

# Analysis of Machine Learning Classifiers for Breast Cancer Diagnosis

<sup>1</sup>Prof. Satish Manje, <sup>2</sup>Mr.Sainathreddy Bonthu, <sup>3</sup>Mr.Arun Elavarasan, <sup>4</sup>Mr.Aniket Gupta

<sup>1</sup>Asst.Professor, <sup>2,3,4</sup>UG Student, <sup>1,2,3,4</sup>Computer Engg. Dept. ShivajiraoS.Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India. <sup>1</sup>*satishmanje93@gmail.com*, <sup>2</sup>*bonthusaireddy@gmail.com*, <sup>4</sup>*arunelavarasan168@gmail.com*, <sup>3</sup>*aniket.gupta.r@gmail.com*

**Abstract** - Breast cancer is a form of cancer which originates in breasts of women. Cancer is caused by uncontrolled cell division or expansion. Breast cancer cells usually form a tumor that can be seen on an X-ray. Breast cancer has become one of the most well-known illnesses among women, resulting in the death of the most prevalent malignancies in women. Early treatment helps to cure malignant growth and prevent its recurrence. The focus of this project is to develop a model that can predict whether a cancer is malignant or benign at an early age. This will be performed using a variety of machine learning algorithms (1) Logistic Regression Algorithm (2) Support Vector Machine Algorithm (3) Random Forest Algorithm Our objective is to detect the tumor and classify it as malignant or benign center on the discrete feature given in the dataset. This is frequently referred to as a classification problem. To achieve this, data mining techniques are used to clean up the dataset that has been gathered and then apply certain Machine Learning classification techniques to construct a model that tells that which of the candidate is Benign or Malignant. This also analyze which features are more useful in the prediction and classification of the malignancy and perform model selection by observing certain general patterns by correlating the features.

**Keywords-** Machine Learning, Healthcare, Breast Cancer Diagnosis, Support Vector Machine (SVM) Classification.

## I. INTRODUCTION

Breast cancer is most common cancer type in women it is second major reason of cancer death after lung cancer it starts off when malignant cancer cells start to grow from the breast cells. There is sometimes the doctor might diagnosis patient to having a benign tumor is not cancer instant of the malignant and hence advanced systems supporting machine learning should be used to help with early detection machine learning has become a widely used approach in the healthcare field due to high performance in predicting results reducing the cost of drug resulting patients good health valuing the quality of medical care being provided and in making concrete choice to save lives it is a kind of artificial intelligence that works or deals with the development computer programs with the aid of computer models and information from different sets of data to help in the process of classification predication and the deduction process this paper mainly focuses on the variety of machine learning algorithm used to classify breast cancer as benign or malignant based upon many other factors or terms. Early detection of breast cancer can raise the prediction and chance of survivors significantly and machine learning in support of evidence results with good accuracies have helped achieve this moreover a good classification technique aids doctors in detecting benign tumor s at an early stage, saving patients

from needless therapies. the whole experiment was Implemented using rstudio and the data set which was used the Wisconsin data set which was of computers found from the uci machine learning repository and this data head is created by William who was is a physician at the organization of Wisconsin hospital in Madison Wisconsin united stated of America.

## II. AIMS AND OBJECTIVE

### a) Aim

The focus of this work is that make an efficient system able to classify breast tumor s as malignant and benign. This system is split in two stages. The primary stage is the normalization of the data. The secondary stage is the classification of tumor s. This project gives accuracy 98.42%. The overall result showed that the DL outperformed the previous studies where the same data set was used.

### b) Objective

- To analyses the WBCD dataset for finding relation between the features.
- To apply different established classifier with the WBCD dataset for comparing them.

- To find most satisfactory approach supporting the dataset with good prediction accuracy.
- To classify of breast cancer whether benign and malignant using machine learning system.
- Using a machine learning system to detect breast cancer early.
- Using a machine learning system, reduce breast cancer classification errors during processing.

