University of Mumbai

Breast Cancer Detection and Classification using a Transfer Learning Approach

Submitted at the end of semester VII in partial fulfilment of requirements

For the degree of

Bachelor of Technology

By

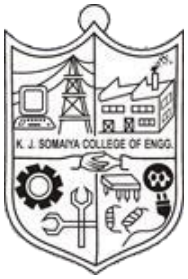
Ayush Gupta – 1913012

Saachi Dholakia – 1913013

Pranab Mehrishi – 1913042

Under the guidance of

Dr. Anudeepa S. Kholapure



Department of Electronics and Telecommunication Engineering

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Batch 2019 – 2023

K. J. Somaiya College of Engineering, Mumbai-77
(Autonomous College Affiliated to University of Mumbai)

Certificate

This is to certify that the dissertation report entitled **Breast Cancer Detection and Classification using a Transfer Learning Approach** submitted by Ayush Gupta, Saachi Dholakia and Pranab Mehrishi at the end of semester VII of LY B. Tech. is a bona fide record for partial fulfilment of requirements forthe degree of Bachelors in Technology in Electronics and Telecommunication Engineering ofUniversity of Mumbai.

Guide	Head of the Department

Principal

Date: December 14, 2022

Place: Mumbai – 77

K. J. Somaiya College of Engineering, Mumbai-77
(Autonomous College Affiliated to University of Mumbai)

Certificate of Approval of Examiners

We certify that this dissertation report entitled **Breast Cancer Detection and Classification using a Transfer Learning Approach** is bona fide record of project work done by Ayush Gupta, Saachi Dholakia and Pranab Mehrishi during semester VII.

This project work is submitted at the end of semester VII in partial fulfilment of requirements for the degree of Bachelors in Technology in Electronics and Telecommunication Engineering of University of Mumbai.

_____

Internal Examiners

_____

External/Internal Examiners

Date: December 14, 2022

Place: Mumbai – 77

K. J. Somaiya College of Engineering, Mumbai-77
(Autonomous College Affiliated to University of Mumbai)

Declaration

We declare that this written report submission represents the work done based on our and / or others' ideas with adequately cited and referenced the original source. We also declare that we have adhered to all principles of intellectual property, academic honesty and integrity as we have not misinterpreted or fabricated or falsified any idea/data/fact/source/original work/ matter in my submission.

We understand that any violation of the above will be cause for disciplinary action by the college and may evoke the penal action from the sources which have not been properly cited or from whom proper permission is not sought.

_____ Signature of Student _____ Roll Number	_____ Signature of Student _____ Roll Number
_____ Signature of Student _____ Roll Number	

Date: December 14, 2022

Place: Mumbai – 77

Abstract.

The most frequently occurring cancer among women is breast cancer. There is a chance of fifty percent fatality in a case as one in two women diagnosed with breast cancer deaths in the cases of Indian women. This paper aims to use transfer learning technique to work with a model giving highest amount of accuracy without the issue of overfitting. The InBreast and MIAS data set was used as a training set to compare the performance of the various deep learning techniques in terms of key parameters such as accuracy and precision. The results obtained are very competitive and can be used for detection and treatment.

Content.

Chapter 1. Introduction

Worldwide, breast cancer comprises 10.4% of all cancer incidences among women, making it the second most common type of non-skin cancer (after lung cancer) and the fifth most common cause of cancer death. In 2004, breast cancer caused 519,000 deaths worldwide (7% of cancer deaths; almost 1% of all deaths). Breast cancer is about 100 times more common in women than in men, although males tend to have poorer outcomes due to delays in diagnosis. Cancer cells are very similar to cells of the organism from which they originated and have similar (but not identical) DNA and RNA. This is the season why they are not very often detected by the immune system, in particular, if it is weakened.

