**Breast Cancer Classification using ResNet50 and comparing results pre-trained vs dataset user trained**



**Abstract.** Breast cancer develops when a malignant tumor forms as a result of some breast cells growing abnormally quickly. Breast cancer is discovered in one-eighth of all women at some point in their lives. Breast’s cancer, one of the top prevalent causes for discontinuation of life for women, can be avoided and, if detected early enough, can significantly increase the prognosis and chances of survival by enabling patients to get timely clinical therapy. Using the **Database for Screening Mammography (DDSM)** dataset, we developed an accurate method in this study for classifying mammograms or breast cancer images into labels for cancer and normal images. Two types of deep learning and transfer learning models were used, a pretrained resnet50 and a trained resnet50 and the test results were compared using a confusion matrix.

**Extended Abstract.** 2.3 million women globally received a breast cancer diagnosis in 2020, and 685,000 of them passed away. Breast cancer, the most frequent cancer in the world, will have been detected in 7.8 million people living by the end of 2020. Breast cancer causes more disability-adjusted life years (DALYs) to be lost for women globally than any other type of cancer. After adolescence, breast cancer affects women of all ages in every country on earth, though it becomes more common as people age.

Between the 1930s and 1970s, there was minimal change in the mortality of breast cancer. In nations with early detection programmes including various treatment methods to rule out invasive disease, improved survival rates started to emerge in the 1980s.

The proposed approach analyzed the accuracy value for pretrained model and compared it with a model we trained from scratch only to prove pretrained model (with an accuracy of 43.08300395256917)is not very accurate as compared to the other ResNet50 model, trained from scratch which has an accuracy of 80.23715415019763.

**Keywords:** Breast Cancer, ResNet50, DDSM, Deep Learning, Transfer

Learning

# 1 Introduction

Breast Cancer develops when a malignant tumor forms as a result of some breast cells growing abnormally quickly. One out of every 8 women unfortunately may receive a breast cancer diagnosis at some time in their life. This second most prevalent cause of mortality for women can be avoided, significantly improving the prognosis and chances of survival by enabling patients to obtain timely clinical therapy. However, if diagnosed in the early stages, this condition can be avoided. In India, one of the deadly diseases that is currently responsible for a large number of fatalities is breast cancer. Women are experiencing an increase in cancer cases due to dietary and lifestyle changes. In the entire world, it is indeed second biggest fatality claiming diagnosis for women. Deep learning principles are employed in this.

Some related work has been done on breast cancer prediction including models using Convolutional Neural Network (CNN), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) Algorithm, Random Forest, Naïve Bayes, Artificial Neural Network (ANN).

This paper however takes care of the problem using a different approach, using pre-trained and user trained resnet models and comparing the two for their accuracies.

Globally, the amount of women population that had to unfortunately give in to the fatal diagnosis, i.e. Breast cancer in 2020, was six hundred eighty-five thousand unfortunate lives, with 2.3 million receiving a diagnosis. By the end of 2020, breast cancer was a very common cancer in the world’s entirety, and would have been diagnosed in 7.8 million of the living women. It is also the reason for the most disability-adjusted life years (DALYs) to be lost for females worldwide. After puberty, breast cancer affects women of all ages in every country on earth; however, as women age, the likelihood of developing the disease increases.

Between the 1930s and 1970s, there were few changes in breast cancer mortality. Countries with early detection programmes that included various treatment modalities to rule out invasive disease started to see improved survival rates in the 1980s.

# 2 Related Work

Some related work has been done in the past to classify breast cancer at early stages.