### III. LITERATURE SURVEY

#### **Paper 1: Statistical and Local Binary Pattern Features for Breast Mass Classification:**

Breast cancer is the known cancers among women, and it is typically fatal. Imprecision or a delay in diagnosis are the major causes of death. Early treatment aids in the cure and prevention of malignant development. The purpose of the project is designing a model that can accurately detect and categorize tumors. In order to achieve this, three Machine Learning (ML) algorithm are: Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF), which are used for accuracy. Algorithms are broadly used for diagnostic purposes in the medical industry because of their efficiency in data classification. Precision, recall, ROC area, and accuracy were used to assess the effectiveness of SVM. Through testing, the SVM algorithm method is shown to have the highest accuracy and give the best results. In research paper that this identify the Breast Cancer using the perfect suitable method and got 99.5% as most authentic result using random forest method.

#### **Paper 2: Breast cancer recognized from biopsy images with highly reliable random subspace classifier ensembles:**

The rejection option was implemented for both ensembles by connecting the majority voting consensus degree to a confidence measure and refusing to classify ambiguous samples if the consensus degree declined below a certain threshold. The efficiency of the recommended cascade classification technique was tested with a breast cancer biopsy image data. The combined feature representation using Gray Level Co-occurrence Matrix LBP texture description,

and Curvelet Transform takes advantage of the complimentary characteristics of several feature extractors; In the job of classifying biopsy images, the combined characteristics was shown to be useful. The two-stage ensemble cascade classification technique achieved high classification accuracy (99.25%) while also ensuring high classification reliability (97.65%) and a low rejection rate

(1.94 percent). The cascade architecture also includes a technique for balancing classification accuracy and rejection rate. Although the suggested approach has shown decent results in the classification of biopsy images, there are still certain elements that need to be examined further. The cascade system's parameters, such as ensemble size and number of steps, are discussed in this study. The rejection threshold was determined by trial and error; this may not have yielded the best results. In all application settings, satisfactory performance is achieved. As a result, several self-adaptive rules or algorithms for automatically adjusting the parameters for the cascade system, such as ensemble size and rejection threshold, were determined empirically in this paper; this may not have resulted in the best results. In all application settings, satisfactory performance is achieved.

#### **Paper 3: Breast Cancer Histopathological Image Classification Using Convolutional Neural Networks:**

The research shows a set of experiments utilizing a deep learning strategy to avoid created characteristics on the Breast dataset. This demonstrated that it could modify an existing CNN architecture, in this case AlexNet, that was built for categorizing color photos of objects to classify BC histopathology images. In this project also proposed several training strategies for CNN architectures, based on the extraction of patches obtained randomly or by sliding a window mechanism, that allow it to take care the high-resolution of these textured images without having to change CNN architectures designed for low-resolution images. The major purpose is to divide histology photos into separate histopathology patterns that correspond to whether the tissue is non-cancerous or malignant. Dealing with the inherent complexity of histopathology pictures is the system's main concerns. The CNN has been widely employed to produce outcomes in a types of pattern recognition problems. CNN is able to outperform standard textural descriptors in photos of microscopic and macroscopic textures. This employ multiple ways to cope with high resolution texture images without affecting the CNN architecture used for low resolution images, in addition to evaluating different CNN designs. The ambition is to retain the tissue's natural structure and molecular. The rejection threshold was determined by trial and error; this may not have yielded the best results. As a result, several self-adaptive rules or algorithms for automatically adjusting. The goal is to preserve the original tissue structures and molecular composition allowing to observe it in a light microscope. CNN has achieved success in image classification problem including medical image analysis. CNN consists of multiple trainable stages stacked on top of each other followed by supervised classifier and feature maps.

