How to Cite:
Das, L. N., Saini, S., Kataria, P., & Dipanshu. (2022). Breast cancer detection from histopathological images using machine learning models. *International Journal of Health Sciences*, *6*(S3), 9542–9553. https://doi.org/10.53730/ijhs.v6nS3.8254

Breast cancer detection from histopathological images using machine learning models

Prof. L. N. Das
Department of Applied Mathematics, Delhi Technological University
Email: lndas@dce.ac.in

Sachin Saini
Department of Applied Mathematics, Delhi Technological University
Email: sachinsaini_2k18mc098@dtu.ac.in

Puneet Kataria
Department of Applied Mathematics, Delhi Technological University
Email: puneetkataria_2k18mc084@dtu.ac.in

Dipanshu
Department of Applied Mathematics, Delhi Technological University
*Corresponding author email: dipanshukumar_2k18mc034@dtu.ac.in

***Abstract*---**Breast cancer is the most common cancer in women worldwide, accounting for more than 25% of all cancer cases and affecting more than 2.1 million people each year. According to the WHO, early detection is crucial to improving patient outcomes and survival. However, prognosis by histology of biopsy tissue is a complex procedure and the ultimate interpretation can be controversial. Therefore, machine learning algorithms are deployed to generate techniques that can be used by technicians, radiologists, and physicians as tools to unequivocally detect and diagnose breast cancer at an early stage. This will help to significantly increase the survival rate of the patients and their subsequent quality of life.

***Keywords*---**breast cancer, histopathology, tumors, machine learning techniques.

Introduction

Breast cancer is considered one of the leading causes of death in women. However, the deplorable trend of this rapid increase in the number of breast cancer cases worldwide to also provides a wealth of data that is of great use in advancing clinical and medical research aimed at early detection and diagnosis of

the disease. It is therefore a very valuable research topic. Previous studies have recognized this importance and have proposed the use of machine learning and deep learning algorithms to classify suspicious neoplasms using data from previous patients as training modules, albeit largely on an academic scale. Therefore, our goal is to advance this research by not only creating tools that can help to better classify tumor growths as cancerous or non-cancerous, but also to integrate the trained models into the hic et nunc through our applications for use in scenarios integrate today's patients and physicians a reality. Screening for breast cancer involves several tests, a few of which are preliminary in nature and can provide an indication of whether or not more stringent testing is needed. The time lag between receiving the test results and seeing a specialist can be a concern for the patient. The initial problem in this project was the detection of breast cancer with high precision based on functional parameters of the breast tissue cells. The problem with the functional parameter dataset was the expertise and time required to detect and quantify cellular parameters, requiring a system to automatically detect parameters from histological images. The use of this screening procedure should not be limited to clinicians and laboratory experts only, but the facility should be made available to patients via a portal where laboratory results can be uploaded and patients can receive the results without the intervention of a specialist.

The Datasets

The project used two different datasets, one for each of the two phases performed during its course. The first phase, related to the determination of the presence of carcinomas in a patient, used the widely available and tested Wisconsin breast cancer diagnostic record. In the second phase, the breast histology imaging dataset was used as a training and validation input for the development of his convolutional neural network.

Wisconsin Diagnostic Breast Cancer dataset

The dataset used right here is publicly available for all. To create the data set, Dr. Wolberg used fluid samples from sufferers with solid breast tumors and an easy-to-use graphical computer application known as Xcyt, which is in a position to investigate cytological functions primarily based totally on a virtual scan. The application makes use of a curve-fitting set of rules to calculate ten functions for every of the instance cells, then calculates the mean, extreme, and standard error values for every function withinside the picture and returns a vector with an actual price of thirty.

