: object
```

Fig 4: Columns with non empty value

So, column Unnamed: 32 has 569 missing values so we drop it. So, the new shape of the data is (569, 32) which means 569 rows and 32 columns. Now we can see the number of Malignant (M) (harmful) or Benign (B) cells (not harmful) cells and plot it in a graph (Fig 5). We can see that id column acts as the identifier of the patient and it is of integer type and it cannot be used as a feature to predict the tumor. Next, we encode categorical data values (Fig 6). Here the value 1 represents Malignant (M) (harmful) and value 0 represents Benign (B) cells (not harmful) cells. Now, we visualize a correlation between the different attributes.

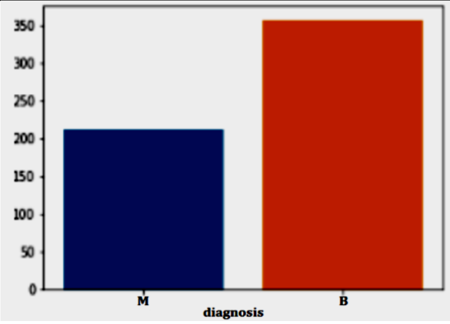

Fig: 5: Graph of Malignant and Benign data

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1
0 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 1 1
0 1 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0
0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 0 0 1 0
0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 0 0 0
0 1 0 0 0 1 0 0 1 1 0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1
0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1
1 1 1 1 1 1 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0
0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0
0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0
1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0
0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 1
1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 1 1 1 1 1 0]
```

Fig 6: Encoding of categorical data values

Confusion Matrix: Confusion matrix is used to describe the performance of the classification model. It is basically a N*N matrix which compares the actual target values with those predicted by the machine learning model.

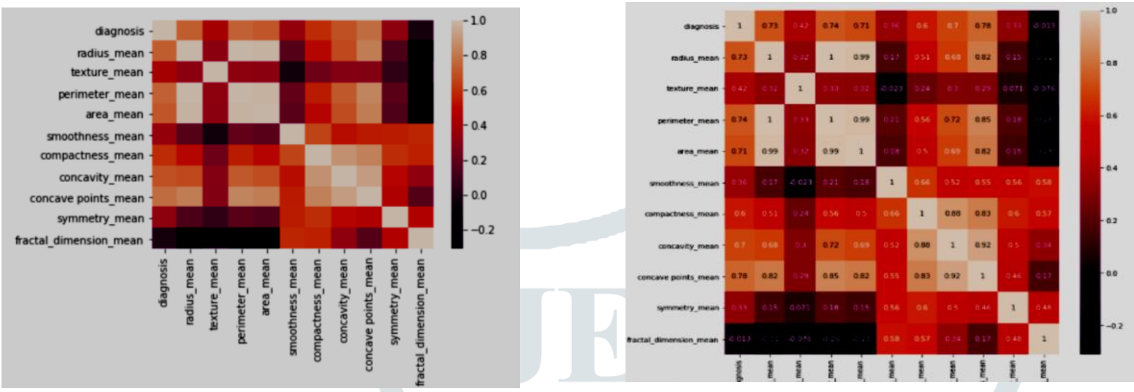

Fig 7: Convolution matrix of the models used.

In the above heat map we can see how much one column influences all the other columns (e.g., radius mean has 32% influences on texture meanTraining and Testing. Next we split the datasets into independent (X) and dependent (Y) datasets. They are of type array.

X = [: , 2 : 31].values Y =[: , 1].values.

The dependent data set(Y) has the diagnosis whether the patient has cancer and the independent dataset(X) has the features that are used to predict the outcome. Now we split the dataset into 75% training and 25% testing and use different machine learning models such as logistic regression, random forest classifier, decision tree to the training set. After that we print the accuracy on the training data

VII. CLASSIFICATION REPORT

A classification report is used to measure the quality of predictions from a classification algorithm. The report shows the main classification metrics precision, recall and f1- score on a per- class basis. The matrices are calculated by using true and false positives, true and false negatives. Positive and negative in this case are generic names for the predicted classes. There are four ways to check if the predictions are right or wrong. Here the matrices are of form [TP FP] [FN TP] where,

- TP/True Positive: when a case was positive and predicted positive.
- TN/True Negative: when a case was negative and predicted negative.
- FN/False Negative: when a case was negative but predicted negative.
- FP/ False Positive: when a case was negative but predicted positive

Prediction of Model:

pred = model[5]. predict(X-test)
print(pred)

•

So here we printed the predictions. The first data shows the actual result of which patient had cancer and the second data is the one predicted by the model. The accuracy of the model is 95% so we can see a few wrong predictions but mostly this model is successful in predicting a tumor Malignant (M) (harmful) or Benign (B) (not harmful) based upon the features provided in the data and the training given

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 0
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]

