Automatic detection of invasive ductal carcinoma with Convolutional Neural Networks

Nakshtra Popli

Electrical and Electronic Engineering Department

Dr. Akhilesh Das Gupta Institute of Technology and Management

New Delhi, India

Shubham Popli

Computer Science Department

NorthCap University

Gurgaon, India

*Abstract*— A profound learning device for recognizing and imagining intrusive ductal carcinoma (IDC) tissue areas in bosom malignant growth entire slide pictures (WSI) is characterized in this article (BCa).Deep learning techniques incorporate computational displaying of the taking in measure and are gain from-information strategies. To get familiar with a progressive part-based portrayal, the CNN is prepared on countless picture patches (tissue locales) from WSI. Utilizing various models like VGG16, ResNet50, EfficientNetB0, and DenseNet121, the model helps us in accomplishing the best presentation.

**Keywords:** breast cancer, neural network, cancer

# Introduction

Breast cancer is the most common cancer in women, and the most common form of breast cancer is invasive ductal carcinoma (IDC). Automated approaches can be used to save time and reduce error when defining and categorising breast cancer subtypes, which is a crucial clinical role.

The most common subtype of all breast cancers is Invasive Ductal Carcinoma (IDC). Pathologists usually concentrate on the regions of a whole mount sample that contain the IDC when assigning an aggressiveness score. As a consequence, delineating the exact regions of IDC within a whole mount slide is a standard pre-processing phase for automatic aggressiveness grading.

In this Python project, we'll build a classifier that will be trained on 80% of a breast cancer histology image dataset. We'll hold 10% of the data for validation purposes. We'll build a CNN (Convolutional Neural Network) called CancerNet and train it on our images using Keras. After that, we'll create a confusion matrix to evaluate the model's results.

IDC stands for Invasive Ductal Carcinoma, a form of breast cancer that begins in a milk duct and spreads to fibrous or fatty breast tissue beyond the duct. It is the most common type of breast cancer, accounting for 80% of all diagnoses. Histology is the analysis of tissue structure at the microscopic level.

# LITERATURE SURVEY

Invasive breast cancer spreads to the surrounding area from two original sites: milk ducts or lobules. This accounts for about 70% of all breast cancer cases. [1,2].

The pathologist's first move in diagnosing breast cancer is finding tumour cells in a histologic segment [1]. The whole procedure will take a long time if performed manually by a pathologist, so it should be automated. [3][4]

An automated and repeatable procedure for detecting invasive breast cancer on tissue slides may theoretically cut the time it takes to diagnose a breast case in half and eliminate some of the inter- and intra-observer variability. [5] [6]

The process of digitising tissue slides is known as digital pathology, and it makes the whole process more effective, thereby increasing the overall quality of routine diagnostic pathology workflow. [nine]

Identification of mitoses, tubules, cores, and lymphocytes, malignancy scoring, and the relationship of quantitative histologic picture highlights and sub-atomic highlights of bosom disease forcefulness have all as of late been stretched out to issues in bosom malignancy pathology.

Recognition of histologic image features that forecast breast cancer aggressiveness, recognition of histologic image features that predict breast cancer aggressivenessThese previous approaches have typically limited their analysis to only small portions of tissue or tissue microarrays (TMAs) as opposed to larger whole slide images.[8]

With the help of neural networks we can analyse the entire digital pathology image.

A neural network is made up of layers of artificial neurons that communicate with each other through connections. In recent years, it has been demonstrated that neural network models with thousands of neurons organised in many layers perform exceptionally well in computer vision and pattern recognition tasks..[9][10][11][12]

Using a Convolutional network classifier, we present a classification method for detecting the presence and extent of invasive breast cancer on whole slide digitised pathology images in this study. [13][14][15]

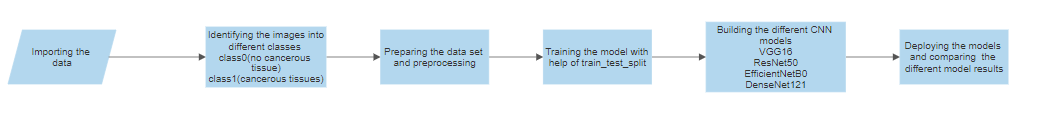
The aim of this study was to quantify the accuracy and robustness of a deep learning-based machine classifier for detecting the extent of invasive breast cancer on digitised whole slide pictures.