The H&E dataset from Kaggle 162 was used to build the model. Automating mammography cancer detection to increase patient care has been a tough inspiring task. The dataset consists of both malignant and benign photos. The current study suggests a CNN strategy for automatically detecting this malignancy by analyzing IDC tissue areas in WSI. In this post, three different CNN architectures have been discussed along with accurate comparisons. The accuracy of the suggested system utilizing CNN Model 3 is 87°. The Model 3 consisting of five-layer CNN is appropriate considering this task, despite the fact that it is deeper than Models 1 and 2. A sizable data set of over 275,000 5050 pixel RGB image patches powers all designs.The accuracy of the recommender model increased by 8% in comparison to the results of the machine learning (ML) algorithm when we compared it to the algorithm results. The suggested methodology has been successful in obtaining precise results that can lessen human error in the diagnostic procedure and lower the cost of cancer diagnosis. The usage of a secondary database like Kaggle is the main drawback for the effortful research, and future research should focus on the original database for more accurate results about cancer identification. breast symbol.

7:3 division ratio was chosen to separate train and validation samples by The Vellore Institute of Technology's Ramik Rawal School of Computer Science and

Engineering (SCOPE), located at 632014 Gorbachev Road in Vellore, Tamil Nadu,

India.

Replicating published research findings is challenging because of inconsistent evaluation datasets to support decision systems in such tasks.

The data set provided by Lee and team, **The Curated Breast Imaging Subset of Database for Screening Mammography (**DDSM-CBIS) is formatted according to contemporary computer vision data sample collections and consists of unzipped image samples cooked by skilled medical personnels, visualization of normal and cancer area separation by plotted encovering lines, and mammographic identification for training data to be used. The samples collection includes 891 mass cases and 753 calcification instances, giving it a size that can be used to analyze mammography decision support systems.

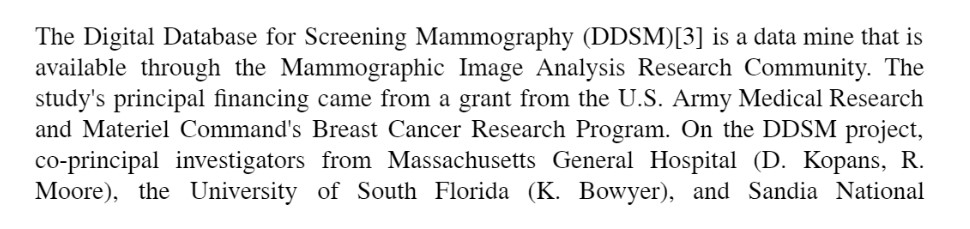
For each of its 2620 cases, the DDSM contains a wealth of data. While some information is scarce, others can be challenging to find. By revising the ROI separations and collection with rearrangement of the metadata into better readable way, they have overcome the challenge.

The detection of breast cancer has become much easier thanks to the development of approaches for computer vision and deep learning (DL). In this study, a novel digital mammogram-based breast cancer diagnosis model known as OMLTS-DLCN (Optimal Multi-Level Thresholding-based Segmentation with DL Enabled Capsule Network) is presented. The model (OMLTSDLCN) is used to get rid of the unusable disturbance present in the mammograph dataset images.

In order to automatically classify breast cancer using information from mammograms and ultrasounds—two distinct types of imaging—this study recommends utilizing a convolutional neural network (CNN) model that has depth in layers. The model has five learnable layers that are made up of just four CNN layers and one completely linked layer. The model makes it simpler to automatically extract noteworthy characteristics from the photographs because there are less configurable factors. The so produced model gives better results than the most recent methods, according to extensive simulation results on the MIAS dataset, along with DDSM accompanied by IN breast mammography datasets as well as the BUS\_1 along with BUS\_2 ultrasound

datasets.

# 3 Dataset



Laboratories collaborated (P. Kegelmeyer). More examples from the Washington University of Medicine were provided by Peter E. Shile, MD, Assistant Professor of Radiology and Internal Medicine. Other partners include Sacred Heart Hospital, ISMD, Inc., the Dept. of Medical Engineering and Radiology in University of Medicine at Wake Forest School. The database's main goal was to be heavily utilized for research by the academic community in order to assist creation of intelligent

logical lines of code that would facilitate detection.

