An unstructured mammogram analysis for feasible classification and detection of breast cancer using a convolutional approach

Abstract. Currently, different methods are available for the purpose of breast cancer classification and detection. Most of these techniques are well appreciated by society and in response to the demand of society, almost every year the different techniques are introduced by different researchers. But it does not satisfy the demand of current requirements. Under such a situation, we are going to propose a new breast cancer classification and detection algorithm using a convolutional approach. This technique starts with the mammogram preprocessing step. It is followed by the convolutional model architecture design step. In the next step, the segregation of the dataset into the training and testing phase is performed. Then the convolutional model architecture is trained using the training dataset and pre-masked images. After that our proposed algorithm predicts, the breast cancer detection and classification result. We have found that our proposed algorithm can be used for breast tumor detection and classification from mammogram images with the average approximate accuracy of 98.5% and the average approximate F1 Score of 0.98. Novel preprocessing steps and modifications in the convolutional architecture make the proposed methodology unique. Due to high performance, novelty, ease of use, our proposed method is useful to develop any mobile or web application in the future.

Keywords: Breast Cancer, Mammogram, Pre-masked image, Convolutional, Accuracy Score, F1 Score.

1 Introduction

Breast Cancer is considered to be most common nowadays. According to the December 2020 statistics of the World Health Organization (WHO) among 24.5% of world's female cancer population is affected by this disease. It starts spreading from breast. At the early stage of this disease, the cancerous cells we may be found as a tumor or lump, which is clearly visible from the X-Ray images. Most of the breast tumors are found as non-cancerous in nature and often observed that they do not spread outside the breast, whereas malignant breast tumors are cancerous in nature and may spread outside the breast. Patients of this disease are often suffered from symptoms like breast or nipple pain, swelling of all or part of a breast, skin dimpling, nipple retraction, reddish nipple or breast skin, nipple discharge, swollen lymph nodes, etc. Among many types of this disease ductal carcinoma in situ (DCIS) and invasive carcinoma are considered to be the most common and phyllodes tumors and angiosarcoma are considered as rare.

Hence, it is considered as a serious issue of society and an automated computer-based method is required to detect and classify breast cancer with precision from mammogram images. This method helps the medical practitioner to start early treatment and also helps to reduce the fatality rate. The inconclusive, incomplete, and dissatisfactory results of the previously proposed techniques encourage us to develop a novel breast cancer classification and detection algorithm using a convolutional approach that works on unstructured data such as mammogram images. It displays the classified output using convolutional model architecture along with the satisfactory accuracy rate and F1 score.

Our paper is divided into various sections. In section 2, we explain the advantages and disadvantages of some pre-existing techniques, in section 3 we explain our proposed methodology which consists of the main architecture of the method and the algorithm, whereas the experiment result and analysis are explained in section 4.

2 Literature Survey

This section represents various pre-existing methods [7] [8] [9], along with their advantages and disadvantages as shown in the Survey Table (**Table. 1**). All these methods are well appreciated but in context with our problem they are producing inclusive, incomplete, dissatisfactory results. In-depth analysis of these methodologies has proven to be very competent to identify the downsides. Identification of these drawbacks helps us to update and modify our algorithm and code and to calculate the accuracy rate. The Survey Table is given below:

Table. 1. Existing Methodology Analysis Table

Methodology	**Advantages**	**Disadvantages**
Decision Tree [1]	Decision trees require less effort for data preparation during pre-processing.	A small change in the data can cause a large change in the structure. According to our problem, it causes instability.
K-Nearest Neighbor [2]	No assumption about data.	KNN needs a huge amount of memory.
Random Forest Classifier [3]	Random forest minimizes the overfitting issue and tries to increase the accuracy score.	This method needs high computational power and resources.
Support Vector Machine [4]	It is effective in high-dimensional spaces.	Due to a huge time consumption issue, the performance of this method is not satisfactory when we apply this strategy in a large dataset.
Gaussian Naïve Bayes [5]	Naive Bayes is more appropriate for categorical input variables than numerical variables. In the context of our problem, this property is suitable.	It presumes that every feature of the dataset is independent. This property limits the application of this technique in real-world cases.
Multi-Scale Fusion U-Net [6]	It solves the problem of multi-scale variation in breast lesions and boundary pixel blurring.	For the different modes of images, segmentation effects will be reduced.