Description of Dataset

162 whole mount slide images of Breast Cancer (BCa) specimens were scanned at 40x in the initial dataset. 277,524 50 x 50 patches were collected from that (198,738 IDC negative and 78,786 IDC positive). uxXyYclassC.png — > example 10253idx5x1351y1101class0.png uxXyYclassC.png uxXyYclassC.png uxXyYclassC.png uxXyYclassC.png uxXyYclassC.png uxXyYclassC.png uxXyY Where u is the patient ID (10253idx5), X is the patch's x-coordinate, Y is the patch's y-coordinate, and C is the patch's class, with 0 representing non-IDC and 1 representing IDC.

# Proposed Methodology

Our project structure is going to be in the following way as described in the flow chart:



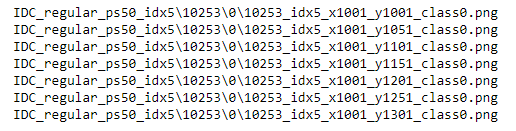
1. Importing the data:

As you have seen the Description of the dataset, our data is 3GB which is huge.

It comprises of 250+ different folders each of which is named after individual patient IDs

Each patient ID folder further has two folders ‘0’ and ‘1’ for patients not having breast cancer and for patients having breast cancer, respectively

To import such a huge data we have used glob library which recursively gives us the file names of all the images



We save the images of class 1 in class1 array

We save the images of class 2 in class2 array



1. Pre-processing the data:

As we have seen above, Class 0 has 198738 images and Class 1 has 78786 images, we need to have size of class 0 and class 1 to be same so that we don’t mess up our accuracy metric.

So we randomly sample class 0 to have same size as class 1



We reshape our images to be 50\*50 size as CNN accepts constant size input

Then we concatenate our data into single array and shuffle it for randomisation



When training the data, lambda layer is added to the model to divide the image pixel value by 255 as this is uint8 image which has value from 0 i.e black to 255 i.e white.

1. Saving the data:

We split our data into train and test data

80% data is train data

20% data is testing data

We have our X\_train, y\_train, X\_val, y\_val into .npy file so that later we can directly start our execution of code from this point without doing any pre-processing again and again

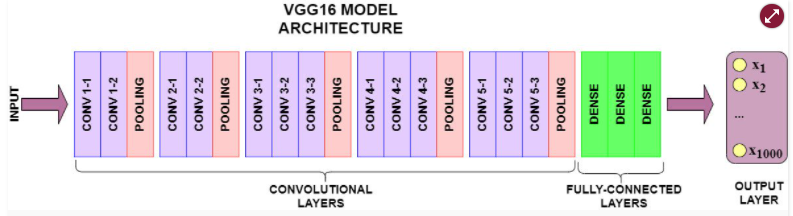
# Classification:

For classification we are using four models which have proven themselves at Google’s imagenet competition

We will be using transfer learning to train our models Our weights are going to be same as those of imagenet

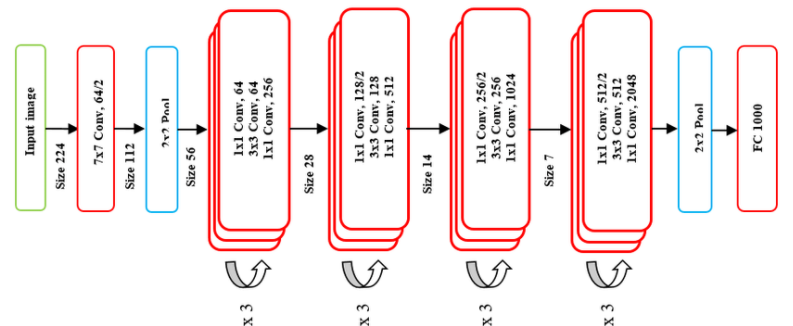
Last 5 layers will be trained by us.