[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]
```

Fig 7: Prediction of the models used

VIII. RESULT AND DISCUSSION

The performance of the algorithms differed with and without principal component analysis implementation on the dataset. Analysis and comparison of the performance of different models implemented on the test portion of the dataset was evaluated across various performance metrics.

*Performance Metrics***:** This paper deals with classification problem and therefore the chosen performance metrics primarily focus on classification. For the detection of breast cancer, if the target variable is 1 then it is a positive instance, meaning the patient has a malignant tumor and therefore cancer. And if the target variable is 0, then it a negative instance, meaning the tumor is benign and the patient does not have cancer.

*Confusion matrix***:** The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for classification problems where the output can be of two or more types of classes which make it perfect for this paper. The table layout or the matrix layout helps to visualize the performance of an algorithm. Each row of the matrix in table 4.1 represents the instances in an actual class while each column represents the instances in a predicted class or vice versa.

Table 1: Confusion Matrix

	PREDICTIVE NEGATIVE	**PREDICTIVE POSITIVE**
ACTUAL NEGATIVE	TRUE NEGATIVE (TN)	FALSE POSITIVE(FP)
ACTUAL POSITIVE	FALSE POSITIVE (FN))	TRUE POSITIVE(TP)

True Positives (TP): True positives are the cases when the actual class of the data point was True(1) and the predicted is also True(1) Ex: The case where a person is actually having malignant (1) tumor and the model classifying his case as malignant (1) comes under True Positive.

True Negatives (TN): True negatives are the cases when the actual class of the data point was False (0) and the predicted is also False (0). Ex: The case where a person having benign (0) tumor and the model classifying his case as benign (0) comes under True Negatives.

False Positives (FP): False positives are the cases when the actual class of the data point was False (0) and the predicted is True(1).False is because the model has predicted incorrectly and positive because the class predicted was a positive one (1). Ex: A person having a benign (0) tumor and the model classifying his case as malignant (1) comes under False Positives.

False Negatives (FN): False negatives are the cases when the actual class of the data point was True (1) and the predicted is False (0). False is because the model has predicted incorrectly and negative because the class predicted was a negative one (0).Ex: A person having malignant (1) tumor and the model classifying his case as benign (0) tumor comes under False Negatives. The ideal scenario for the model would be when it gives 0 False Positives and 0 FalseNegatives.

a. **Accuracy:** Accuracy in classification problems is the number of correct predictions made by the model over the summation of all different types of predictions made. Accuracy is a good measure when the target variable classes in the data are nearly balanced. Ex: 60% of the data are benign and 40% are malignant.

   Accuracy =TP+TN/TP+ FP+ T N +FN

b. **Precision:** Precision is the ratio of True Positives to the summation of True Positives and False Positives. Ex: Precision is a measure of proportion of patients that has been diagnosed as having malignant tumor, actually had malignant tumor. The predicted positives (People predicted as having malignant tumor are TP and FP) and the people actually having a malignant tumor are TP.

   Precision =T P/TP+FP

c. **Recall or Sensitivity:** Recall is a measure that shows the proportion of patients that actually had malignant tumor was diagnosed by the algorithm as having malignant tumor. The actual positives (People having malignant tumor are TP and FN) and the people diagnosed by the model having a malignant tumor are TP. Therefore, if we want to focus more on minimizing False Negatives, we would want our Recall to be as close to 100% as possible.

   Recall = TP/TP + FN

d. **F1 Score:** F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0,1]. It shows how precise the classifier is and how robust it is at the same time.

   F1= 2 ×Precision ×Recall/ Precision + Recall

After completing the implementation of all seven algorithms for detecting breast cancer from the dataset, the results can be compared from the table 1 using the performance metrics.

Table 2: Comparison of Scores of Various Models (Average)

	Accuracy	**Precision**	**Recall**	**F1 score**
Decision tree	95%	95%	95%	95%
Logistic Regression	95%	95%	95%	95%
Random Forest Classifier	95%	95%	95%	95%

Snapshots:

Out[8]:

	patient_id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	ra
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	...	
5	843786	M	12.45	15.70	82.57	477.1	0.12780	0.17000	0.15780	0.08089	...	
6	844359	M	18.25	19.98	119.60	1040.0	0.09463	0.10900	0.11270	0.07400	...	
7	84458202	M	13.71	20.83	90.20	577.9	0.11890	0.16450	0.09366	0.05985	...	
8	844981	M	13.00	21.82	87.50	519.8	0.12730	0.19320	0.18590	0.09353	...	
9	84501001	M	12.46	24.04	83.97	475.9	0.11860	0.23960	0.22730	0.08543	...	

10 rows × 32 columns

Fig 8: Snapshot of data processing

```
Model 0
              precision    recall  f1-score   support

           0       0.97      0.96      0.96        90
           1       0.93      0.94      0.93        53

    accuracy                           0.95       143
   macro avg       0.95      0.95      0.95       143
weighted avg       0.95      0.95      0.95       143

0.951048951048951
Model 1
              precision    recall  f1-score   support

           0       0.97      0.96      0.96        90
           1       0.93      0.94      0.93        53

    accuracy                           0.95       143
   macro avg       0.95      0.95      0.95       143
weighted avg       0.95      0.95      0.95       143