There are 2,620 reports of mammography tests in the database and includes instances with certified pathology data for normal, benign, and cancerous conditions. The DDSM is a helpful tool and resource in the creation and for the purpose of validation of systems responsible for decision making, due to the size of the database and ground truth checking. The dataset is in jpeg format, in the form of black and white images and is available at the following web-site in the form of different labelled volumes each containing set of labelled images.

For our project, we have created a database using a few volumes of the data, labelled as normal and cancer samples .All the images are black and white in scale. The data samples labelled as cancer contain markings which denote where the cancer cells can be found in the mammography sample. The curated dataset used contains 3,596 jpeg File samples in the cancer data folder and 2,728 jpeg File samples in the normal label data folder. Further the data has been divided into train set for training, test set for testing and valid set for validation in 6:2:2 splitting partitions ratio. After the splitting, the training set contains 2,157 Files for the cancer subset and 1,636 Files for the normal subset. The testing set contains 720 Files for cancer subset and 547 Files for normal subset. The validation set contains 719 Files for cancer subset and 545 Files for normal subset. Further, we used Image Data Generator to cope up with the limitation of low data availability for model training. All the images have been resized to a common value of 224x224 pixels.

The DDSM outlines, as shown in the aforementioned figures, only offer an approximate area and not a specific mass boundary box. The separation algorithm applied at CBIS-DDSM was created to ensure precise separation of the bulk from the tissue in its immediate surroundings. Only masses were segmented in this analysis; no calcifications were included.

# 4 Proposed Approach

The process flow of the proposed method used in the project to classify the data image samples into cancer and normal is shown in a descriptive way below :

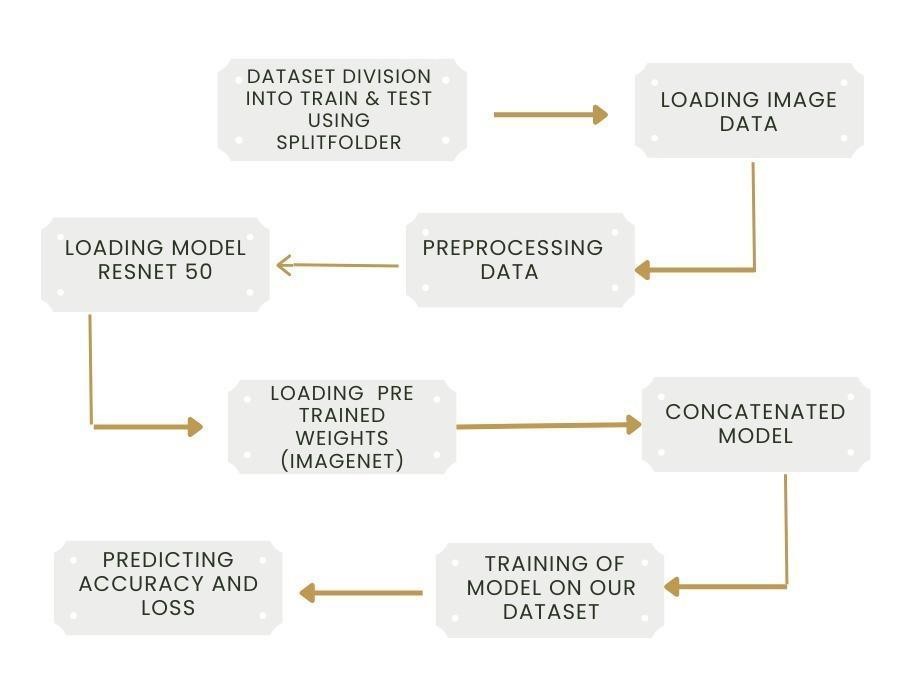


Figure 3 : proposed algorithm flowchart

## IMPLEMENTATION

**Libraries**

The use of *Google Colaboratory* made it a convenience to do this project. The packages and libraries imported during running the python code in the colab file are as shown below:

***Keras:*** Keras is an API designed library of python not for machines but made for human beings. This library is used to make the computation effective and efficient in the way as it helps in the backend of the neural networks created in this project.

