Diagnosis of Breast Cancer Using CNN on Mammography Images Enhanced with CLAHE and Morphology Methods

1 Introduction

Breast cancer is a major form of cancer affecting large number of populations. But, unlike other form of cancer, it is highly diagnosable with mammography. Most form of breast cancer can be classified into benign and malignant. Malignant cancer poses greater risk as compared to the benign as it has high risk of overspreading. So, an early and proper diagnosis can be a matter of life and death. Today, these diagnoses are based on the human radiologist, although this is effective but, human error as well as eye fatigue can result in number of misdiagnosis including false negative and false positive. [1]

So, oracle system that takes input mammography images and provides a clear diagnosis with high accuracy can be beneficial for early detection of breast cancer. In this report, the process behind developing such a system is stated with some preliminary result.

Biomedical images are generally enhanced either using the morphological methods or Contrast Limited Adaptive Histogram Equalization (CLAHE). In contrast, to local adaptive histogram equalization, CLAHE takes the overamplification of contrast into account to clip the intensities at a certain threshold. And, different morphological methods like binarization on a certain threshold, and erosion and dilation can be used to extract a certain region of interest. In paper [2], instead of just relying on one or other form of image processing of mammography images, the author proposed to use a hybrid model that takes the benefits for both of the methods to provide a better processed image.

CNN is one of the widely used form of artificial neural network having exceptional result when working for classification problem. On CNN, the image is passed through several convolution layer that works to extract the features from the images, which then can be passed to fully connected networks and finally to map the probabilities to class thus classifying the image. Since, this a computationally expensive task, generally smaller input image is given, but in the case of mammography, even the small calcification needs to have impact on the network, so using large image size does pose benefit as well as some challenges.

2 Methodology and Implementation Details

The image information is contained in the .dcm file which needs read and further image processing task needs to be performed. Both the morphological operation and CLAHE is performed on the image. All the images are labelled and an image datastore is created. This processed image datastore then can be used using transfer learning on pretrained alexnet. The alexnet need to be modified to input larger image size. And, since we will only be classifying malignant and benign i.e two classes the output layer also needs to be updated accordingly.

