BREAST CANCER CLASSIFICATION AND PREDICTION USING HISTOPATHOLOGY IMAGES

SRI LAKSHMI. H
2022510053

ABSTRACT

Breast Cancer is ranked the number one cancer among Indian females with a rate of 25.8 per 100,000 women and a mortality rate of 12.7 per 100,000 women. It is also one of the most common causes of cancer worldwide. There have been many biological and non-biological research in the past and present to be able to prematurely detect breast cancer. In this project, I have taken an approach using Deep Learning Neural Network CNN to try to predict whether a histopathology image of the human tissue is cancerous or non-cancerous, and if cancerous, what stage does it belong to: in-situ, Invasive Ductal Carcinoma Stage 1 (IDC 1), IDC 2, IDC 3 and IDC 4. The results show higher accuracy than the most state of the art methods in this field.

1.INTRODUCTION

The aim for this Project is:
1)To create an Image Classification Model that can accurately identify and categorize breast cancer cells from the non-cancerous images.
2)Implement deep learning neural network Convolutional Neural Networks (CNN) to build the Image Classification Model.
3)To Predict the stages of the identified cancerous cells.
4)Analyze and Optimize the model and evaluate with performance metrics.

Histopathology refers to the microscopic examination of tissue in order to study the manifestations of disease. Specifically, in clinical medicine, Histopathology refers to the examination of a biopsy or surgical specimen by a pathologist, after the specimen has been processed and histological sections have been placed onto glass slides. Histopathology slides, on the other hand, provide a more comprehensive view of disease and its effect on tissues, since the preparation process preserves the underlying tissue architecture. As such, some disease characteristics, e.g., lymphocytic infiltration of cancer, may be deduced only from a histopathology image. The diagnosis derived from a histopathology image remains the gold standard in diagnosing a substantial number of diseases including almost all types of cancer.

Due to the anatomy of the human body, women are more vulnerable to cancer than men. Among the different reasons for breast cancer, age, family history, breast density, obesity, and alcohol intake are reasons for breast cancer.

Statistics reveal that in the recent past the situation has become worse. As a case study, shows that breast cancer accounts for 14% of cancers in Indian women. It is reported that with every four minutes, an Indian woman is diagnosed with breast cancer. Breast cancer is on the rise, both in rural and urban India. A 2018 report of Breast Cancer statistics recorded 1,62,468 new registered cases and 87,090 reported deaths. In 2007, the number of new cases for breast cancer was 12775, while the expected number of new cancer patients in 2018 will be 18235. Statistics show that, in the last decade, the number of new cancer disease patients increased every year at an alarming rate.

Breast cancer tumors can be categorized into two broad scenarios.

(i) Benign (Noncancerous)
(ii) (ii) Malignant (Cancerous).

Identification of the benign and malignant tissues is a very important step for further treatment of cancer. Based on the penetration of the skin and damage of the tissue medical photography techniques can be classified into two groups.

(i) Noninvasive.
(ii) Invasive.

The benefits of this project are; It helps in improvement in early diagnosis. It helps in reducing healthcare costs, and a more beneficial influence on worldwide breast cancer management is all possible because of the abilities of the machine learning techniques in breast cancer diagnosis.

2.EXISTING WORKS

Breast Cancer Detection and Prevention Using Machine Learning [1] **by**
1 Faculty of Computing, Islamia University of Bahawalpur, Bahawalpur 63100, Punjab, Pakistan
2 Department of Information Systems, College of Computer and Information Science, King Saud University, Riyadh 11543, Saudi Arabia
3 Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
4 Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

In their research, the scholars proposed an efficient deep learning model that is capable of recognizing breast cancer in computerized mammograms of varying densities. Their research relied on three distinct modules for feature selection: the removal of low-variance features, univariate feature selection, and recursive feature elimination. The craniocaudally and medial-lateral views of mammograms are incorporated. They tested it with a large dataset of 3002 merged pictures gathered from 1501 individuals who had digital mammography performed between February 2007 and May 2015. The researchers applied six different categorization models for the diagnosis of breast cancer, including the random forest (RF), decision tree (DT), k-nearest neighbors (KNN), logistic regression (LR), support vector classifier (SVC), and linear support vector classifier (linear SVC).

Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms [2] **by**
1 Habib Dhahri

This study was based on genetic programming and machine learning algorithms that aim to construct a system to accurately differentiate between benign and malignant breast tumors. The aim of the study was to optimize the learning algorithm. In this context, they have applied the genetic programming technique to select the best features and perfect parameter values of the machine learning classifiers. The performance of the proposed method was based on sensitivity, specificity, precision, accuracy, and the roc curves.

Breast Cancer Survival Prediction by
Sathipati and Ho used an optimized SVM regression to identify miRNA signatures associated with survival time in patients with lung adenocarcinoma. They used a novel feature selection algorithm called IBCGA and these features were then fed into traditional SVR. Although their custom SVR outperformed other regression methods, it did not generalize well to unseen validation data. Another issue with this paper was the size of datasets.

3.DATASET

Breast cancer can develop at any different part of the breast. The most common form of breast cancer is Invasive Ductal Carcinoma (IDC). In order to detect IDC, it is through various methods such as mammography, ultrasound, biopsy and so on. Through biopsy, histopathology images are derived. The Dataset that is going to be used for training and testing for the image classification model will be the Breast Histopathology Images dataset. Since the dataset is too large, I have taken 1/8 th part, i. e. 46253 images.

The dataset was originally uploaded on the Gleason Case website:

http://gleason.case.edu/webdata/jpi-dl tutorial/IDC_regular_ps50_idx5.zip

The dataset consists of 277,524 50x50 pixel RGB digital image patches that were derived from 162 H&E-stained breast histopathology samples. Within these patches, there are 198, 738 IDC negative and 78,786 IDC positive. These images are small patches that were extracted from digital images of breast tissue samples. The breast tissue contains many cells but only some of them are cancerous. Patches that are labeled "1" contain cells that are characteristic of invasive ductal carcinoma. For more information about the data, see

https://www.ncbi.nlm.nih.gov/pubmed/27563488

and
http://spie.org/Publications/Proceedings/Paper/10.1117/12.2043872.

There is a compilation of all the images into one folder which is named IDC_regular_ps50_idx5. For this project, the data from IDC_regular_ps50_idx5 folder will be directly extracted. It is a folder full of folders that is named after the patients' id, which also consists of the IDC positive and IDC negative photos. There are 279 patients, and in each file contains images of IDC positive and negative. The way the images are stored is in two different arrays, where one is to store the images, the other is to store its type of class, IDC positive or negative, indicated with the numbers 0 and 1.

4.MY WORK

Firstly, the dataset is imported entirely in my Jupyter notebook. Then the patches are visualized and some observations are drawn out.

Data Visualization:

What does the image look like? And to what ratio is the IDC positive and IDC negative?

A single image from the dataset is visualized:

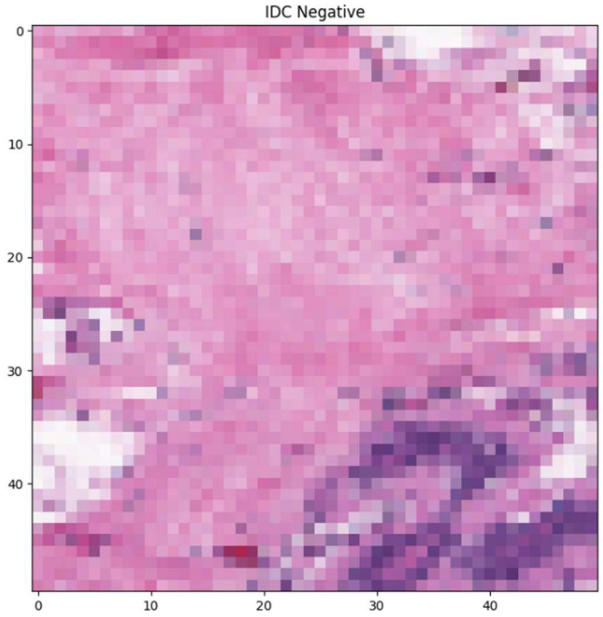


Fig. 1: A sample of IDC Negative sample i. e. Invasive

Healthy and Non healthy patches are visualized then,

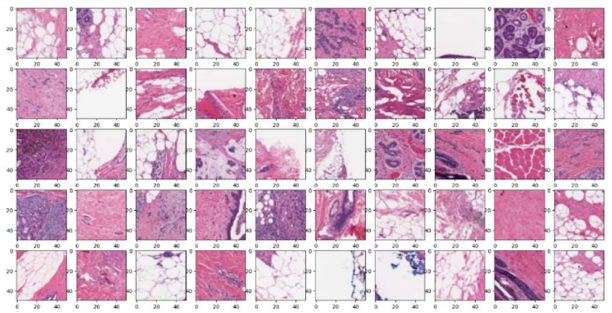


Fig. 2: Healthy Patches

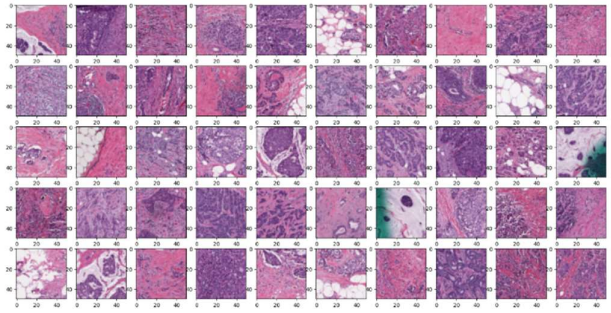


Fig. 3: Cancerous Patches

Observations:

- There might be a chance that not all the images are 50x50pixels.
- Comparing the Healthy Patches and the Cancer Patches, the Cancer patches seems to have more Purple-ish look to it.

To avoid any issues during training, I have resized all images to follow the 50x50 size to ensure fairness. After resizing, normalization, shuffling and splitting the data have been carried out.

The Model used for this project is a custom Convolutional Neural Network model. The Optimizer used for this model is Adam and the evaluation metrics is Accuracy and Confusion Matrix.

Application or Use of CNN in Breast Cancer Diagnosis:

CNNs and AI can improve medical image quality by enhancing low-contrast features, reducing noise, removing artifacts, and optimizing image registration. They also assist in image, segmentation, and ROI detection, enabling precise analysis and diagnosis of anatomical structures or lesions. AI algorithms can adjust image contrast, brightness, and intensity levels and apply contrast-limited adaptive histogram equalization (CLAHE) techniques to improve image quality. Additionally, CNNs recognize and remove common imaging artifacts, ensuring accurate interpretation. AI algorithms optimize image alignment, while segmentation and ROI detection enable precise analysis and diagnosis of specific areas. CNNs is also used for super-resolution imaging, enhancing image resolution and quality beyond the original acquisition. AI-driven super-resolution techniques use Deep learning models to generate high-resolution images from low-resolution inputs, providing enhanced detail and diagnostic
information.

CLASSIFICATION OF CANCEROUS AND NON CANCEROUS IMAGES

Custom CNN Model for classifying between Cancerous and Noncancerous:

My Custom CNN has 4 Convolution and Max Pooling layers, 1 Flattening Layer and 2 Dense Layers. The Conv2D layers apply convolutional operations with different filter sizes to capture image features. MaxPooling2D layers down sample the spatial dimensions to reduce computational complexity. The flattening layer flattens the output from the previous layer into a 1D array. The Dense layers are responsible for combining the extracted features and making the final classification. The first layer uses Relu Activation function and the last layer uses the Softmax activation function to produce probability scores for each class, cancerous and non cancerous.

The number of Epochs was 25 with a batch size of 75. The loss function used here is Binary Cross Entropy.

The Model has an Accuracy of 85 percent.
Accuracy: 0.85
Precision: 0.73
Recall: 0.66
F1-score: 0.70

The Model Accuracy Curve and the Loss curve has been plotted in addition to the Confusion Matrix.