* VGG16

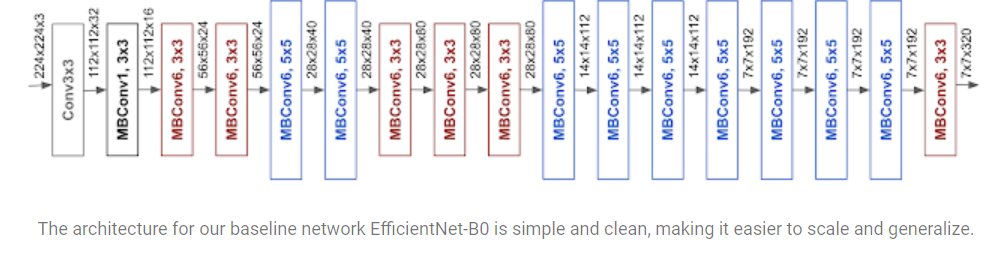


* ResNet50

Each of the five stages of the ResNet-50 model has its own convolution and identity block. Every convolution block has three convolution layers, and each identity block has three convolution layers. The ResNet-50 has over 23 million trainable parameters.

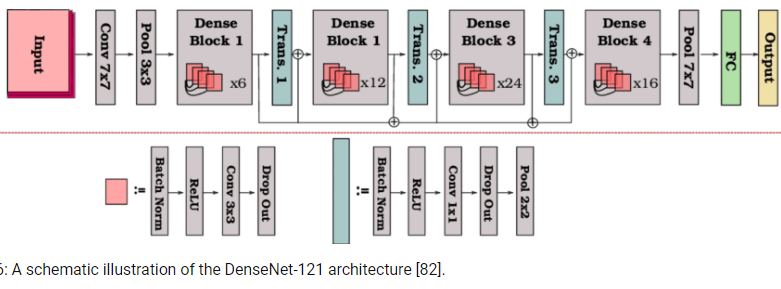


* EfficientNetB0

EfficientNet is a convolutional neural network architecture and scaling approach that uses a compound coefficient to scale all depth/width/resolution dimensions uniformly. The EfficientNet scaling approach uniformly scales network distance, depth, and resolution with a collection of defined scaling coefficients, unlike traditional practise, which scales these factors arbitrary.

* DenseNet121

DenseNet (Dense Convolutional Network) is an organization design that centers around developing profound learning organizations while likewise making them more successful to prepare by utilizing more limited associations between layers. DenseNet is a convolutional neural organization in which each layer is associated with any remaining layers further in the organization; for instance, the main layer is associated with the second, third, fourth, etc, and the subsequent layer is associated with the third, fourth, fifth, etc. This is accomplished to guarantee that everything of information can stream between the organization's layers.



1. Performance Metrics:

The metric we are using to evaluate our Models are:

* Accuracy:

Plotting the loss or accuracy vs epochs graph for both the training and validation sets will give you the exact amount you need to train the model.

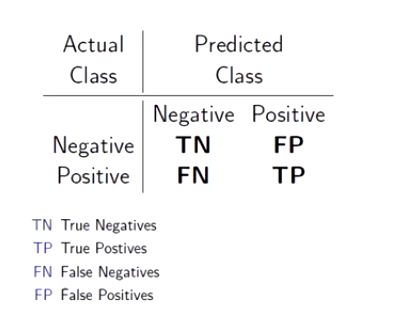
The validation-set loss increases after the early stopping condition, but the training set value continues to decrease. Both training and validation accuracy must be declining in an accurate model.

So our exact epoch number is which epoch value corresponds to the early stopping value.