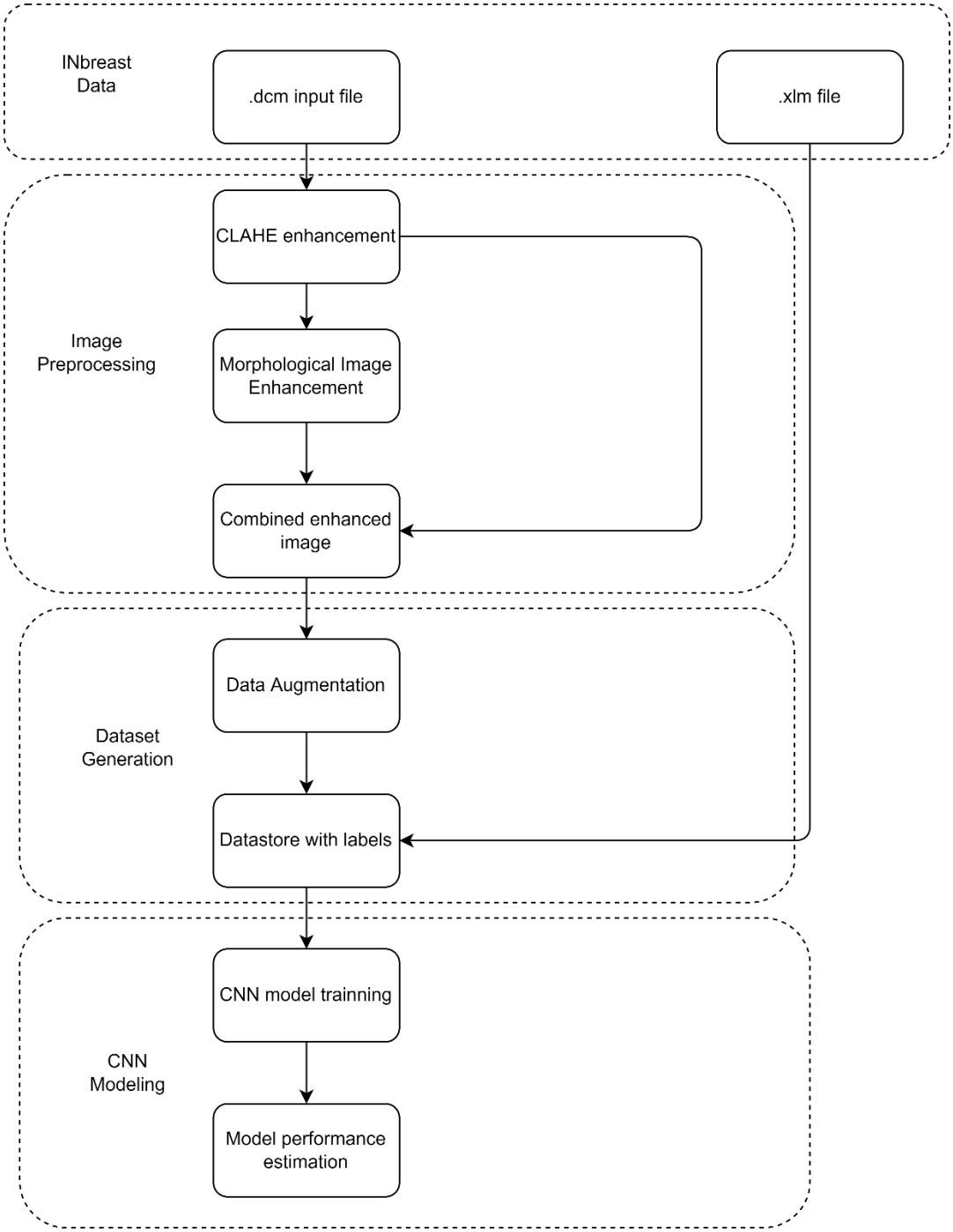

Figure 1 Block diagram of proposed system

.

2.1 INBreast Data

Most of the mammography data set contains the images that are later digitalized, which contains a lot or artifact and generally have other noises associated with it. So, instead of using those datasets, this project use INBreast dataset which is a digital dataset with .dcm files, and .xlm file

that contains the labeling of the files by experts. This xml files also have information on whether there are masses present on the images.

The code present on the original data set creates a number of variables and since are only working with the images and classification label. So, to preprocess that tablePreprocessing.m code creates a table containing PatientID, DataID, RoiID/Filename, Label. And, since there were a number of missing values those data were also removed.

```
%% Make a table tbl_inbreast that contains filename, id and classification label
%Run the code on the INBreast Release 1.0\ALLDICOMs\ folder before this

close all

s = size(INbreast);
tbl_inbreast = zeros(s);

for i = 1:s(1)
    tbl_inbreast = [INbreast(:,3),INbreast(:,7),INbreast(:,8)];
end

roi_name = extractBefore(tbl_inbreast(:,2), 9);
tbl_inbreast = [tbl_inbreast,roi_name];
tbl_inbreast = table(tbl_inbreast(:,1),tbl_inbreast(:,2),tbl_inbreast(:,4),...
               tbl_inbreast(:,3),'VariableNames',["PatientID","DataID","RoiID","Label"]);

tbl_inbreast(any(ismissing(tbl_inbreast),2), :) = [];
```

Figure 2 Code in tablePreprocessing.m

Also, INbreast.xls was exported to INbreast1 variable and this contained the mass information as well as other diagnosis. Since, the data were from 117 patients and each patient had several views of breast. Among all that information, we are currently working on the images that had mass present. So, to do so, we need to eliminate the data that didn't had defined mass. So, another table, tbl_inbreast_mass was created using the information in INbrest.xls.