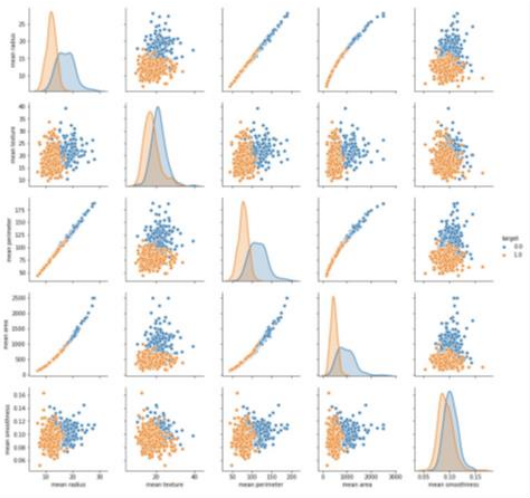

Fig 2.1. Concurrent mapping of mean values of radius, texture, perimeter, area, and smoothness showing relation between cancerous and non-cancerous tumors features

Breast Histology Images dataset

This set of data consists of histological breast images selected from Andrew Janowczyk's website. Our objective here is to classify the cancer photos (IDC: invasive ductal carcinoma) from non-IDC photos. The original set of data has 277,524 50x50 pixel RGB digital image spots (198,738 negative IDCs and 78,786 positive IDCs) derived from 162, hand-stained histopathological breast specimens. These photos are small blobs derived from virtual photos of breast tissue samples. Breast tissue carries large number of cells, however, just a few of them are cancerous. Patches marked "1" incorporate the cell function of invasive ductal carcinoma.

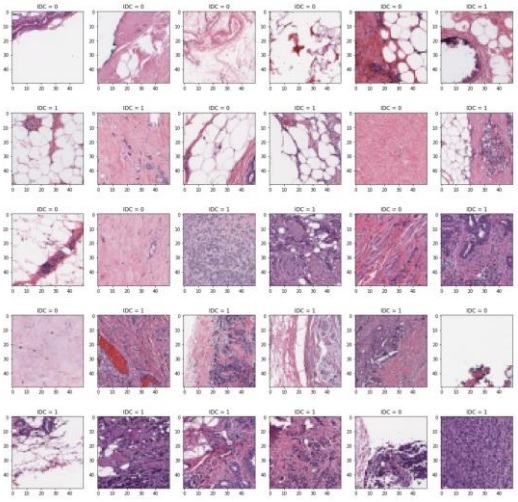

Fig 2.2. Histology images from the dataset; 0 signifies a non-cancerous tissue patch and 1 signifies a cancerous tissue patch

Machine Learning Techniques
K-Nearest neighbours (KNN)

K-Nearest Neighbors is a simple but crucial category algorithm in Machine Learning. It belongs to the supervised learning area and unearths intense applications in pattern recognition, statistics mining, and intrusion detection. It is extensively applied in real-lifestyles eventualities when you consider that it's far non-parametric, i.e., does not make any underlying assumptions approximately the distribution of statistics. We used the KNN set of rules at the preprocessed WDBC dataset, varying the cost of K from one to 10 if you want to determine the set of rules' performance over the complete range. Maximum accuracy becomes received on the cost of K = 7 for our dataset. The KNN set of rules has a computationally pricey trying out section and because of our constrained time and processing power, we determined to execute dimensionality discount at the statistics previous to reapplying KNN. This made the gap metric greater meaningful. As KNN is a higher ideal for datasets with a smaller variety of features, we created a correlation matrix to useful resources in characteristic extraction to perform the dimensionality discount.

Dimensionality Reduction

Reducing the range of impartial variables to a fixed of important variables via means of removing the variables which are much less substantial in predicting the final results is referred to as Dimensionality reduction. It is used to get much less dimensional statistics in order that higher visualization of the machine learning models may be performed via way of means of plotting the prediction areas and the prediction limit for every model. Regardless of the range of impartial variables, we frequently gain impartial variables via the means of making use of the appropriate dimensionality reducing technique. The two commonly used methods in the downsizing process are feature selection and feature extraction. Feature choice is used to filter out the redundant capabilities of your dataset. The principal distinction between feature selection and feature extraction is that feature selection preserves a subset of the unique capabilities, whilst feature extraction creates totally new ones. Feature extraction on the other hand is used to create a new, smaller set of capabilities that captures the maximum beneficial information. Once more, feature selection preserves a subset of the unique capabilities, whilst feature extraction creates new ones.