IV. COMPARTIVE STUDY

SR NO.	PAPER TITLE	AUTHOR /PUBLICATION	TECHNOLOGY	PURPOSE	ADVANTAGES
1.	Breast Mass Classification using Statistical and Local Binary Pattern Features.	Dmitri Krioukov, 2019 XLV Latin American Computing Conference (CLEI)	Local binary pattern (LBP), SVM, k-nearest neighbor (KNN), Digital Database for Screening Mammography (DDSM),	To build a model to detect and correctly classify the tumor with high accuracy.	The system classifies normal from abnormal cases with high accuracy rate.
2.	Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles	Fabiano Teixeira, IEEE, 2019	Local Binary Pattern (LBP), Gray Level Co-occurrence Matrix (GLCM), Random subspace ensemble.	To Classify Breast cancer, the parameters for the cascade system, such as ensemble size and rejection threshold, were decided empirically	Provides a comprehensive biopsy image characterization by taking advantages of their complementary strengths. a high classification accuracy of 99.25 % was obtained (with a rejection rate of 1.94 %) using the proposed system.
3.	Breast Cancer Histopathological Image Classification Using Convolutional Neural Networks	David A. Omondiagbe, Amandeep S. Sidhu, IICSPI, 2019	AlexNet, CNN, Speech Recognition, Signal Processing, Object Recognition.	The objective is to retain the tissue's natural structure and molecular makeup so that it may be examined under a light microscope.	CNN Performance is better when it compared with previously obtained by other machine learning model trained with hand-crafted textural descriptor.
4.	Machine Learning Classification Techniques for Breast Cancer Diagnosis	David A. Omondiagbe, Shanmugam Veeramani , Amandeep S. Sidhu,	Computer Aided Detection (CAD), Linear Discriminant analysis (LDA), CFS, RFE, PCA and LDA. SVM, ANN and NBC	The purpose of this paper is to combine techniques of machine learning algorithm with the methods of feature extraction/feature selection and to identify the most suitable approach compare their performances.	By this research obtained accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07%

Table no.01: Comparative Analysis

V. PROBLEM STATEMENT

Cancer the fatal diseases that has posed a danger to the industry. According to the Organization (WHO), cancer causes around 12.5 percent of all deaths worldwide, which is higher than the proportion of fatalities caused by HIV/AIDS, tuberculosis, and malaria combined. As a result, the growth in breast cancer strikes and mortality in Edo State raises an important question about the effect of breast cancer campaigns on Edo women, particularly in light of their poor response to early breast cancer presentation. The researcher critically evaluated the effectiveness of the system mastering system to easily classify which type of breast most cancers (Benign or Malignant) is present in light of the above and given the confirmation of the ACS (2013) that breast most cancers deaths remain preventable on an early level. Breast cancers will be split into three categories in the study: (Benign tumor and Malignant tumor). Manually dividing cancers into benign and malignant categories can be difficult, prone to human error, and time consuming. The following is a comprehensive list of the new system's features: Classification errors could be significantly reduced, early disorder analysis could be performed, human errors could be eliminated, and the device would no longer perish. The researcher, on the contrary, wants to use machine learning to detect and classify breast cancer.

VI. PROPOSED SYSTEM

A comparison of machine learning algorithm (ML) is shown in this suggested system: Support machine vector (SVM), Logistic Regression Algorithm, and Random Forest (RT). The Wisconsin datasets were utilized to generate the data collection. The dataset was categorized into training sets and testing sets in order to implement the machine learning methods. There will be a comparison of all six algorithms. The website will be given a model of the algorithm that produces the best results. The website will be built using the flask python framework. The database will be hosted on Xampp, Firebase, or the native Python and flask libraries. The UCI Machine Learning Repository has this data set available. It is form up of 32 multivariate real-world properties. The sum total of cases in this data collection is 569, and there are no missing values. The proposed system's procedure is as follows:

- The patient fixes an appointment using the website.
- The patient will next meet with the doctor in person for the appointment.
- The doctor will manually examine the patient before doing a breast mammography or an ultrasound.

## VIII. ALGORITHM

**Step 1:** Start

**Step 2:** Load the important libraries

**Step 3:** Import dataset and extract the X variables and Y

```
df = pd.read_csv("mydataset.csv")
```

```
X=df.loc[:,['Var_X1','Var_X2','Var_X3','Var_X4']]
```

```
Y = df[['Var_Y']]
```

**Step 4:** feature names as a list

**Step 4:** Y includes the labels and x includes features

**Step 5:** Function is used to convert text class to numerical class

**Step 6:** Divide the data into validation, test, &train.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3, random_state=123)
```

**Step 7:** Logistic Regression