Cancer cells are formed from normal cells due to a modification/mutation of DNA and/or RNA. These modifications/mutations can occur spontaneously ill Law of Thermodynamics – an increase of entropy) or they may be induced by other factors such as; nuclear radiation, electromagnetic radiation (microwaves, X-rays, Gamma-rays, Ultraviolet-rays, etc.), viruses, bacteria, fungi, parasites (due to tissue inflammation/irritation, heat, chemicals in the air, water and food, mechanical cell-level injury, free radicals, evolution and aging of DNA and RNA, etc. All these can produce mutations that may start cancer. Cancer can be called therefore "Entropic Disease" since it is associated with the increase of entropy of the organism to the point where the organism cannot correct this itself External intervention is required to allow the organism to return to a stable entropic state.

Cancer develops if the immune system is not working properly and/or the number of cells produced is too great for the immune system to eliminate. The rate of DNA and RNA mutations can be too high under some conditions such as; an unhealthy environment (due to radiation, chemicals, etc.), poor diet (unhealthy cell environment), people with genetic predispositions to mutations and people of advanced age (above 80).

During the early stages of the disease, the symptoms are not presented well and hence diagnosis is delayed. It is recommended by the NBCF (National Breast Cancer Foundation) that women over the age of forty years of age should get a mammogram once a year. A mammogram is an X-ray of the breast. It is a medical technique used for the detection of breast cancer in women without any side effects deeming the procedure safe. Women who get regular mammograms have a higher survival rate as compared to women who do not. In 2018, over six hundred thousand fatalities were caused by breast cancer. The number is approximately fifteen percent of the total deaths resulting from all types of cancer among women. The chances of contracting

this particular type of cancer is usually higher in urban regions; however, the rate ofcontraction seems to be on an upward rising trend globally. The only current method of improving the results of breast cancer cases is early diagnosis and screening.

The current system uses two radiologists to analyze each woman's X-rays. In rare cases where they disagree, a third doctor assesses the images. In the research study, an AI model was given anonymized images, so that the women could not be identified.

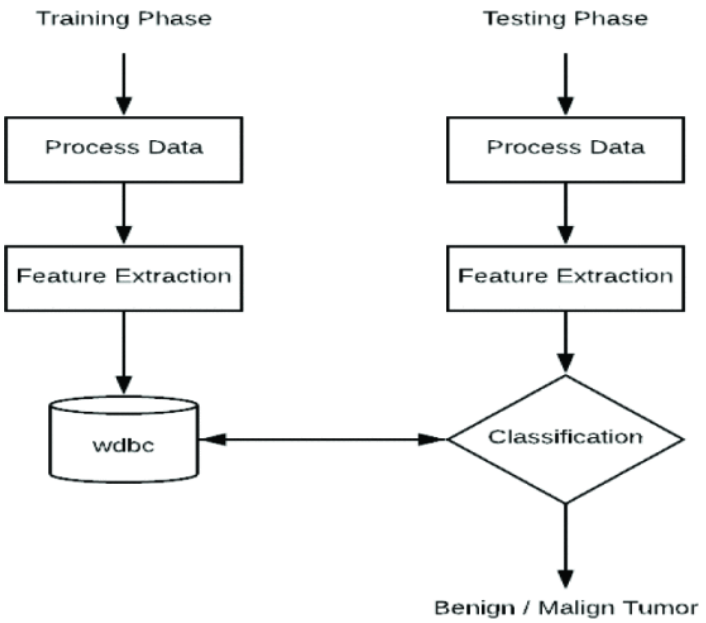


Figure 1. Separation of Training and Testing Phases

1.1 Background.

Now days, breast cancer is the most frequently diagnosed life-threatening cancer in women and the leading cause of cancer death among women. Over the last two decades, research related to breast cancer has led to extraordinary progress in our understanding of the disease, resulting in more efficient and less toxic treatments. Increased public awareness and improvedscreening have led to earlier diagnosis at stages amenable to complete surgical resection and curative therapies.