Support vector machine

Support Vector Machine is a widespread supervised Learning algorithm, that is employed for Classification and Regression issues, and given the labeled coaching knowledge (supervised learning), the algorithmic program generates the best hyperplane that classifies new examples. In 2-dimensional space, this hyperplane is defined as a line which divides a plane into two parts, with every class lying on either aspect of the line. Support vector machine is a brand-new technique to supervised pattern classification and is implemented with success in a number of problems with pattern recognition. It is an educational set of rules for getting to know classification and regression regulations from data. In the SVM technique, we aim to increase the margin between the information points as much as

possible and therefore the hyperplane. The function of loss that helps maximize the margin is known as hinge loss.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

Hinge Loss Function, simplified

The cost turns out to be 0 given that the predicted value and the actual value have the same sign. If that does not turn out to be the case, we will find out the value of the loss. We additionally add a regularization parameter for the cost function. The intention of the parameter is to stabilize the margin loss and maximization.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Loss Function for SVM

And now we can find out the gradient and then we can update the weights.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \left(1 - y_i \langle x_i, w \rangle\right)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Gradients

When there's no misclassification, we can just update the gradient by using the regularization parameter.

$$w = w - \alpha \cdot (2\lambda w)$$

Gradient Update - no misclassification

If our model makes an error within the prediction of the category of our data point, we perform a gradient update.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Gradient Update - misclassification

Deep Learning Techniques
Convolution Neural Network (CNN)

Deep learning CNN models use pictures to coach and test themselves. Every input picture passes by a series of convolutional layers with filter pooling, totally connected layers, and so on. The Softmax function is then applied using the model obtained to classify an object with probabilistic values between zero & one. A CNN might have many layers, each of which learns to acknowledge totally different options of an image. Filters at different resolutions are applied to every coaching image, and therefore the output of every convolved image is employed as input for succeeding layer. Filters will begin as very easy options, such as edges and brightness, and grow in complexness to features that unambiguously outline the object. Using CNNs for deep learning has become progressively well-liked thanks to 3 necessary factors:

- It terms out the requirement for feature extraction (manually) - the options are directly learned by the CNN.
- It produces progressive recognition results.
- It is retrained for tasks of new recognition, which allows you to make on networks which already exists.

A CNN is well trained on lots and lots of pictures. Once operating with giant amounts of information and sophisticated network architectures, GPUs will considerably speed up the time interval to coach a model.

Neural Network Architecture

The network is designed as to give associate accommodative model and design to train our neurons. The structure includes five key alternatives which can be defined below. The most effective unit of a neuron is called a node which acts as an affiliation among passing information.

- Input Layer: In this layer every input features are processed and also passed to other layers.
- Hidden Layer: The calculation of weighted information which comes from the input layer takes place in this layer and then it also transmits that data to the next hidden layer or output layer.
- Output layer: Finally, in this layer an activation function is used to Map the output to format that we desire.
- Connections and weights: Connection helps with the transfer of neuron p to the input of neuron k and serves as links between different nodes.
- Activation function: This function helps in mapping the provided input and the output. It is seen as signal of 0 or 1 which depends upon threshold function output of activation or non-activation of neuron. We use nonlinear activation functions to compute nontrivial bounded-node problems.

Results and Discussion

Preprocessing the Data

The measurements obtained are from digitized fine-needle aspirate (FNA) images of a mass of breast suspected of harboring cancerous tissue. The dataset includes 569 cancer biopsy cases; Each record has 32 attributes. Characteristics included in the data are identification number, cancer diagnosis (coded as “M” for malignant or “B” for benign), and 30 laboratory measurements with numerical values. Unknown and zero values were removed to avoid biasing results. The normalization of the data forms to a numeric type has also been done.

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.1444
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374

Fig 5.1. Data sample obtained after preprocessing

KNN Before Dimensionality Reduction

The algorithm used corresponds to KNN with a worth of k between one to 10 (k increasing at a step rate of one), that by default uses euclidian metric that operates with real (double) numbers.