Since, this table we have the filenames and label for only the images with mass, this was then used to reference every image and create a datastore. (code for this in massData.m)

2.2 Image Preprocessing

The image processing steps comprises of CLAHE enhancement and morphological enhancement. The morphological enhancement provides us with a good enough region of interest (ROI) i.e. suspicious location for cancer growth. This segmentation can then be used to extract the particular ROI, which then can be added to the CLAHE image output to produce image with distinct information.

To see the actual implementation, we can run the MATLAB code to see the differences in the morphological enhancement on original data and CLAHE enhanced image. To implement that, the following processing were done -

2.2.1 Normalizing the image

The original image just read using dicomread function in not normalized so, the image was normalized to range of 0 to 255. This makes easier to do further processing.

2.2.2 Flip all images to same direction

Some of the image are flipped right where as some are flipped left. So, to make a consistency in the original data all images were flipped to left side.

2.2.3 CLAHE processing

The image was then enhanced using contrast limited adaptive histogram equalization. This is one of the effective methods that helps define the masses present on the image. Clip limit of 2/256 was set up based upon visual inspection of the result.

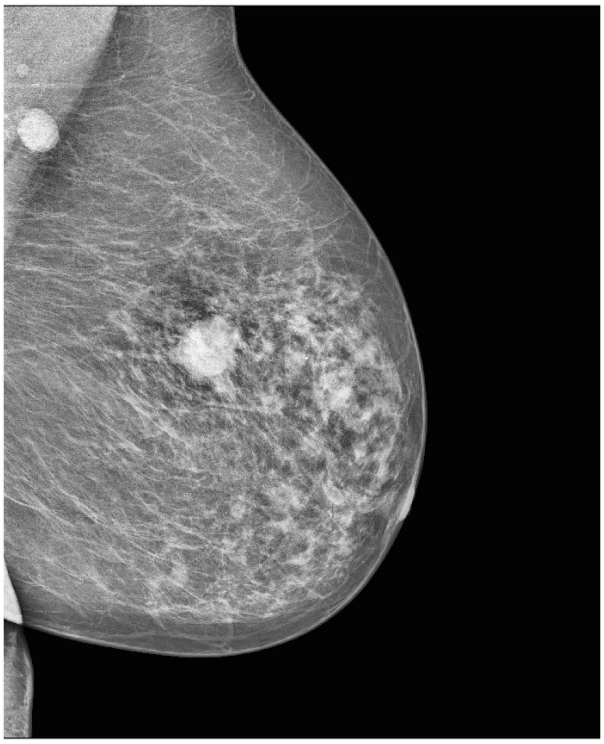

Figure 3 CLAHE processed

2.2.4 Binarize Image

The image was binarized that defines every pixel as either high or low based on a particular threshold. This is done to do morphological enhancements.

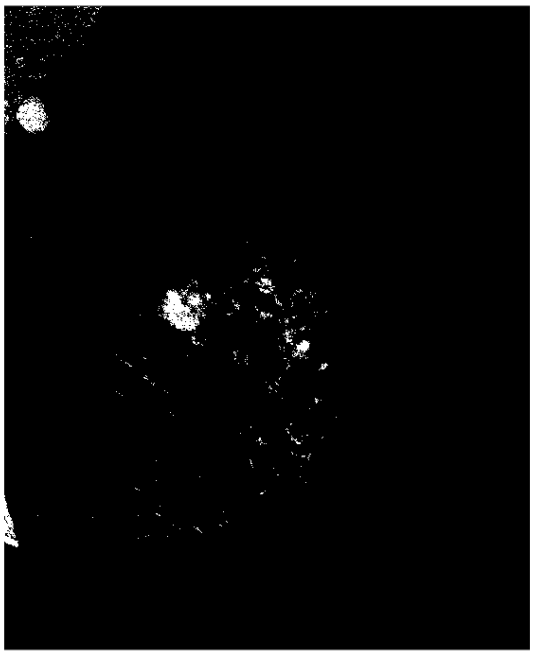

Figure 4 Binarized Image

2.2.5 Morphological enhancement

If the original image is enhanced using morphological enhancement, there are several artifacts that affect the final learning process. So, instead this was applied the CLAHE output, which reduces the many artifacts present in the image, and help to distinguish a region of interest (possible tumors and masses) on image. First the image was dilated, this fill up the hole or missing values around