Consequently, survival rates for breast cancer have improved significantly, particularly in younger women. This article addresses the types, causes, clinical symptoms, and various approaches both non- drug (such as surgery and radiation) and drug treatment (including chemotherapy, gene therapy, etc.) of breast cancer.

Breast cancer is the most common cause of cancer in women and the second most common cause of cancer death in women in the U.S.

Breast cancer refers to cancers originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk.

1.2 Motivation.

Unlike the human experts, who had access to the patient's history, AI had only the mammograms to go on. The results showed that the AI model was as good as the current double-reading system of two doctors. And it was actually superior at spotting cancer than a single doctor.

Compared to one radiologist, there was a reduction of 1.2% in false positives, when a mammogram is incorrectly diagnosed as abnormal. There was also a reduction of 2.7% in false negatives, where cancer is missed. This was a research study, and as yet the AI system has not been let loose in the clinic. Even when it is, at least one radiologist would remain in charge of diagnosis. But AI could largely do away with the need for dual reading of mammograms by two doctors, easing pressure on their workload, say, researchers.

Prof Are Darzi, report co-author and director of the Cancer Research UK (CRUK) Imperial Centre, told the BBC: "This went far beyond my expectations. It will have a significant impact on improving the quality of reporting, and also free up radiologists to do even more important things."

Women aged between 50 and 70 are invited for breast screening every three years - those who are older can ask to be screened. The use of AI could eventually speed up diagnosis, as images can be analyzed within seconds by the computer algorithm.

Sara Hiom, director of cancer intelligence and early diagnosis at CRUK, told the BBC: "This is promising early research which suggests that in future it may be possible to make screening more accurate and efficient, which means less waiting and worrying for patients and better outcomes."

1.3 Scope of the project.

The following are the scopes of this project.

- To perform a literature review of existing models.
- To survey and find out various data pre-processing and training techniques.
- Comparing various algorithms' accuracies.
- Building a model for automatic detection and classification.
- Making a notification system.

1.4 Summary

Chapter 1 spoke about the background of breast cancer and understanding of the detection of breast cancer and how it affects us and especially doctors.

Over the course of the next chapter, literature survey is carried out by the team, in order to understand, the accuracy provided by different models and the need for deep learning was established.

Chapter 2. Literature Survey

A literature review is a comprehensive summary of previous research on a topic that was performed by the team on the topic of breast cancer classification and detection using the transfer learning technique. The literature review surveys scholarly articles, books, and other sources relevant to a particular area of research. Over the course of the month, multiple literature surveys were read and analyzed, of which the following 4 are the ones with the most accurate findings.

I] Paper 1

Comparative study of machine learning algorithms for breast cancer detection and diagnosis

Author(s): Dana Bazazeh, and Raed Shubair

Published in: 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016

It is not an easy one due to several uncertainties in detection using mammograms. Machine Learning (ML) techniques can be used to develop tools for physicians that can be used as an effective mechanism for early detection and diagnosis of breast cancer which will greatly enhance the survival rate of patients. This paper compares three of the most popular ML techniques commonly used for breast cancer detection and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN).

II] Paper 2

A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique

Author(s): Abeer Saber, Mohamed Sakr, Arabi Keshk, Osama M. Abo-Seida, and Huiling Chen.

Published in: IEEE Access

The aim of the paper was to address reducing training time, improving classificationperformance, and the problem of overfitting.

III] Paper 3

Evaluate the Malignancy of Nodules Using the 3D Deep Leaky Noisy-or Network

Author(s): Liao, Fangzhou & Liang, Ming & Li, Zhe & Hu, Xiaolin & Song

Published in: IEEE Transactions on Neural Networks and Learning Systems.

It addresses the low-performance problem faced in the 2D CNN Model. So, in order to overcome the problem faced by 2D CNN Models, the paper suggests 3D Models. A comparison between the two is provided. As conclusion, the 3D Model approach comes on top.