3 Proposed Methodology

Our proposed methodology is focused on breast cancer classification and detection from an unstructured dataset (BUSI) [10] using breast mammogram classification and detection algorithm using a convolutional approach. We have proposed a block diagram to show the main concept of the methodology at a glance as shown in **Fig. 1.**

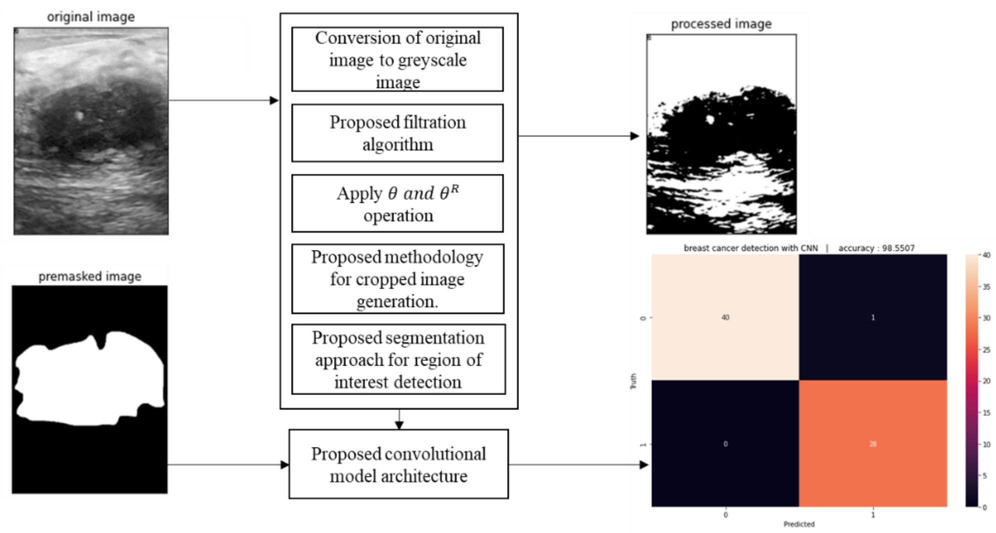

original image
premasked image
Conversion of original image to greyscale image
Proposed filtration algorithm
Apply θ and θ^R operation
Proposed methodology for cropped image generation.
Proposed segmentation approach for region of interest detection
Proposed convolutional model architecture
processed image
breast cancer detection with CNN | accuracy : 98.5507
Truth
Predicted
0
1
40
1
0
28
0
5
10
15
20
25
30
35
40

Fig. 1. The built-in architecture of mammogram classification and detection algorithm using a convolutional approach

3.1 Algorithm

Our algorithm is divided into 16 steps. The algorithm takes a mammogram image, $I(u,v)$ as an input from the dataset [10] and produces a binary output (Yes/No) and the accuracy rate. The algorithm is as follows:

Algorithm:	mammogram classification and detection algorithm using a convolutional approach
Input:	$I(u,v)$
Output:	Yes/No and the accuracy rate

1. Read an image $I(u,v)$ from the dataset.
2. Convert the image to a grayscale image,$G(u,v)$.
3. Preprocessed $G(u,v)$ using the following formula and 5x5 kernel:

$$G_1(u,v) = \frac{1}{2\pi d^2} e^{\frac{-u^2+v^2}{2d^2}} \tag{1}$$

   Where,
   $G_1(u,v)$ = Processed Image
   d = Standard Deviation
   u = uth row in the grayscale image
   v = vth column in the grayscale image

4. Perform segmentation on $G_1(u,v)$ using the following equation:
   If pixel $(u,v) \geq \tau$ then,
   $(u,v) \leftarrow 1$,
   Otherwise,
   $(u,v) \leftarrow 0$, provided maximum pixel value = 255

   Where,
   $G_1(u,v)$ = grayscale image
   (u,v) = a pixel of an image in (u,v) position

τ = selected threshold value in our algorithm. It is tested after the trial-and-error method.
1 = Light
0 = Dark

5. Apply θ operation on each pixel (u, v) of $G(u, v)$ using a 3x3 kernel window, to probe and reduce the shape stored in $G(u, v)$.

$$G_2(u, v) \leftarrow G_1(u, v)\,\theta\,\Delta(u, v) \leftarrow \{\Phi \in \varepsilon \mid \Delta\,(u, v)\varphi \; subset\ of\ G_1(u, v)\} \qquad (2)$$