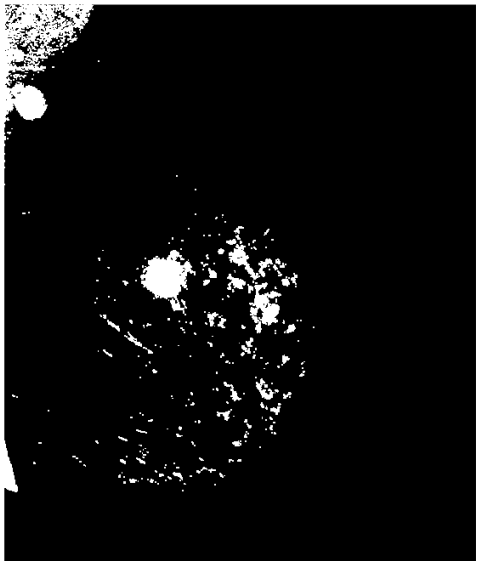

Figure 5After Dilation

Figure 6 After Erosion

the ROI present on the image and then erosion removes/erode from the boundary. This process removes the unnecessary artifacts present on the image.

2.2.6 Masking and Blurring

The result was then used to create an inverted mask, which was then blurred using gaussian blurring at sigma of 5. This help to remove sharpness from the mask which when subtracted from original image provide us with the region of interest in the original image

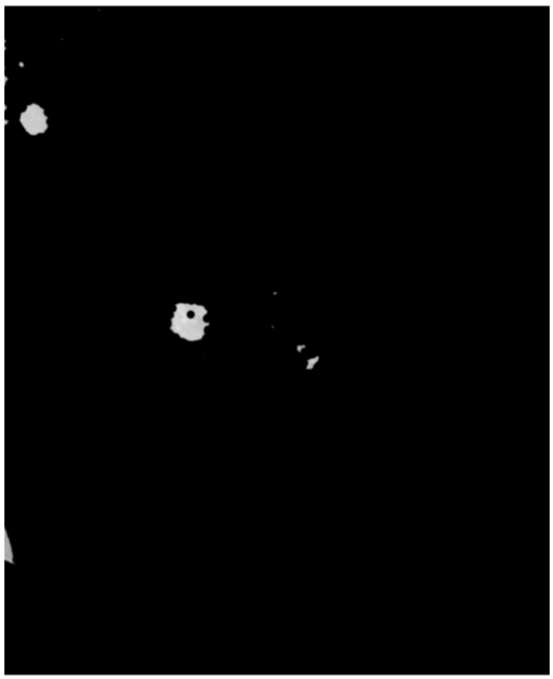

Figure 7 Original Image after Masking

2.2.7 Weighted Addition

The result from masking was weighted and added to CLAHE output. This gives us with a good output image to be used for the next step.

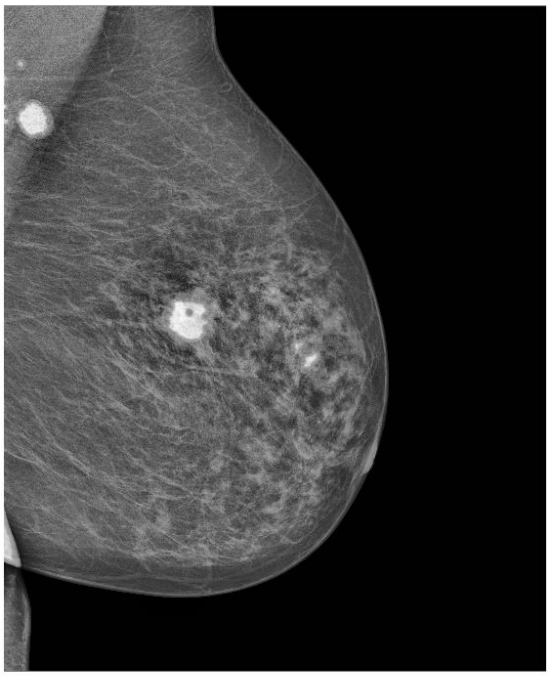

Figure 8 Preprocessed Image

2.3 Dataset Generation

2.3.1 Padding

The preprocessed image is then padded to form a square image and resized to 512x512 as most of the pretrained algorithm works with square images.