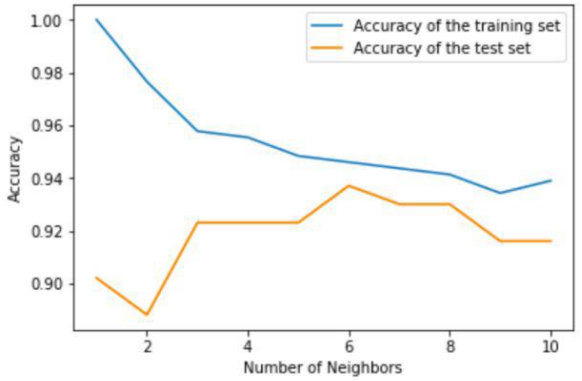

Fig 5.2. Accuracy trend with an increase in the number of neighbors

Dimensionality Reduction - Feature Extraction

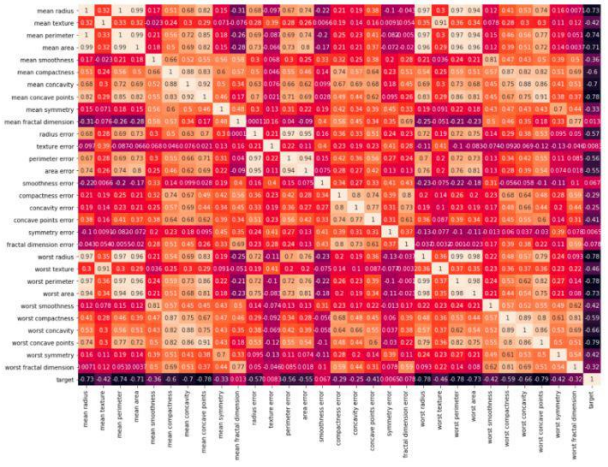

Fig 5.3. Correlation Matrix used for Feature Extraction

K-Nearest Neighbour Post Dimensionality Reduction

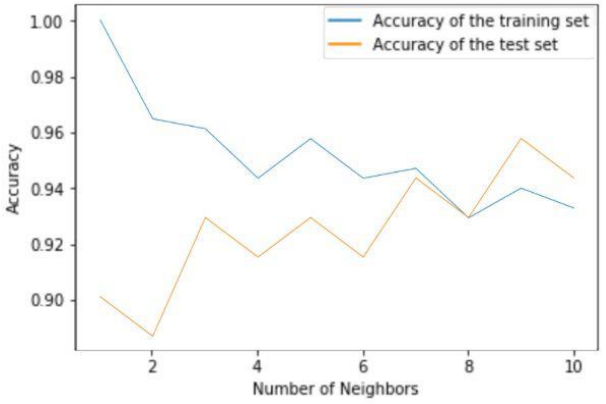

Fig 5.4. Trend of accuracy with increase in number of neighbors post dimensionality reduction

Support Vector Machine

```
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_breast_cancer

cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(cancer.data, cancer.target, random_state=0)

svm = SVC()
svm.fit(X_train, y_train)

print('The accuracy on the training subset: {:.3f}'.format(svm.score(X_train, y_train)))
print('The accuracy on the test subset: {:.3f}'.format(svm.score(X_test, y_test)))
```

```
The accuracy on the training subset: 1.000
The accuracy on the test subset: 0.629
```

Fig 5.5. Accuracy of SVM before feature scaling

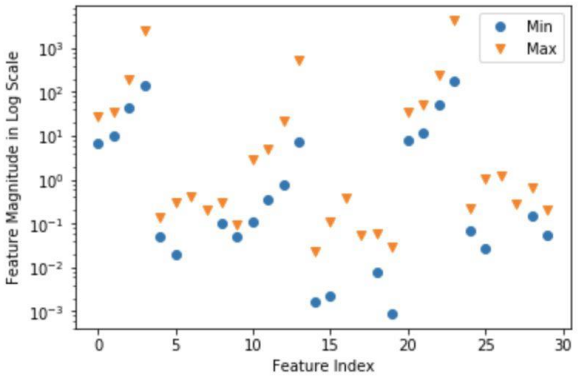

Fig 5.6. Map of min max values of features

```
min_train = X_train.min(axis=0)
range_train = (X_train - min_train).max(axis=0)

X_train_scaled = (X_train - min_train)/range_train
```

```
The accuracy on the test subset: 0.951
```