Where,
$\varepsilon \leftarrow$ A Euclidean space
$G_1(u, v) \leftarrow$ a binary image in ε
$\Delta(u, v)_\varphi \leftarrow \Phi$is the translation of Δ by the vector Φ

6. Apply θ^Roperation on each pixel (u, v) of $G_2(u, v)$ as follows:

$$G_3(u, v) \leftarrow G_2(u, v)\theta^R\Delta(u, v) \bigcup_{\gamma\ \epsilon\ G_1(u,v)} \Delta\,(u, v)_\gamma \qquad (3)$$

Where,
$\varepsilon \leftarrow$ A Euclidean space and $G_2(u, v)$ a binary image after θ operation in ε
$\Delta(u, v) \leftarrow$ The structuring element
$\Delta(u, v)_\gamma \leftarrow$ The Translation of $\Delta(u, v)$ by γ

7. Find the best-bounded box of image $G3(u, v)$ using the following function.
7.1.

$$C_{u,v} \rightarrow \rho(G_3(u, v), \alpha(u, v), \beta(u, v) \qquad (4)$$

Where,
$C_{(u,v)} \rightarrow$ collection of boundary points
$\rho() \rightarrow$ a method used for boundary points generation
$G_3(u, v) \rightarrow$ processed image after step 6
$\alpha(u, v) \rightarrow$ a method to retrieve outer boundary
$\beta(u, v) \rightarrow$ a method to store the endpoints of the horizontal, vertical and diagonal boundary

7.2. Merge all the boundary points together & produce the final boundary box.

8. Calculate the extreme left, right, top, and bottom-most corner points.
9. Create a cropped image $G_5(u, v)$ by segregating the blank area from the image $G_4(u, v)$ using the following equation.

$$G_5 \leftarrow C(t(u, v), b(u, v), l(a, v), r(u, v)) \qquad (5)$$

Where,
$C() \rightarrow$ a method to convert cropped images.
$t(u, v) \rightarrow$ a method to extract the extreme topmost corner point.
$b(u, v) \rightarrow$ a method to extract the extreme bottommost corner point.
$l(u, v) \rightarrow$ a method to extract the leftmost corner point.
$r(u, v) \rightarrow$ a method to extract the rightmost corner point.
$(u, v) \rightarrow$ the u^{th} row and v^{th} column.

10. Apply τ operation on $G_5(u, v)$ using the following equation:

$$G_5(u, v) \leftarrow \tau(G_5(u, v), \text{selected } \tau \text{ value, maximum } \tau \text{ value, } \tau_1, \beta(u, v)) \qquad (6)$$

Where,
$G_5(u, v) \rightarrow$ a processed image
$\tau() \rightarrow$ a thresholding method
$G_5(u, v) \rightarrow$ a cropped image
selected τ value $\leftarrow$ 100 in our algorithm
maximum τ value $\leftarrow$ 255 in our algorithm
$\tau_1 \leftarrow$if a pixel $(u, v) \geq$ Selected τ value then $(u, v) \leftarrow 1$, otherwise $(u, v) \leftarrow 0$
$\beta(u, v) \leftarrow$ a method to store the endpoints of horizontal, vertical, and diagonal boundary points

11. Read the pre-masked image $\rho(u, v)$ from the dataset (D).
12. Resize $G_4(u, v)$ into a 128 X 128 image.

13. Declaring $A[n][2]$ list for each image for storing true and false sample values.

```
$n \leftarrow$ total number of images
$end\_of\_iterations \leftarrow n - 1$
$i \leftarrow no\_of\_iterations \leftarrow 0$
While ($i \leq end\_of\_iterations$) then,
        If $G_6(u, v)$ is a tumourous image then,
                $G_6(u, v) \leftarrow$True
                $A[i][0] \leftarrow 1$ and $A[i][1] \leftarrow 0$
                Go to Otherwise
        If$G_6(u, v)$is a non-tumorous image then,
                $G_6(u, v) \leftarrow$ False
                $A[i][0] \leftarrow 0$ and $A[i][1] \leftarrow 1$
                Go to Otherwise
        Otherwise
                $i \leftarrow i + 1$
                Go to While
        End If
End While
```