2.3.2 Augmentation

Then the images were augmented by rotating at an interval of 30 degree and were flipped to give a good enough number of datasets. All the images were saved in a particular folder to create a data store. This datastore was then fed into the training model.

2.4 CNN Model

Before starting the training process, the dataset was divided into training and testing set. There is also a color processing step that changes the gray image to be represented as color image. This needs to be done as most pretrained are trained for RGB image. Also, to equate the number of images per classes, random selection to minimum available data was selected.

The testing set was used for validation of the system. The pretrained alexnet was used to create a model and some modification i.e. changing the input layer to match the image size of out datastore and the final classification layer was changed to classify only two categories of malignant and benign. Alexnet was the choice because of the fact that it has shown to provide better accuracy

compared to other pretrained models [3]. Accuracy and loss performance metric are used here during the training process which quantifies the quality of the trained model.

Figure 9 CNN layers used

3 Result

The model was trained with out training data set and the progress of the training can be seen in the figure below. And, the training was stopped at epoch 40 to avoid overfitting of the data. The final validation accuracy was obtained to be 79.90%.

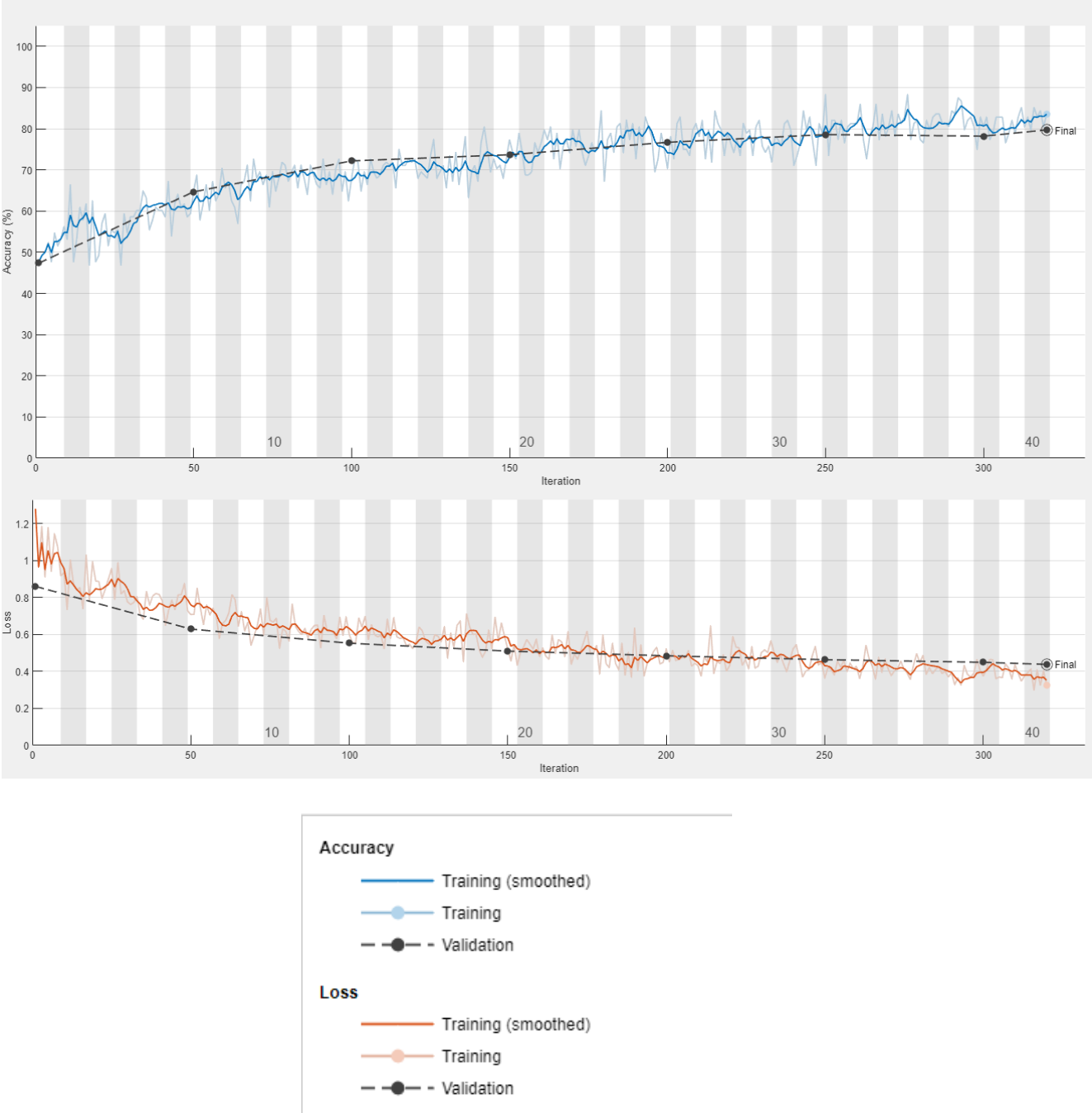

Figure 10 Training progress

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Validation Accuracy	Mini-batch Loss	Validation Loss	Base Learning Rate
1	1	00:00:03	47.66%	47.37%	1.2809	0.8610	1.0000e-05
7	50	00:01:10	64.84%	64.66%	0.7086	0.6295	1.0000e-05