Fig 5.7. Accuracy of SVM model post feature scaling

Feature scaling was thus successfully used to substantially increase the accuracy of the trained SVM to 95.1%

Summary of Results obtained from Sklearn Base Models

The results obtained show that although the highest accuracy for the trained data is obtained from the logistic regression model, the random forest model performs best on validation data with an accuracy of 75.5%. These results also show that each of these models has a higher accuracy rate on the training data than on the validation data, as expected.

```
logreg : averaged train/valid accuracy = 0.859/0.682
random_forest : averaged train/valid accuracy = 0.786/0.755
extra_trees : averaged train/valid accuracy = 0.773/0.754
gaussianNB : averaged train/valid accuracy = 0.715/0.715
```

Fig 5.8. Training and Validation Accuracy for Sklearn Models built

Confusion Matrix for the SVM Model

The confusion matrix for the SVM model provides the accuracy of the model used. The results show 27% false-negative and 20.8% false-positive and 72.9% and 79.15% true positive and true negative respectively.

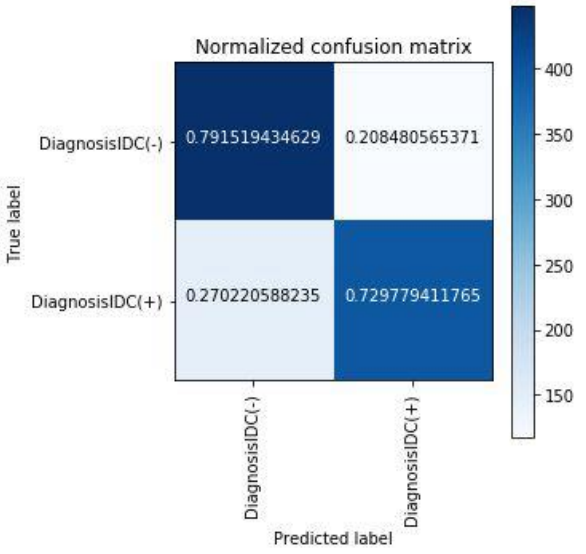

Fig 5.9. Confusion matrix for the SVM model

Confusion Matrix for the CNN

The confusion matrix for the CNN model provides the accuracy of the model used. The results show 15% false-negative and 22% false-positive and 85% and 78% true positive and true negative respectively, which is a major improvement over the SVM model.

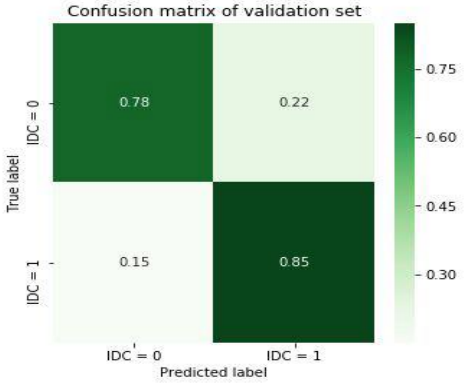

Fig 5.10. Confusion matrix for CNN model

Accuracy Metrics for CNN model on histology image dataset

```
/kaggle/working/nn0.meta
INFO:tensorflow:Restoring parameters from nn0
final train/valid loss = 0.4411/0.4291, train/valid accuracy = 0.8018/0.8126
```

Fig 5.11. Accuracy prediction for CNN model

The accuracy of a machine learning model may be calculated by the application of

the subsequent formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Thus, the accuracy obtained can be confirmed as follows:
Accuracy = (0.78 + 0.85) / (0.78 + 0.85 + 0.22 + 0.15)
= 0.815
~= 0.8126