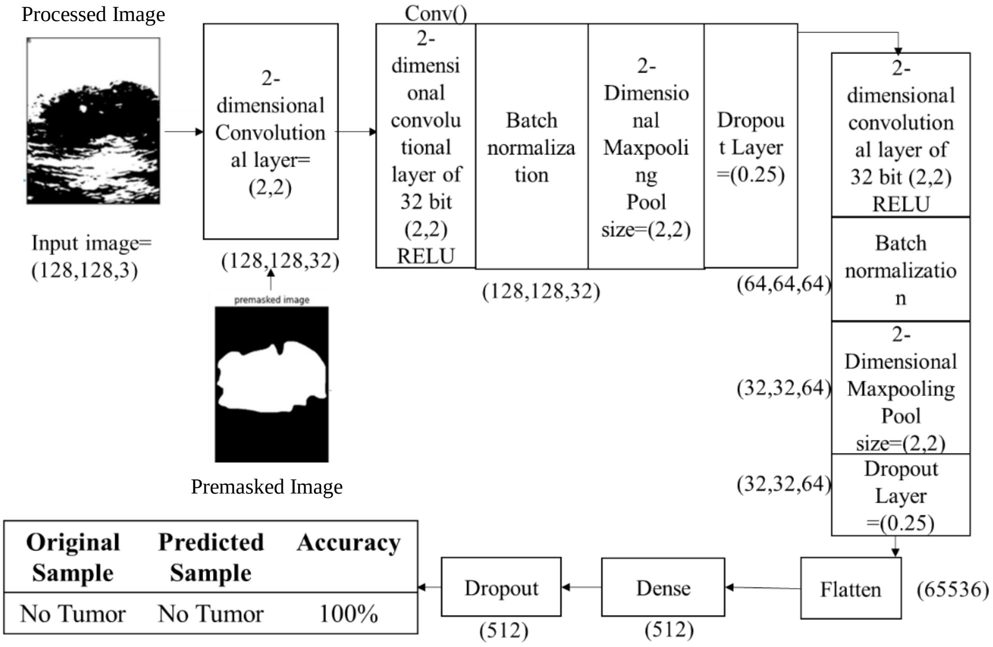
14. The proposed convolutional architecture each has an input x of size 128x128x3 is used to feed into it. In the convolutional architecture following layers are observed:

    a. One 2-dimensional convolution layer with kernel size 2x2
    b. One 2-dimensional convolution layer with kernel size 2x2 and a rectified linear activation function
    c. One batch normalization layer
    d. One 2-dimensional max pooling layer with pool size 2x2
    e. One dropout layer
    f. One 2-dimensional convolution layer of 64-bit input and a rectified linear activation function
    g. One 2-dimensional convolution layer of 64-bit input and a rectified linear activation function
    h. One batch normalization layer to calculate the mean output (close to 0) and the standard deviation output (close to 1)
    i. One 2-dimensional max pooling layer with pool size 2x2
    j. One dropout layer
    k. One flatten layer
    l. One dense layer and activation layer
    m. One dropout layer
    n. One dense layer with an activation softmax function

15. Segregate the whole dataset [10] into two sections. We select 80% mammogram images from the dataset (BUSI) [10] for training purposes and 20% mammogram images from the dataset (BUSI) [10] for testing purposes.
16. We feed the training and testing data into the designed convolutional model architecture to generate result.

3.2 Model Architecture

Mammogram training and testing phase can be explained with the help of a model architecture as shown in **Fig.2.** the architecture accepts preprocessed mammogram image and premasked image as output and displays predicted sample results and accuracy for the same.



Original Sample	Predicted Sample	Accuracy
No Tumor	No Tumor	100%

Fig.2. Mammogram training and testing phase using proposed model architecture

4 Experimental results

We consider the mammogram image dataset (BUSI) [10], which consists of 410 malignant and 130 non-malignant images collected from the dataset [10] and 209 malignant and 133 non-malignant pre-masked images to calculate the performance of mammogram classification and detection algorithm using a convolutional approach. The size, colour, and format of breast tumor images in the dataset [10] are similar in nature, whereas the resolutions of the images are different. The format of the mammogram images are '.png' by nature.

We have applied our algorithm in the python environment, version 3.8, with the hardware configuration of the Intel Core i3 5th Generation processor,4GB DDR3 primary memory (RAM), and an integrated graphics card. Anaconda as a distributor of Python version 3.8 is used. jupyter notebook version 6.3.0 as an open web interface is used as a programming platform for the implementation of our algorithm.