IV] Paper 4

A Review on Recent Progress in Thermal Imaging and Deep Learning Approaches for Breast Cancer Detection

Author(s): Roslidar Roslidar, Aulia Rahman, Rusdha Muharar, Muhammad Rizky Syahputra, Fitri Arnia, Maimun Syukri, Biswajeet Pradhan, and Khairul Munadi.

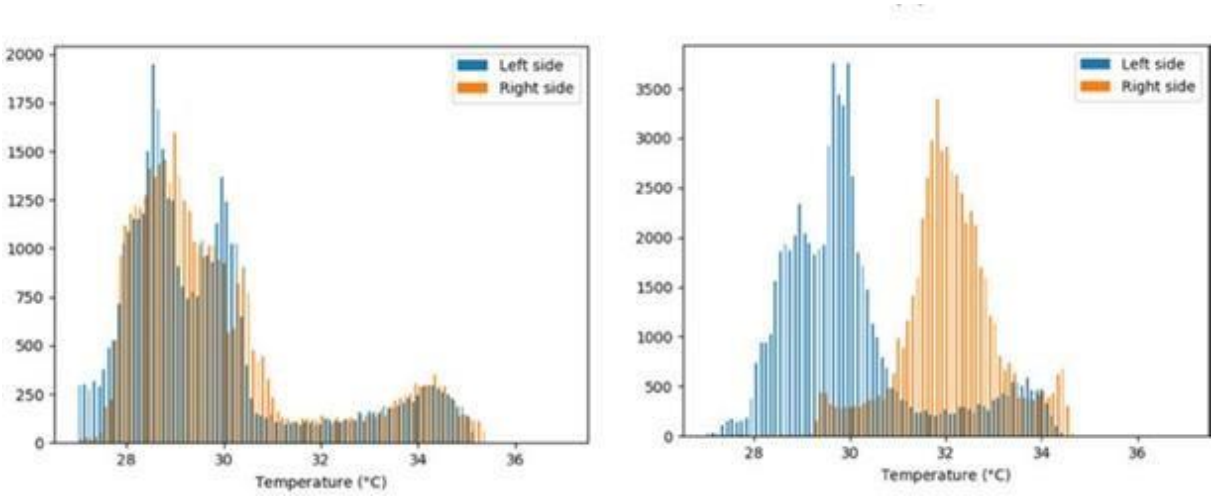
The following paper discusses thermography, a non-invasive and non-contact cancer screening method that can detect tumors at an early stage even under precancerous conditions by observing temperature distribution in both breasts. The thermograms obtained on thermography can be interpreted using deep learning models such as convolutional neural networks (CNNs). They covered most research related to the implementation of deep neural networks for breast thermogram classification and propose future research directions for developing representative datasets, feeding the segmented image, assigning a good kernel, and building a lightweight CNN model to improve CNN performance.

Conclusion.

In this chapter, understanding of different research papers and models performed by various authors takes place. The research provides information to with respect to different models where transfer learning technique can be applied to provide the best accuracy, without the trouble of overfitting the dataset.

Chapter 3. Project Design & Implementation

The first dataset has been taken from the InBreast and the second is from MIAS. Combining the dataset and performing data augmentation on it, provided us with a dataset with a larger number of values to work our model on.



Graph 1. First-order histogram-based features and co-occurrence matrix-based features

An image is assumed as the function f (x, y) of two space variables x and y. The value of the function is any discrete value of i within the range i != [0, L − 1].
For the first order histogram-based features, the intensity-level histogram is,

$$h(i) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \delta(f(x, y), i)$$

where δ is the Kronecker delta function,

$$\delta(j, i) = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$$

and p(i) is the probability value of image intensity,

$$p(i) = \frac{h(i)}{NM}.$$

From the research paper *Review on Recent progress in Thermal Imaging and Deep Learning Approaches for Breast Cancer Detection*[4] we can clearly understand the mathematical logic used by the authors