```
lr_classifier=LogisticRegression(random_state = 51, penalty = 'l1')
```

```
lr_classifier.fit(X_train, y_train)
```

```
y_pred_lr = lr_classifier.predict(X_test)
```

```
accuracy_score(y_test, y_pred_lr)
```

**Step 8:** Support Vector Machine

```
svc_classifier = SVC()
```

```
svc_classifier.fit(X_train, y_train)
```

```
y_pred_scv =svc_classifier.predict(X_test)
```

```
accuracy_score(y_test, y_pred_scv)
```

**Step 9:** Random Forest

```
rf_classifier=RandomForestClassifier(n_estimators = 20,
```

```
criterion = 'entropy', random_state = 51)
```

```
rf_classifier.fit(X_train, y_train)
```

```
y_pred_rf = rf_classifier.predict(X_test)
```

```
accuracy_score(y_test, y_pred_rf)
```

**Step.10:** End

## IX. MATHEMATICAL MODEL

### Classification Methods

#### 1.Simple Logistic Regression Model

The LR classification model is a popular choice for modeling binary classifications. For this model, the conditional probability from one of the two output classes is assumed to be equal to a linear combination input feature. The logistic equation used for this classification model is:

$$Z_i = \ln\left(\frac{p_i}{1-p_i}\right)$$

where P indicates probability of occurrence of the event i.

#### 2. SVM Learning with Stochastic Gradient Descent (SGD) Optimization

Optimization is used SGD for support vector machine classification algorithm with hinge loss function. Gradient descent is well known optimization algorithms used in deep learning and machine learning algorithms. Stochastic gradient descent selects random samples from a dataset instead of selecting the batch data as in batch gradient descent. It computes gradient and performs weight updates

for each selected training sample  $x^i$  and the label  $output^i$  until a minimum cost ( $J_{\min}(w)$ ) is reached.

Randomly mix the training set's samples Until approx. cost minimum ( $J_{\min}(w)$ ) reached:

For each training sample i: Calculate gradients and update weights

For each weight j :

$$w_j = w + \Delta w_j$$

where :

$$\Delta w_j = \eta(\text{target}^i - \text{output}^i)x_j^i \quad \eta \text{ is learning rate}$$

#### Stage 1: Pre-processing

In this study to scale the characteristics using the module of Standard Scaler. Dataset standardization is typical prerequisite for different machine learning algorithms. It transforms the attributes Gaussian distribution to its standard based on

$$\frac{x_i - \text{mean}(x)}{\text{stdev}(x)}$$

where stdev is the standard deviation.

The Robust Scaler depends on the interquartile range to transform the features using,

$$\frac{x_i - Q1(x)}{Q3(x) - Q1(x)}$$

where Q1, Q2, and Q3 represent quartiles. All the transformations used are included in scikit-learn machine learning library.

#### Stage 2: Features Selection

Usually, feature selection is applied as a pre-processing step before the actual learning. However, no algorithm can give excellent predictions without informative and discriminative features, therefore, in this paper used randomized SVD to implement PCA and maintain the most important features while reducing the dataset size. The scikit-learn library in Python was used to create the feature selection module. Feature selection in our study was depend on the given modules: deleting low-variance features, univariate selection, and recursive elimination.

#### Stage 3: Machine Learning Algorithm

When compare to a single model, ensemble machine learning methods usually provide greater predictive performance. This was a machine learning competition, and the winning answer was used as a breast cancer diagnosis model. In this paper, the following heterogeneous ensembles machine learning algorithms were used to classify the given data set: Support Vector Machine (SVM), logistic regression (LR), Random Forest (RF).

## X. SYSTEM ARCHITECTURE



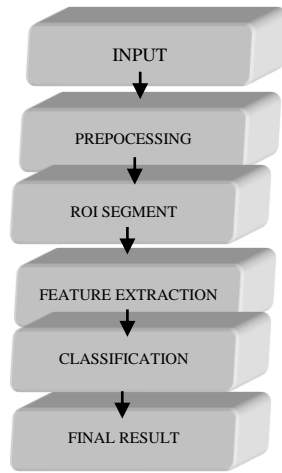


Fig.no.01: System Architecture

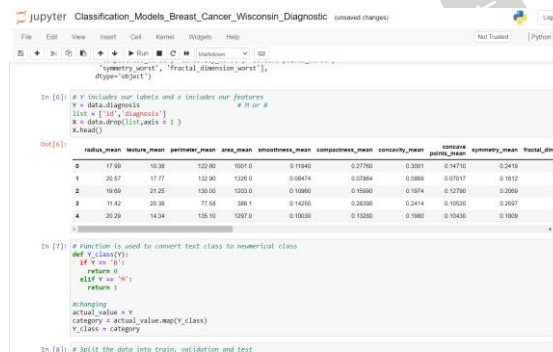
REFERENCE

**Description:** The first step is pre-processing here cleaned the data set that is removed any missing value tests and outliers and test first independent variables which was ID or sample code number was removed without making use of feature selection at first Fig represents the system architecture of breast cancer detection. Framework design is the calculated model that characterizes the structure, conduct, and more perspectives on a framework. Framework engineering can be made up of framework segments and subframeworks that work together to complete the overall framework.

XI. ADVANTAGES

- It increases accuracy and efficiency
- It is effective and faster.
- It reduces overfitting.
- It minimises the complexity of the model and helps to understand.

XII. DESIGN DETAILS