After reading the images from the local machine, we partitioned them into two categories, i.e., "yes" (malignant) and "no" (normal) and we have achieved the following results as shown in **Fig. 3.** and **Fig. 4**.

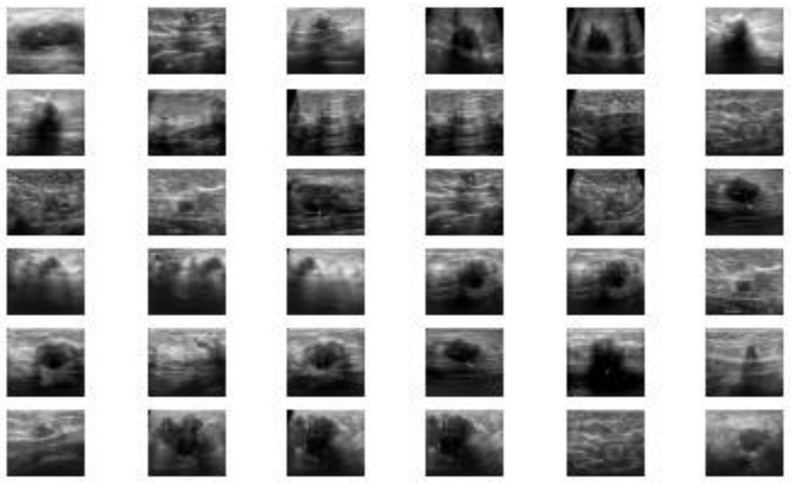


Fig. 3. Malignant images

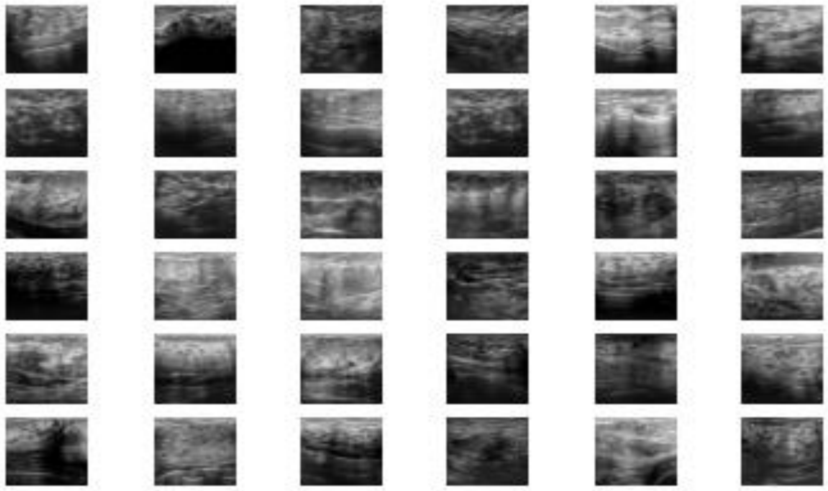


Fig. 4. Non-malignant images

After reading all the images from the dataset each image will pass through the various steps of the algorithm as shown in **Fig. 5.**