Histogram based features	**Formulae**
Mean	$\mu = \sum_{i-0}^{L-1} ip(i)$
Variance	$\sigma^2 = \sum_{i=0}^{L-1} (i-\mu)^2 p(i)$
Skewness	$\mu_3 = \sigma^{-3} \sum_{i=0}^{L-1} (i-\mu)^3 p(i)$
Kurtosis	$\mu_4 = \sigma^{-4} \sum_{i=0}^{L-1} (i-\mu)^4 p(i) - 3$
Energy	$E = \sum_{i=0}^{L-1} p^2(i)$

Table 1. Mathematical Functions used for Graph 1

Image pre-processing aims to improve image data/features by suppressing unwanted data and enhancing important image features to increase the performance of the NN model. Image pre-processing is crucial for NNs given that the success of the learning process depends on feature learning from input images. Generally, image pre-processing includes mean subtraction, normalization, PCA whitening, and local contrast normalization. The normalization of thebreast thermogram temperature matrix was undertaken previously. The study compared the classification accuracy of normalized datasets with the non-normalized breast thermogram temperature matrix. The result showed that the normalized input has a 16% better accuracy rate.

The segmented ROI of a breast thermogram shows a significant increase in temperature compared with that in the neighboring area.

Following are the pre-trained models that we have used to evaluate, compare and benchmark the performance of the model that we shall propose in the future.