```

In [6]: # Y includes our labels and x includes our features
Y = data_diagnosis # @ or @
list = ['id', 'diagnosis']
x = data.drop(list,axis = 1)
X = x.drop('id',axis = 1)
dtype='object')

Out[6]:
radius_mean  texture_mean  perimeter_mean  area_mean  smoothness_mean  compactness_mean  concavity_mean  convexity_mean  symmetry_mean  fractal_dim
0      17.99      10.38      122.80      1001.0      0.11040      0.27700      0.3051      0.24710      0.2419
1      20.57      17.77      132.90      1326.0      0.08474      0.07884      0.0880      0.07617      0.1812
2      19.69      21.25      133.00      1203.0      0.10960      0.15900      0.1974      0.12790      0.2069
3      11.42      20.39      77.58      386.1      0.14250      0.28390      0.2414      0.10520      0.2967
4      20.29      14.34      136.10      1287.0      0.10030      0.13290      0.1980      0.15420      0.1989

In [7]: # function is used to convert text class to numerical class
def Y_class(Y):
    if Y == 'M':
        return 0
    elif Y == 'B':
        return 1
    else:
        return -1

# changing
actual_value = Y
category = actual_value.map(Y_class)
Y_class = category

In [8]: # Split the data into train, validation and test
  
```

Fig.no.02: Data Description

XIII. CONCLUSION

Thus, we have attempt to implement the paper “Anjana Ivaturi, Ankita Singh, B. Gunanvitha, Chethan K S”, “Soft Classification Techniques for Breast Cancer Detection and Classification”, ICIEM 2020. According to this paper machine learning classifiers such as Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM). RF which give the accuracies 98%. Hence, RF Classifier Performing good on the train and test data. So RF Model could take for further prediction on unseen data.

[1] “Anjana Ivaturi, Ankita Singh, B. Gunanvitha, Chethan K S”, “Soft Classification Techniques for Breast Cancer Detection and Classification”, ICIEM 2020.

[2] J. B. Harford, "Bosom malignancy early identification in low-pay and center salary nations: do what you can versus one size fits all," Lancet Oncology, vol. 12, no. 3, pp. 306–312, 2011.

[3] C. Lerman, M. Daly, C. Sands, A. Balshem, E. Lustbader, T. Heggan, L. Goldstein, J. James, and P. Engstrom, "Mammography adherence and mental trouble among ladies in danger for bosom disease," Journal of the National Cancer Institute, vol. 85, no. 13, pp. 1074–1080, 1993.

[4] P. T. Huynh, A. M. Jarolimek, and S. Daye, "The bogus negative mammogram," Radiographics, vol. 18, no. 5, pp. 1137–54, 1998.

[5] N. Saidin, U. K. Ngah, H. Sakim, and N. S. Ding, Density based bosom division for mammograms utilizing diagram cut and seed-based area developing methods. IEEE Computer Society, 2010.

[6] R. D’mello, V. Pardeshi, H. Mehta and S. Dhage, "Comparative Study of Breast Cancer Detection Techniques," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-6.

[7] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2018, pp. 997-1002.

[8] P. Tumuluru, C. P. Lakshmi, T. Sahaja and R. Prazna, "A Review of Machine Learning Techniques for Breast Cancer Diagnosis in Medical Applications," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 618-623.

[9] M. Amrane, S. Oukid, I. Gagaoua and T. Ensar, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4.

[10] Mohamed A. Berbar, Yaser. A. Reyad, Mohamed Hussain, “Breast Mass Classification using Statistical and Local Binary Pattern Features” 2012 16th International Conference on Information Visualisation.