***NumPy:*** NumPy is a scientific library of python under which computations like manipulations, calculations, random number generation fields are occupied. It helped in computing the dimensions of a particular image.

***TensorFlow:*** TensorFlow itself is a big library which is represented as the backbone of the deep learning projects. TensorFlow is the biggest module in the project and libraries or modules like *keras , image, config* are a part of it.

***Matplotlib:*** NumPy is the numerical mathematics extension of Matplotlib, a graph making plotting library resource in the programming language of python. It provides the object-oriented API so that charts may be included into applications. wxPython, Qt, GTK, or other general-purpose GUI toolkits are examples.

***Pandas:*** Pandas is a module which is used for data analysis and it helps in representing multidimensional arrays and working with time series and mathematical tables.

## Components

**Convolution Neural Networks (CNN):** The Deep Learning**.** Algorithms ingest a picture as input, extract features (understandable weights and biases) to**.** different**.**products / features of the picture and also being able to locate one from another. The pre-processing needed in a**.** ConvNet is**.** dipped as analogized to other algorithms used for classification. **.** For in ancient method**.**filters were made from hand, with a lot of training. ConvNet has the power and abilities to understand these features.

The design of a ConvNet is similar to that of the connected neuronal arrangement as seen in the normal human’s brain and was outstanding organization ofthe Visual Cortex. Particularly, every neuron only reacts to stimuli in a bounded region of the visual area called the Receptive Field.

A combination of such areas overrules tocoverthe full visual area. CNN networks were made to locate image information to an output variable. They have estimated so effectively by which they were the goto method for any type of divination problem including image information as an input.

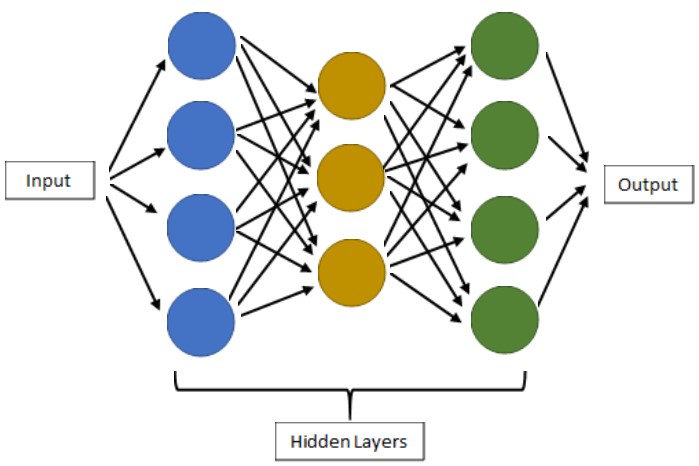


Figure 4. Diagrammatic Representation of CNN

## 4.1 Loading the Dataset

The DDSM dataset publicly available on eng.usf.edu website was downloaded. Few libraries were downloaded for each of the classifications, cancer and normal. The dataset for mammograms was loaded into google drive for ease of access and was divided into different folder categories for further categorization into training, testing and validation datasets. The dataset samples were divided to the 3 subsets (train, testing and validation) considering ratio of 6 : 2 : 2 , using a random state for division . Splitfolders functionality from split\_folder library was used for implementation in the local python environment.

## 4.2 Processing the Data

The dataset had to be cleaned for any empty paths which were to be found. Eventually these were detected and removed. A data frame was created that contained the final paths to all the image data samples stored, to be used for different purposes in the project such as model training and validation.

The train, test and validation directories were set up and Image Data Generator was used for image processing. Image samples were applied to properties such as shear\_range, zoom\_range, horizontal\_flip and randomization to generate more samples with unique yet diverse features in the sample set for training and testing of the model.

The batch size was set as 512 so that the process can utilize the full computing power of the system and flow\_from\_directory to ensure a steady supply of samples to the generator.

## 4.3 The Model

The Model used in the project is ResNet50. ResNet50 is a CNN architecture which contains 50 layers of depth in its structure, with 48 layers as the convolution layers and other two layers are termed as pooling layers.