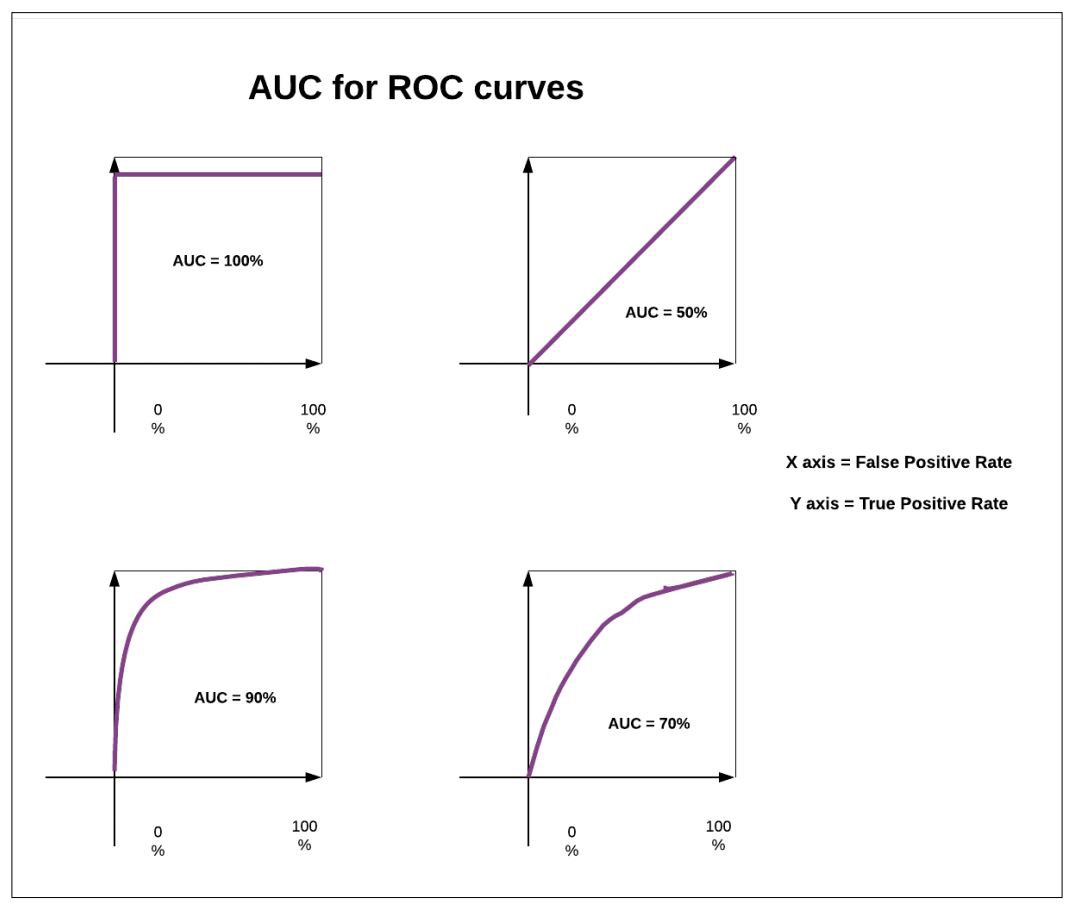
* Precision:

In pattern recognition, information retrieval, and classification, precision (also known as positive predictive value) is the fraction of suitable instances among the retrieved instances (machine learning).Recall:

while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.



* AUC:



V. Experimental Analysis

From the metrics which we have used so we can clearly see that VGG16 works the best among all the other models on this dataset

As you can see, the validation-set loss increases after the early stopping condition, while the training set value continues to decrease. Both preparation and validation accuracy must be diminishing in an accurate model.

So our precise epoch number is which epoch value corresponds to the early stopping value.