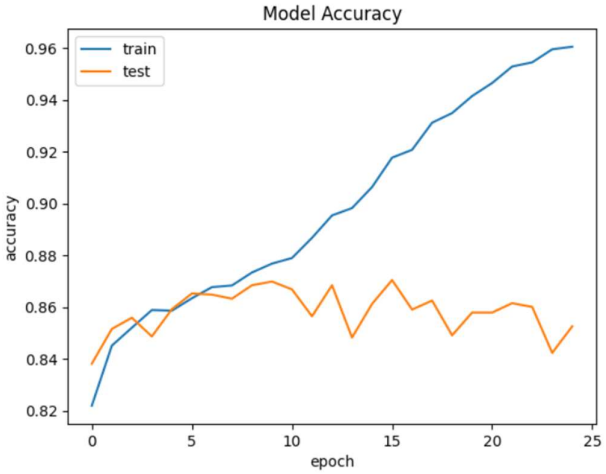


Fig. 4: Accuracy Curve for Model 1

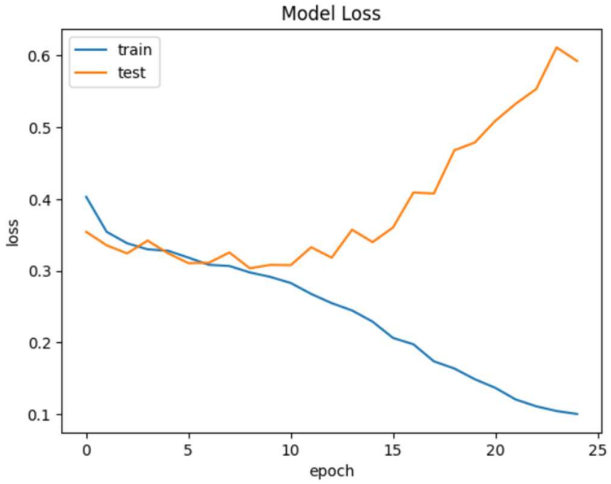


Fig. 5: Loss Curve for Model 1

To further increase the accuracy of the model, another dense layer was added and early stopping was used. The Model 2 had 4 Convolutional Layers, 1 Flattening layer and 3 Dense Layers. The number of Epochs was 25 with a batch size of 256. The loss function used here is Binary Cross Entropy.

The improved Model's Accuracy has been increased from 85 to 87 percent.
Accuracy: 0.87 Precision: 0.76
Recall: 0.70 F1-score: 0.73

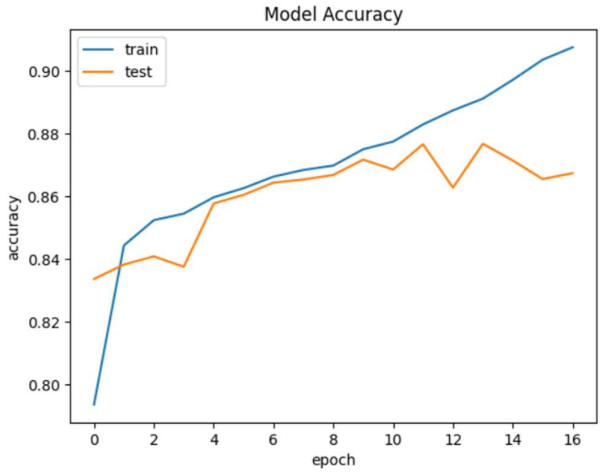


Fig. 6: Accuracy Curve for Model 2

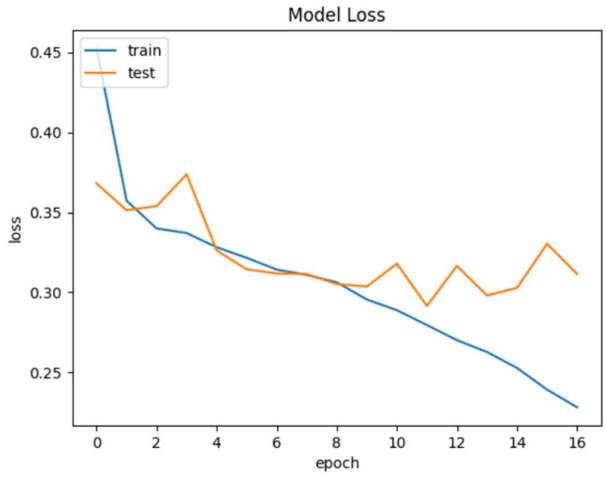
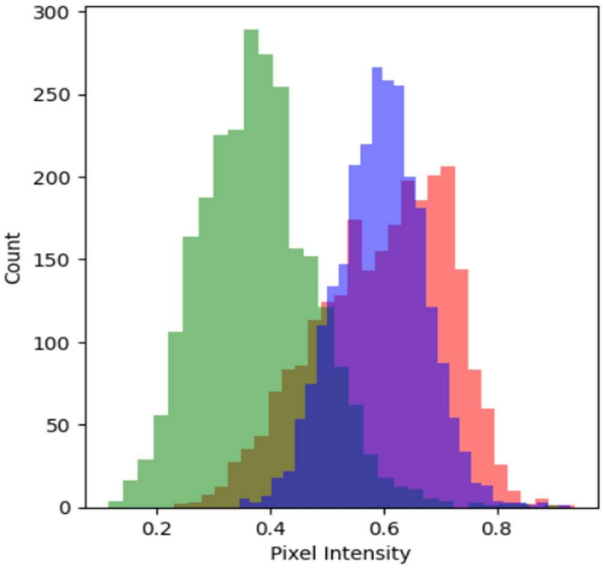