Graphical representation of Accuracy and Loss Metrics

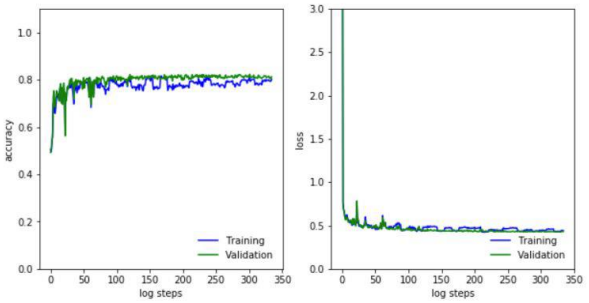

Fig 5.12. Accuracy and Loss curves for CNN model

From the graphs and analysis:

- ● By the use of 10-fold cross-validation, the best base models can gain an accuracy of approximately 76% at the validation sets.
- ● Implementing a 90% cut-up for coaching and validation understanding and educating the neural network for thirty epochs while the usage of data augmentation we're capable of reaping an accuracy of 80% at the validation tests.

Conclusion

In this work, various histological images were carefully examined and then used to try to predict whether a particular patient's neoplasm was an occurrence of breast cancer or not, that is, whether a detected tumor was benign or malignant. Various features such as cell shape, cell size, etc. were extracted. careful examination of images of tissue taken from the patient by fine-needle aspiration. If the tumor cells were found to be normal, the tumor was classified as benign. If, on the other hand, a cell expansion was found that was reminiscent of cancerous growths, the tumor was classified as malignant. The KNN algorithm was originally used to classify tumors using numerical values of certain features of cell nuclei obtained from histological FNA images. Dimensionality reduction was then performed to reduce the number of features (number of dimensions) in the dataset to try to improve the observed performance of the algorithm.

Dimensionality reduction was achieved by feature extraction, which in turn was performed by constructing a correlation matrix of the thirty real-valued features available in the original Wisconsin breast cancer diagnostic dataset. So this matrix defines the correlation of each feature with all other features in the dataset and also with itself the special case that gives a value of one. The 30 real-valued attributes available in the data set were successfully reduced to 12 features; Therefore, the number of dimensions has been reduced to less than half of their value before dimension reduction. This operation improved the capacity of the algorithm, although not to an extent commensurate with the effort and time required to perform feature extraction in the course of dimensionality reduction. Therefore, a different machine learning model was sought in the following. Because it was desirable to preserve the properties of the dataset and increase the overall efficiency of our program, we used SVM, which uses regularization parameters to avoid overfitting, and kernel tricks to generate expert knowledge. Because our dataset had traits that varied widely in size and range, the trait scale was implemented to weight all traits equally; otherwise, large-scale features would have dominated in model training.

After achieving an acceptable level of accuracy, consideration was given to converting the histology images to a digital format and then directly performing the recognition and classification processes using the digitized images. The goal was to significantly reduce the amount of work associated with the previous approach and improve the efficiency of the layering process. Some basic sci-kit learn models like Logistic Regression, Random Forest, Extra Trees, Gradient Boosting, Decision Tree, GaussianNB, etc. were first applied to image processing. The highest average accuracy achieved by them was a meager 74%. A convolutional neural network was then built using TensorFlow, yielding a validation accuracy of out of 80 for the same data set.

References

1. Leo Breiman. 2001. Random forests. Machine learning 45(1):5–32.
2. Yali Amit and Donald Geman. 1997. Shape quantization and recognition with randomized trees. Neural computation 9(7):1545–1588
3. Borges et al. 2014. Benign and Malignant breast tumor classification based on region growing and CNN segmentation.
4. Lulu Wang 1. 2017. Early diagnosis of breast cancer.
5. Galotkar et al. 2017. New diagnostic tools for Breast Cancer.
6. Cakir and Demirel. 2011. A Software Tool for determination of Breast Cancer Treatment Methods.