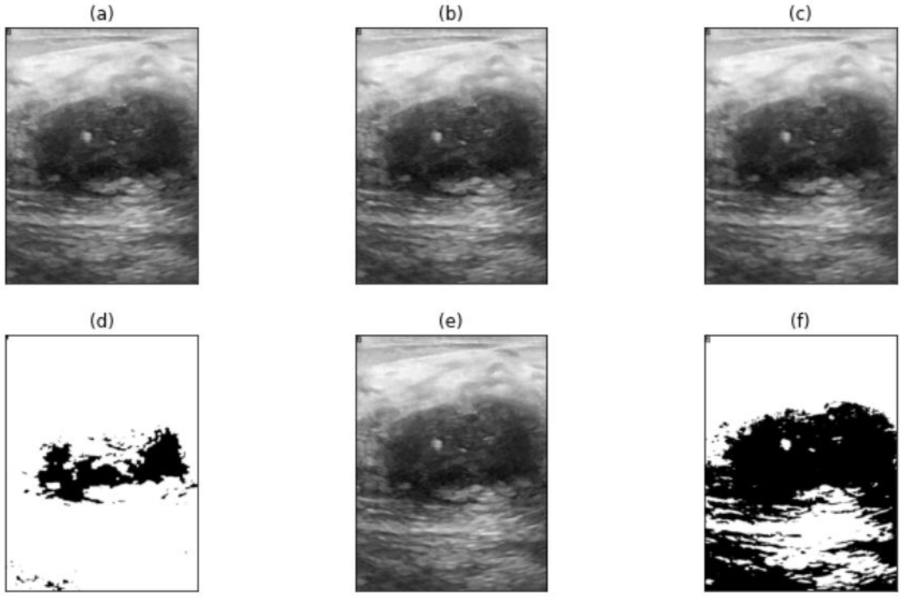


Fig. 5. (a) original image (b) gray-scale image (c) image after applying proposed filtration method (d) image after applying $\theta \, and \, \theta^R$ operation (e) cropped image after applying proposed image cropping methodology (f) proposed segmentation approach for the region of interest detection

We are splitting our dataset (BUSI) [10] into training and testing sets. 80% of the mammogram images are used for training and the rest is used for testing. After the execution of our proposed algorithm, we observed that the number of training samples are 344, number of testing samples are 69 whereas training shape values are (344, 128, 128, 3) and (344, 1). Our proposed convolutional layered architecture is trained. It is used to train on n number of samples (344 in our case). With the changes of several iterations or epochs, the total time consumed by the algorithm is calculated.

Table 2: An example of the training phase.

SNo.	Epoch No.	Time Taken (s)	Loss	Value Loss
1.	1	7.989	0.0000014170	1.0888
2.	2	7	0.0000003455	1.1316
3.	3	7	0.0000011035	1.1742
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
13.	13	7	0.0000001478	1.2158
14.	14	7	0.0000001231	1.2586
15.	15	7	0.0000001321	1.2025

After calculation of accuracy and loss during the testing and validation phase, we have discovered two graphs, as shown in **Fig. 7.** and **Fig. 8.** respectively.

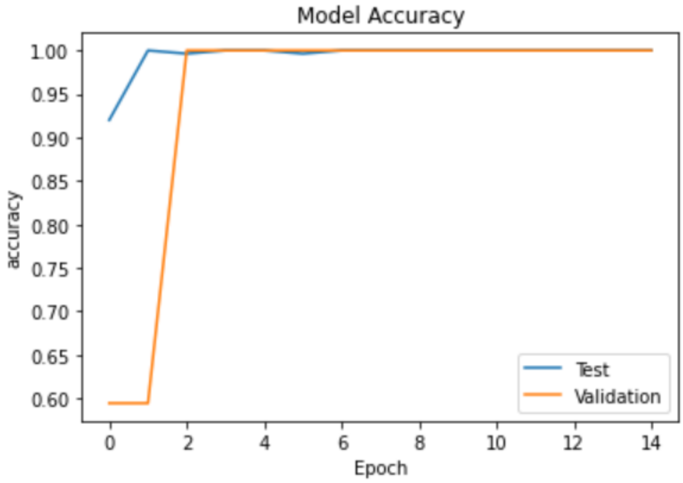


Fig. 7: Accuracy Graph

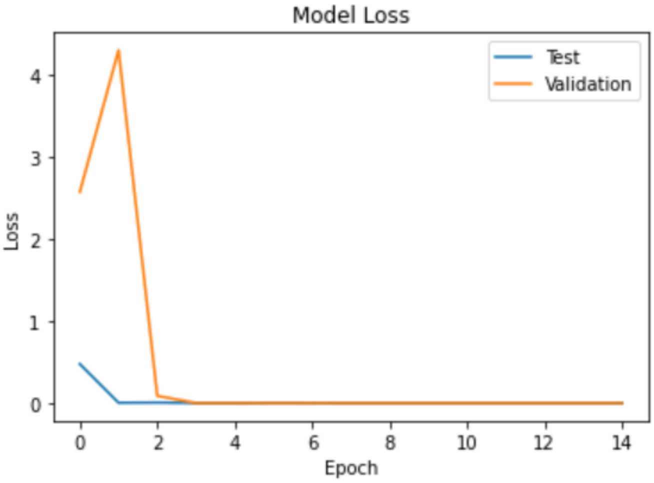


Fig. 8: Loss Graph

After inserting the previously calculated amount of testing samples it is observed that the amount of loss in the test samples or false prediction cases are 0.02 whereas 0.98 or 98% accuracy or right prediction is received from our algorithm. It has been observed that the average F1 Score is 0.98.