Fig. 7: Loss Curve for Model 2

CLASSIFICATION OF STAGES

For the Classification of stages, the RGB Intensity of the patches were analyzed and based on the intensity of the Red and Blue channels, which constitutes the intensities for the purple-ish color of the patches.



For Cancerous:
Average Color (BGR):
[169.7552 130.784 187.0252]
RGB Ratios:
[0.66570667 0.51287843 0.73343216]

Support Vector Machine for the classification of stages

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. The SVM Model was used initially for classifying the stages. The Cancerous Data which had 11,708 images were split in the ratio of 80% for training and 20% for testing.
The accuracy of this model initially was 19% and upon improvement, the accuracy was increased to 0.203245089666951which is 20% percent.

Linear Regression Model

The Linear Regression Model was trained and tested with the same training and testing sets as of the above SVM Model. The accuracy was very low and with Mean absolute Error of 1.6497409805180474.

Custom CNN Model

My Custom CNN has 4 Convolution and Max Pooling layers, 1 Flattening Layer and 3 Dense Layers. The number of Epochs was 2 with a batch size of 32. The loss function used here is Sparse categorical cross entropy.
The Model had an accuracy of about 19 percent.
Then to further Generalize and improve the accuracy of the model, Data augmentation technique was used and the model was trained again.

Data augmentation is a technique commonly used in computer vision tasks to artificially increase the size of the training dataset by applying various transformations to the existing images. This helps improve the generalization of the model and reduces the risk of overfitting.

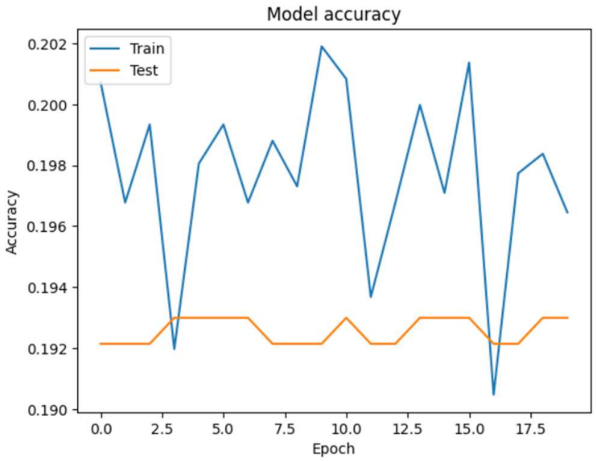


Fig. 8: Accuracy Curve for this Model

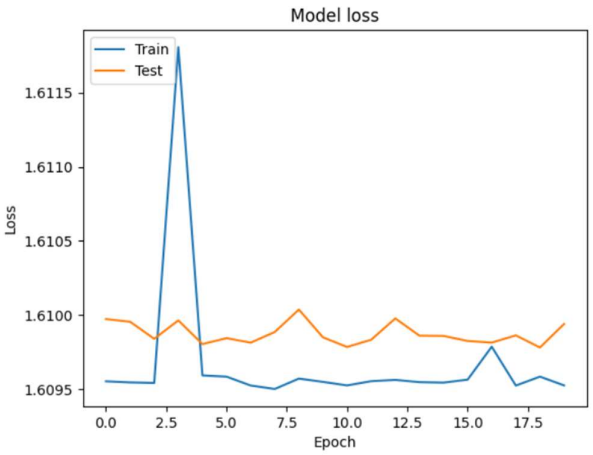


Fig. 9: Loss Curve for this Model

Accuracy: 0.19
Class: In-Situ,
Precision: 0.00, Recall: 0.00, F1-score: 0.00
Class: idc 1,
Precision: 0.19, Recall: 1.00, F1-score: 0.32
Class: idc 2,
Precision: 0.00, Recall: 0.00, F1-score: 0.00