We took two approaches to create a classification model. We used a pretrained ResNet50 model trained over ImageNet data. Also we trained ResNet50 from scratch to obtain different results and compare the two.

In the proposed model ‘ImageNet’ is used for weights and getting pooling on average condition. After getting the result of the model, it was found out that trainable parameters came out to be 23 million+ and non-trainable as 53K+. This model was for training datasets.

ResNet50 served as a base model for our proposed project. We added a

GlobalAveragePooling2D layer, followed by a dense layer using relu for activation and a dense layer for prediction classification, activated using softmax.

Adam optimizer was used in model compilation and categorical cross-entropy has been used to track the loss in functioning model output.

# 5. Experiment Design and Results

The experiment was designed to classify mammogram samples into cancer and normal case labels, and get the best possible results using the best possible model for the same.

As other studies have been done over the years, using CNN, KNN, SVM, Random

Forest, Naïve Bayes, ANN etc.

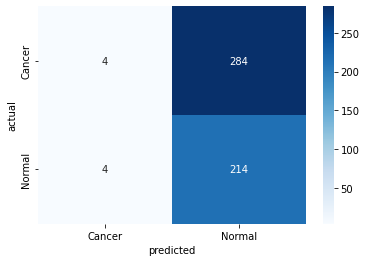
This paper takes care of the problem using a different approach, using pre-trained and trained resnet models and comparing the two for their accuracies.

The pretrained data ResNet50 architecture gives a not so accurate output result as compared to the other scenario.

In the pre-trained model we added a pooling layer over the ResNet50 and as expected got a very low test accuracy of ~43% (43.08300395256917).

Following is the heat map confusion matrix for the performed actions.

Figure5:Confusion matrix of the values predicted by a pre-trained model.



The ResNet50 was used again and was trained over our curated DDSM dataset.

The model was trained for 55 epochs and took around 6 fours for execution. The training accuracy for the model was found to be 94% during execution.

The model gave a test accuracy of ~78% to ~80% (80.23715415019763), way more than the pretrained model.Thus increasing the learning and the output accuracy for the trained model.

Following is the heat map confusion matrix and the training accuracy-loss graph for the performed actions.

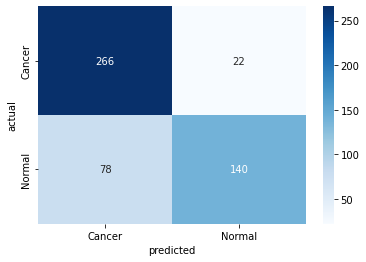


Figure 6: Confusion matrix of the values predicted by a model trained over ddsm.

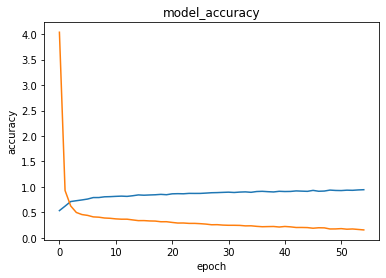


Figure 7: loss and accuracy vs epoch graph for model training

# 6 Discussion and Limitations

In this research, we presented a straightforward and practical approach to the classification of mammograms or breast cancer images using ResNet50 architecture trained over the DDSM dataset. Strong data augmentation and deep convolutional features taken from publically accessible CNNs ResNet50 pre-trained on ImageNet weights were employed to boost the classifier's resilience. We analyzed the accuracy value for pretrained model and compared it with a model we trained from scratch only to prove pretrained model(with an accuracy of 43.08300395256917 ) are not very accurate or ready to use as per differently defined use case situations , compared to the user trained ResNet50 model showed a test accuracy of 80.23715415019763.

The limitations of our model include the fact that the data available for model training was very less. There were around 3000 images in the training set, which in terms of deep learning, can be termed as a low amount of data.

Secondly, ResNet50 is a very deep and complex model which takes a heavy toll on the system. Since images are being used, the amount of RAM allotted and the computation power of the system can be considered as a limitation due to which we couldn't train the model for greater epochs, which could have given better results.