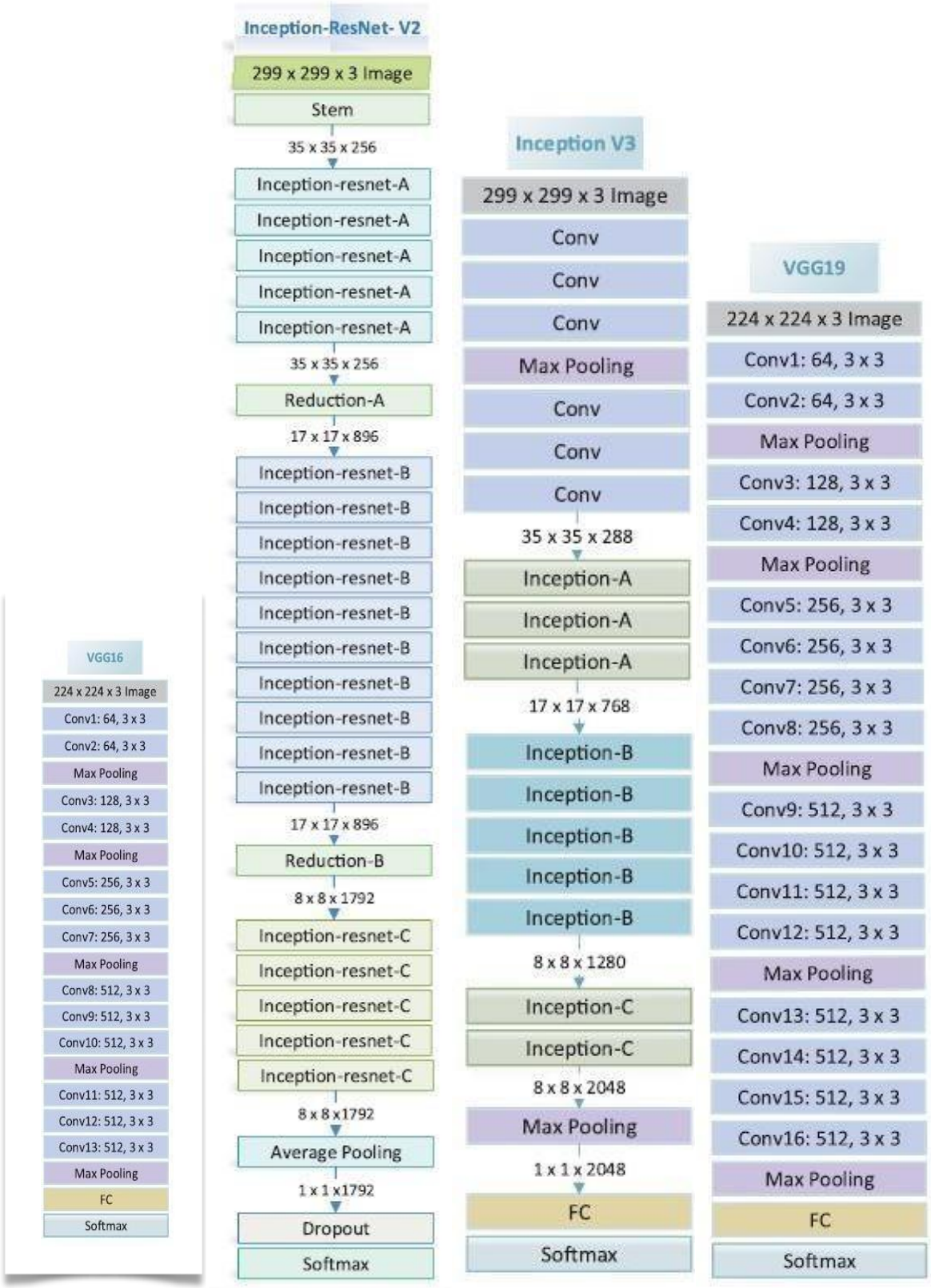


Figure 2. The layers in the pre-trained models used

VGG-16 is a convolutional neural network that is 16 layers deep. You can load a pretrained version of the network trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224.

VGG-19 is a convolutional neural network that is 19 layers deep. You can load a pretrained version of the network trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224.

Inception v3 is an image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. The model is the culmination of many ideas developed by multiple researchers over the years. It is based on the original paper: "Rethinking the Inception Architecture for Computer Vision" by Szegedy, et. al.

The model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concatenations, dropouts, and fully connected layers. Batch normalization is used extensively throughout the model and applied to activation inputs. Loss is computed using Softmax.

Inception-ResNet-v2 is a convolutional neural architecture that builds on the Inception family of architectures but incorporates residual connections (replacing the filter concatenation stage of the Inception architecture).

Of the above models, we obtained maximum accuracy from VGG-19.

Following are some samples from the Dataset.

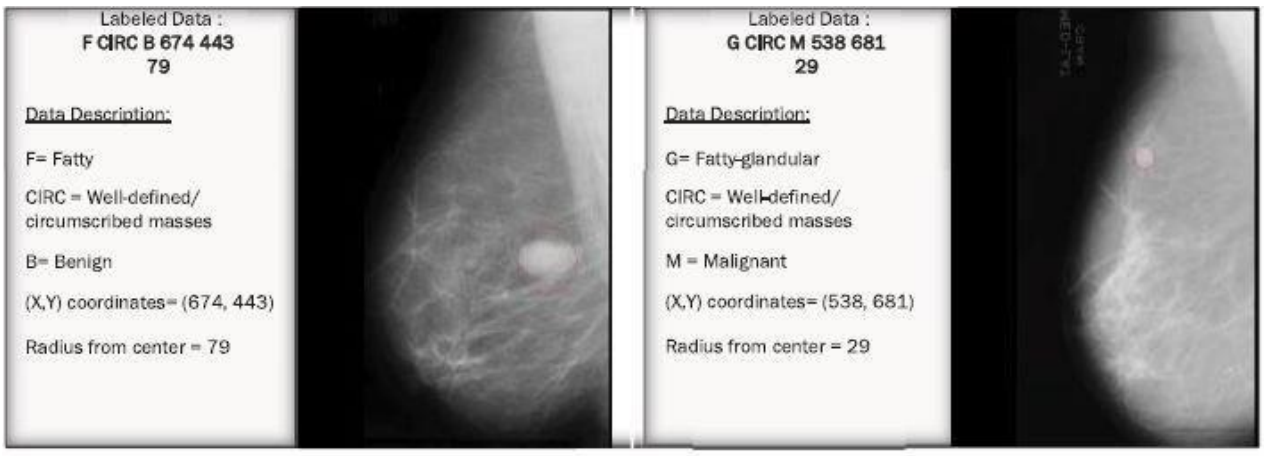


Figure 3. Sample from the dataset.