Class: idc 3,
Precision: 0.00, Recall: 0.00, F1-score: 0.00
Class: idc 4,
Precision: 0.00, Recall: 0.00, F1-score: 0.00

Lastly, the model was tested with an image and the predicted class was obtained and the image was visualized.

```
1/1 [==============================] - 0s 53ms/step
Predicted Label: idc 1
```

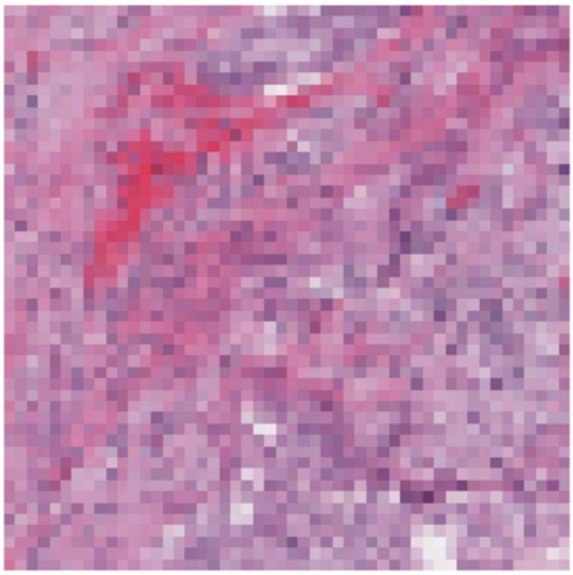


Fig. 10: A Cancerous patch predicted as IDC Stage 1

5.FUTURE WORK

1)I aim to try various machine learning techniques and a standardized dataset to further find out the stage the patch is in and predict the duration it will take to reach the next stage.

6.COMPARISION WITH OTHER WORKS

In the work of Rakhlin et al [3], a Deep Learning CNN approach was used to classify breast cancer histology images. The accuracy achieved was 86% by this process. In my project work, we get an 87% accuracy on the classification, which is an improvement.

7.RESULTS

On assessing the performance of the models of classifying the cancerous and non cancerous images on the test set comprising of images from each class, the two models achieves a remarkable accuracy of 85 and 87 percent respectively.
On assessing the performance of the three models SVM, Linear Regression Model and the CNN Model, the three models give a low accuracy of around 19-20 precent. This can be further improved by using Data Augmentation techniques, using various advanced machine learning models and incorporating several other important factors like tumor size, tissue hardness etc. to high quality standard histopathology image dataset.

8.CONCLUSION

These are the final results obtained from the models. Each model is curated carefully with modifications and improvements at each line of code in order to obtain the desired result.

Table 1: Classification as Cancerous and Non cancerous

Model	Accuracy	Precision	Recall	F1 score
CNN Model 1	0.85	0.73	0.66	0.70
CNN Model 2	0.87	0.76	0.70	0.73

Table 2: Classification of Cancerous Patches into Different Stages

Model	Accuracy
SVM	0.20324508966695132
Regression Model	0.20324508966695132
CNN Model	0.19214347004890442

9.REFERENCES

[1]Khalid, Arslan, Arif Mehmood, Amerah Alabrah, Bader Fahad Alkhamees, Farhan Amin, Hussain AlSalman, and Gyu Sang Choi. 2023. "Breast Cancer Detection and Prevention Using Machine Learning" Diagnostics 13, no. 19: 3113. https://doi.org/10.3390/diagnostics13193113

[2]Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms", Journal of Healthcare Engineering, vol. 2019, Article ID 4253641, 11 pages, 2019. https://doi.org/10.1155/2019/4253641

[3]Rakhlin, A., Shvets, A., Iglovikov, V. and Kalinin, A.A., 2018. Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. arXiv preprint arXiv:1802.00752

[4]ICIAR 2015 Grand Challenge on Breast Cancer Histology Images.

[6]Kerketta Z H, Kujur A, Kumari N, et al. (June 15, 2023) A Cross-Sectional Study on the Epidemiology of Newly Diagnosed Breast Cancer Patients
Attending Tertiary Care Hospitals in a Tribal Preponderant State of India: Regression Analysis. Cureus 15(6): e40489. DOI 10.7759/cureus.40489