0.951048951048951
Model 2
              precision    recall  f1-score   support

           0       0.97      0.96      0.96        90
           1       0.93      0.94      0.93        53

    accuracy                           0.95       143
   macro avg       0.95      0.95      0.95       143
weighted avg       0.95      0.95      0.95       143

0.951048951048951
```

Fig: 9 : Snapshot of Accuracy calculation for different parameter

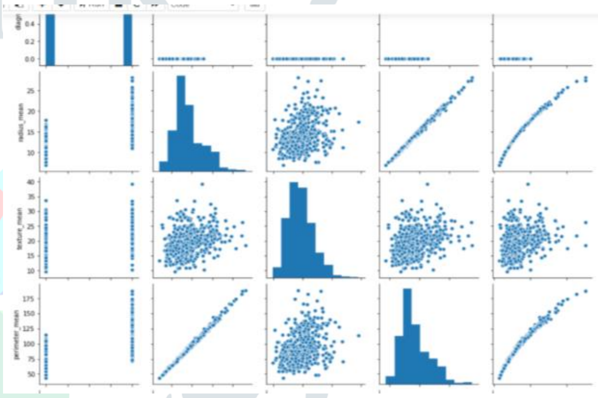

Fig 10: Snapshots of result analysis

IX. Challenges and Future Scope

Machine learning (ML) has attracted much attention with the hope that it can provide the accurate results, but its modeling methods and prediction performance remains controversial sometimes. The aim of this systemic review is to classify and critically appraise current studies regarding the application of ML in predicting the survival rate of breast cancer. In this project in python, we learned to build a breast cancer tumor predictor on the Wisconsin dataset and created graphs and results for the same. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases. In the future, the model can be perfected with the increase in availability of data and most importantly the growth of data. Deep learning models perform proportionately well to the amount of data which means there is space to improve with the availability of a large datasets.

X. Conclusion

Breast cancer, the most common cancer among women, being responsible for 69% of death related to cancer among the same gender gives a glimpse of the magnitude of the problem. The early detection of breast cancer can lead to chances of survival of a large number of these people by receiving clinical treatments on time. This paper shows the comparative analysis of different machine learning algorithms in detecting breast cancer from a digitized image of a fine needle aspirate (FNA) of a breast mass. The simple, safe, accurate, and inexpensive procedure of FNA combined with the predictive model in this paper can be used for prognosis, diagnosis and assist doctors in making the final decision more accurately in shorter time span with less human and monetary resource. The performance of the deep learning models shows promising results for a near perfect detection system

REFERENCES

[1]. **[1]** A Kamboj, P Tanay, A Sinha, -*Breast cancer detection using supervised machine learning: a comparative analysis*, Springer, pp. 263-269, Singapore (2021).

[2]. Hussain, M., Wajid, S. K., Elzaart, A., & Berbar, M. A,- *Comparison of SVM kernel functions for breast cancer detection* Eighth International Conference Computer Graphics, Imaging and Visualization IEEE. August, 2011, (pp. 145–150).

[3]. Janghel, R., Shukla, A., Tiwari, R., and Kala, R. -*Breast cancer diagnosis using artificial neural network models*, Information Sciences and Interaction Sciences (ICIS), 3rd International Conference on, pages 89–94. IEEE.2010

[4]. Delen, D., Walker, G., and Kadam, A - *Predicting breast cancer survivability: a comparison of three data mining methods*, 2005 Artificial intelligence in medicine, 34(2):113– 127.

[5]. Ho, T. K.- *Random decision forests. In Document analysis and recognition*, proceedings of the third international conference on, volume 1, pages 278–282. IEEE.

[6]. Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. - *Applied logistic regression*, volume 398. John Wiley & Sons. 2013

[7]. Hussain, M., Wajid, S. K., Elzaart, A., and Berbar, M. - *A comparison of SVM kernel functions for breast cancer detection*, Eighth International Conference Computer Graphics, Imaging and Visualization, pages 145–150. IEEE, 2011.

[8]. Janghel, R., Shukla, A., Tiwari, R., and Kala, R. - *Breast cancer diagnosis using artificial neural network models*, Information Sciences and Interaction Sciences (ICIS), 3rd International Conference on, pages 89–94. IEEE, 2010

[9]. Mohammed A. N., Sanaa E. F., Kawtar A., EL Habib B., Rachida A. A., Olivier D.- *Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis*, Procedia Computer Science, Volume 191, 2021, Pages 487-492, ISSN 1877-0509

[10]. B.M. Gayathri and C. P. Sumathi,- *Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer*, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1–5.

[11]. A.H. Osman -*An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique*, International Journal of Advanced Computer. Science. Applications., 8 (4) (2017), pp. 158-165

[12]. Tahmooresi, Maryam & Afshar, A & Bashari Rad, Babak & Nowshath, K & Bamiah, Mervat. -*Early Detection of Breast Cancer Using Machine Learning Techniques*, Journal of Telecommunication, Electronic and Computer Engineering. 10. 21-27. (2018).