Table 3: Prediction result and accuracy calculation

SNo.	Original Sample	Predicted Sample	Accuracy (%)
1	No tumor	No tumor	100
2	Tumor	Tumor	100
3	Tumor	Tumor	100
4	No tumor	No tumor	100
.	.	.	.
.	.	.	.
.	.	.	.
66	No tumor	Tumor	0
67	No tumor	No tumor	100
68	Tumor	Tumor	100
69	No tumor	No tumor	100

In **Table 3** displays 69 randomly selected original samples from the dataset, corresponding predicted sample results and the accuracy score. In the end, we have found that our proposed algorithm can be used for breast tumor detection and classification from mammogram images with an average approximate accuracy of 98.5%. As the data is balanced, we are considering the results to be satisfied. Comparing it with other existing methodologies a satisfactory result is observed as shown in **Table 4.** We can conclude that our proposed methodology overpowers the existing classification algorithms [7] [8] [9] in terms of F1 and accuracy score.

Table 4: Comparison chart

Serial No.	**Name of the classification algorithm**	**Accuracy Score**	**F1 Score**
1.	Decision Tree [1]	97%	0.97
2.	K-Nearest Neighbor [2]	94.2%	0.94
3.	Gaussian Naïve Bayes [3]	97%	0.97
4.	Random Forest [4]	97%	0.95
5.	Support Vector Machine [5]	96%	0.96
6.	Multi-Scale Fusion U-Net [6]	96%	0.96
7.	**Proposed Method**	**98.5%**	**0.98**

5 Conclusion

In this paper, we have proposed a mammogram classification and detection algorithm using a convolutional approach. This algorithm is capable of preprocessing unstructured mammogram images from the BUSI dataset [10], collected from the web resource. It is responsible for the prediction of malignant and non-malignant mammogram images in terms of Yes (malignant sample) and No (non-malignant sample) values. It also generates an accuracy and F1 score through which we can compare our proposed method with existing methods [7] [8] [9]. The experimental result shows that after applying the proposed and existing methods [7] [8] [9] on the BUSI dataset [10], our technique is producing approximately 98.5% accuracy and 0.98 F1 scores on average. This result is considered to be satisfactory and based on this result we can say that the proposed algorithm overpowers the efficiency of the existing method as described in **Table 4.** Novel preprocessing steps and modifications in the convolutional architecture using multiple layers make the proposed methodology unique. Due to high performance, novelty, ease of use, our proposed method is useful to develop any mobile or web applications in the future. Our method can be tested on various mammogram datasets to identify the generic performance of the proposed method in the future. The performance of our method may be increased by making necessary modifications in algorithm and code.

6 References

1. L. Yi and W. Yi, "Decision Tree Model in the Diagnosis of Breast Cancer," *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC)*, 2017, pp. 176-179, doi: 10.1109/ICCTEC.2017.00046.
2. Hadidi, moh'd & Alarabeyyat, Abdulsalam & Alhanahnah, Mohannad. (2016). Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm. 35-39. 10.1109/DeSE.2016.8.
3. B. Dai, R. -C. Chen, S. -Z. Zhu and W. -W. Zhang, "Using Random Forest Algorithm for Breast Cancer Diagnosis," 2018 International Symposium on Computer, Consumer and Control (IS3C), 2018, pp. 449-452, doi: 10.1109/IS3C.2018.00119.
4. Shang Gao and Hongmei Li, "Breast cancer diagnosis based on support vector machine," 2012 2nd International Conference on Uncertainty Reasoning and Knowledge Engineering, 2012, pp. 240-243, doi: 10.1109/URKE.2012.6319555.
5. H. Kamel, D. Abdulah and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," 2019 International Engineering Conference (IEC), 2019, pp. 165-170, doi: 10.1109/IEC47844.2019.8950650.
6. J. Li, L. Cheng, T. Xia, H. Ni and J. Li, "Multi-Scale Fusion U-Net for the Segmentation of Breast Lesions," in IEEE Access, vol. 9, pp. 137125-137139, 2021, doi: 10.1109/ACCESS.2021.3117578.
7. S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," *2021 International Conference on Artificial Intelligence (ICAI)*, 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.
8. M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391453.
9. B. Bılgıç, "Comparison of Breast Cancer and Skin Cancer Diagnoses Using Deep Learning Method," *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021, pp. 1-4, doi: 10.1109/SIU53274.2021.9477992.
10. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data in Brief. 2020 Feb;28:104863. DOI: 10.1016/j.dib.2019.104863.