The process of segmentation and feature extraction is visualized below.

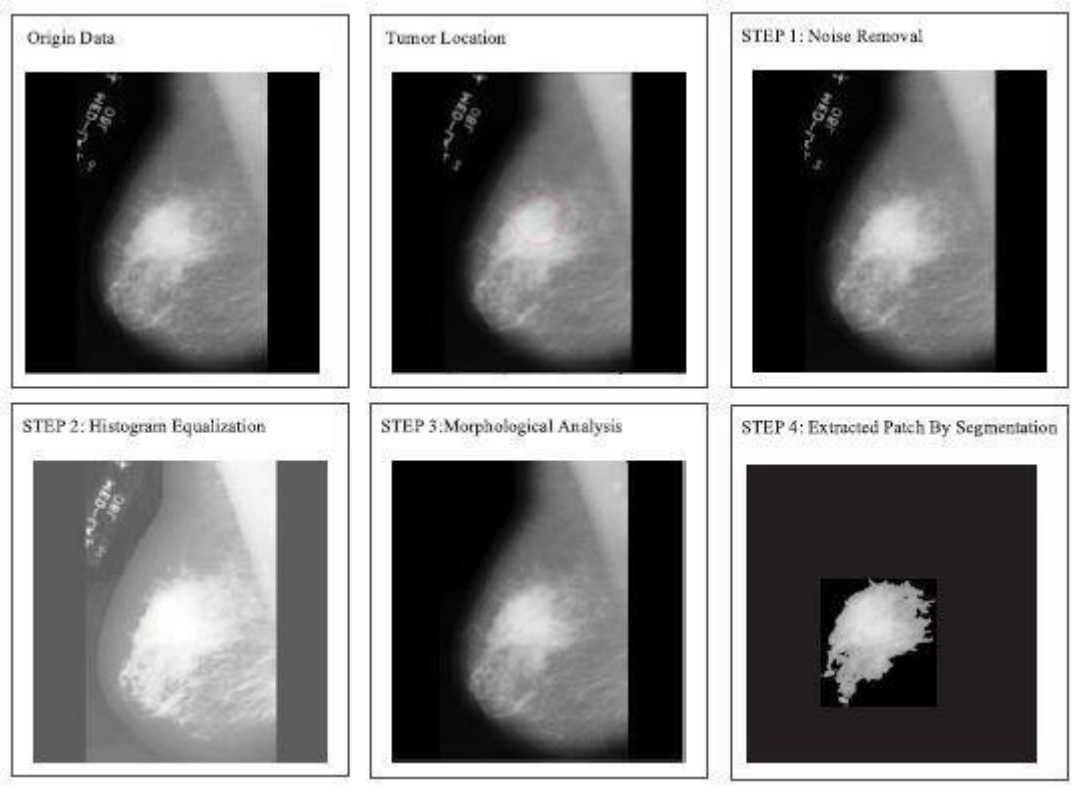


Figure 4. Segmentation and Feature Extraction

The table below shows the summary of the all the models used and the ones with highest accuracy.

Classifier	Result
Multilayer perceptron	Accuracy was about 75.23%
Convolution neural networks	Accuracy is 88.02% for static and 85% for dynamic protocol
VGG-16 VGG-19 Inception-v3 Inceptron-ResNet-v2	VGG-19 performed best with a balanced accuracy of 96.52%.

This chapter describes the models used and how each model is performing. The understanding of mathematic operation taken place, behind the data analysis is presented and theunderstanding of the ROI of breast cancer.

Over the next chapter, conclusions about the report and its understandings are put forward.