13	100	00:02:19	67.97%	72.18%	0.5990	0.5531	1.0000e-05
19	150	00:03:27	77.34%	73.68%	0.5239	0.5103	1.0000e-05
25	200	00:04:35	70.31%	76.69%	0.5216	0.4850	1.0000e-05
32	250	00:05:43	88.28%	78.57%	0.3649	0.4639	1.0000e-05
38	300	00:06:50	80.47%	78.20%	0.3753	0.4497	1.0000e-05
40	320	00:07:17	83.59%	79.70%	0.3261	0.4384	1.0000e-05

Figure 11 Training metrics

4 Conclusion

The model performs with good accuracy, but not best accuracy that can be achieved. This is due to the fact a same parameter was used for the image processing of all the images. Instead, if one is to change parameter based on the images that it is working on, the system might perform better. Also, not all data could be utilized so more augmentation can help with learning process.

References

[1] I. Z.N, I. Jannat-Dastjerdi, F. Eskandari, S. Ghoushchi and Y. Pourasad, "Presentation of Novel Hybrid Algorithm for Detection and," *Computational Intelligence and Neuroscience*, vol. 2021, no. Special Isssue, p. 14, 2021.

[2] N. Kharel, A. Alsadoon and P. Prasad, "Early Diagnosis of Breast Cancer Using Contrast," in *International Conference on Information and Communication Systems (ICICS)*, Irbid, 2017.

[3] E. Omonigho, M. David, A. Adejo and S. Aliyu, "Breast Cancer:Tumor Detection in Mammogram Images Using Modified AlexNet Deep Convolution Neural Network," in *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*, Ayobo, 2020.

Note – Algorithm used for this project are somewhat based on N. Kharel, A. Alsadoon and P. Prasad, "Early Diagnosis of Breast Cancer Using Contrast," in *International Conference on Information and Communication Systems (ICICS)*, Irbid, 2017. paper and https://towardsdatascience.com/can-you-find-the-breast-tumours-part-1-of-3-1473ba685036 web article. Other several web articles and documentaion was used to develop this system.