Altogether, understanding the data can be a tricky limitation as well, since there is no way to confirm the class label for an image just by looking at it. There aren't enough resources available to deal with the posed problem. There are very few image datasets available that scratch the domain of breast cancer mammography, along with being up to date with the current times.

Moreover, It is presently designed only for breast cancer and can be expanded further to cover the whole human body as a full scan for cancerous cells.

# 7 Conclusion and Future Work

From medical professionals to cancer patients, this way of early prognosis will help millions around the globe and save lives. It can be further worked upon by expanding it over other diagnostics, like lung cancer, brain tumors, etc.

Our report concludes with the presentation of a model for detection of breast cancer using images as input to work upon , utilizing transfer learning with Resnet50 as the base model and builds over it , training over 3000+ image data samples obtained from DDSM dataset resource , such that it gives a training and testing accuracy of 94 and 81 respectively. However, some effort can still be put into improving the model's accuracy for better outcomes, which should be tackled in future work dedicated to the current project.

Availability of more data will help further improve the results. Compressing image sizes will help us fit more in the dedicated system resources available. More architectures can be used, such as Inception model, VGG16, Exception model,

ResNext model, comparing the results to find a better fit for the problem.

Multiple datasets can be used together, such as MIAS and DDSM, for better feature extraction and learning of the model.

# References

1. Yue Zhang, Jiandong Ye, Yihao Luo. “Breast Cancer Image Classification on WSIwith Spatial Correlations.” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2019, no. 1219-1223, 2019, [https://www.semanticscholar.org/paper/Breast-Cancer-Image-Classification-o n-WSI-with-Ye-Luo/841b940f9e9e2995f3e9c110ec7186ba890a597e](https://www.semanticscholar.org/paper/Breast-Cancer-Image-Classification-on-WSI-with-Ye-Luo/841b940f9e9e2995f3e9c110ec7186ba890a597e).
2. Bairagi, Anupam Kumar. “Boosting Breast Cancer Detection Using ConvolutionalNeural Network.” *Journal of Healthcare Engineering*, vol. 2021, 2021, p. 11, https://doi.org/10.1155/2021/5528622.
3. F., Hoogi, A, Lee, R., Gimenez. “A curated mammography data set for use incomputer-aided detection and diagnosis research.” *Sci Data*, vol. 4, no.

170177 (2017), 2017, p. 19, https://doi.org/10.1038/sdata.2017.177.

1. J. Avanija& S. Neelakandan, T. Kavitha, Paul P. Mathai C. Karthikeyan, M.

Ashok, Rachna Kohar. “Deep Learning Based Capsule Neural Network

Model for Breast Cancer Diagnosis Using Mammogram Images.” *Interdisciplinary Sciences: Computational Life Sciences volume*, vol. 14, no.

113-129, 2022, https://doi.org/10.1007/s12539-021-00467-y.

1. rawal, ramik. “BREAST CANCER PREDICTION USING MACHINELEARNING:.” *JETIR*, vol. 2020, no. 7, 2020, p. 24, https://www.researchgate.net/publication/341508593\_BREAST\_CANCER\_ PREDICTION\_USING\_MACHINE\_LEARNING.
2. A. Adebanjo, Q. A. Al-Haija. “Breast Cancer Diagnosis in Histopathological

Images Using ResNet-50 Convolutional Neural Network.” *IEEE*

*International IOT, Electronics and Mechatronics Conference*

*(IEMTRONICS)*, vol. 2020, no. 1-7, 2020, https://www.semanticscholar.org/paper/Breast-Cancer-Diagnosis-in-Histopat hological-Images-Al-Haija-Adebanjo/58b5b31fbf5e0748b3ea91eeaae566b2e b00ce17.

1. <https://keras.io/>
2. <https://www.tensorflow.org/>
3. <https://pandas.pydata.org/>
4. <http://www.eng.usf.edu/cvprg/mammography/database.html>