Chapter 4. Conclusion

On testing various models such as VGG-16, ResNet50, Inception V3, VGG19, and combinations using such models to provide the highest accuracy rate of 96.5251%.

While evaluating multilayer perceptron, a lot of false positives were provided resulting in an accuracy of 75.23%. Finally, the accuracy while using convolution neural networks, is 88.02% for static and 85% for dynamic protocol.

Understanding that Breast cancer is one the most common cancer among women, the early detection can lead to the chances of survival for a large number of women by receiving clinical treatment on time.

Concerning the literature review, we understood the importance of using convolution neural networks over just machine learning models to predict the outcome. The above data analysis also showed the region of interest andhow different cells are affected when one is detected with breast cancer.

Chapter 5. Future Works

For the following months of our project, handling various model analysis to testand receive a better accuracy if possible and trying and exploring more deeply to create a pingsystem. A ping system would be created for the doctor when a patient is detected with breast cancer.

Appendix

CNN - Convoluted Neural Network.

Histogram - Graphical representation of data points organized into user-specified ranges

Malignancy - the state or presence of a malignant tumors that is cancer cells.

Mammography - Technique using X-rays to diagnose and locate tumors of the breasts.

SVM - Support Vector Machine.

Transfer Learning - Application of knowledge gained from completing one task to help solve a different, but related, problem.

Thermogram - Photograph that shows differences in temperature. between different parts of an object.

Bibliography.

[1] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818560.

[2] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk and H. Chen, "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," in IEEE Access, vol. 9, pp. 71194-71209, 2021, doi: 10.1109/ACCESS.2021.3079204.

[3] Liao, Fangzhou & Liang, Ming & Li, Zhe & Hu, Xiaolin & Song, Sen. (2017). Evaluate the Malignancy of Nodules Using the 3-D Deep Leaky Noisy-OR Network. IEEE Transactions on Neural Networks and Learning Systems. PP. 10.1109/TNNLS.2019.2892409.

[4] R. Roslidar et al., "A Review on Recent Progress in Thermal Imaging and Deep Learning Approaches for Breast Cancer Detection," in IEEE Access, vol. 8, pp. 116176-116194, 2020, doi: 10.1109/ACCESS.2020.3004056.

[5] Giri P, Saravanakumar K. Breast Cancer Detection using Image Processing Techniques. Orient.J. Comp. Sci. and Technol;10(2)

[6] Gupta, Siddhartha & Sinha, Neha & Sudha, R & Babu, Challa. (2019). Breast Cancer Detection Using Image Processing Techniques. 1-6. 10.1109/i-PACT44901.2019.8960233.

[7] Chaitanya Varma and Omkar Sawant, “An Alternative Approach to Detect Breast Cancer using Digital Image Processing Techniques”, International Conference on Communication and Signal Processing, April 3-5, 2018

[8] Nadeem Tariq “Breast Cancer Detection using Artificial Neural Networks”, J Mol Biomark Diagn, 9:1, 2017

[9] Anuj Kumar Singh and Bhupendra Gupta “A Novel Approach for Breast Cancer Detection and Segmentation in a Mammogram” Eleventh International MultiConference onInformation Processing-2015

[10] Z. A. Abo-Eleneen and Gamil Abdel-Azim, “A Novel Statistical Approach for Detection of Suspicious Regions in Digital Mammogram”, Journal of the Egyptian Mathematical Society, vol. 21(2), pp. 162–168-2013

[11] Monika Sharma, R. B. Dubey, Sujata, S. K. Gupta “Feature Extraction of Mammograms”, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-3 Issue-5 September-2012

Acknowledgement.

We would like to express our deep gratitude to Dr. Anudeepa Kholapure, our research supervisor, for their patient guidance, enthusiastic encouragement, and useful critiques of this research work.

We would like to thank her for her advice and assistance in keeping my progress on schedule.

Finally, we wish to thank our parents for their support and encouragement throughout our study.