From the Metrics table as shown below in FigA that the accuracy metrics is highest in the VGG16 model. And thus comparing to the other metrics VGG16 has significantly outperformed the other models. The AUC for the classification is the highest in VGG16 which help us conclude that it is a better model overall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Classification Accuracy | Classification Precision | Classification Recall | Classification AUC |
| ResNet50 | 87.07% | 87.07% | 87.07% | 93.53% |
| VGG16 | 92.90 % | 92.90 % | 92.90 % | 97.47% |
| EfficientNetB0 | 84.68% | 84.68% | 84.68% | 84.88% |
| DenseNet121 | 89.90% | 89.90% | 89.90% | 95.57% |

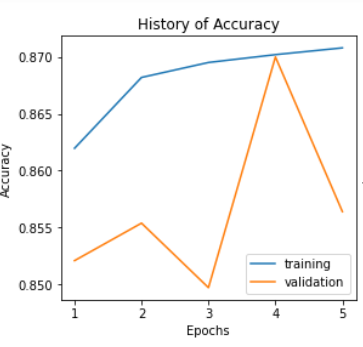


Fig1.ResNet50

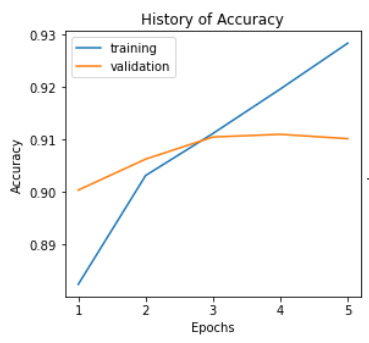


Fig.2 VGG16

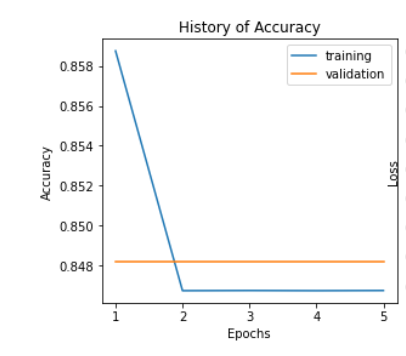


Fig3.EfficientNetB0

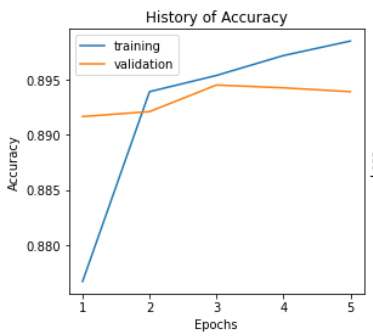


Fig4.DenseNet121

VI. Conclusion

Automated detection of invasive ductal malignant neoplastic disease may be a challenging and significant drawback in the diagnosis of carcinoma. Right and repeatable IDC detection is critical because it is often the first step in the process.

BCa analysis and care most of past histopathology neoplasm location research has tended to this.

By consolidating different kinds of handmade alternatives with AI calculations, we had the option to take care of the issue. Accordingly, we introduced a totally new profound learning framework for machine-driven location of IDC districts in WSI of BCa histopathology.

One of the additional charming outcomes was that the misclassified tissue districts was generally inferable from pathologists' absence of nitty gritty comments instead of mistakes in our proposed cycle. Our methodology's most prominent component is its reproducibility in various concealed WSI results, which is very like abstract paired manual comments from an expert pathologist offering quantitative help for its choices.

##### VII. References

1. Dillon, D. A., Guidi, A. J. & Schnitt, S. J. Pathology of invasive breast cancer. In Harris, J. R., Lippman, M. E., Morrow, M. & Osborne, C. K. (eds) *Diseases of the Breast* chap. Chapter 28, 374–407 4th edition edn (Lippincott Williams & Wilkins, 2010).
2. DeSantis, C., Siegel, R., Bandi, P. & Jemal, A. Breast cancer statistics, 2011. *CA: A Cancer Journal for Clinicians* 61, 408–418 (2011).
3. Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 19, 403–410 (1991).
4. Frierson, H. F. et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. *American journal of clinical pathology* 103, 195–8 (1995).
5. van Baardwijk, A. et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. *International Journal of Radiation Oncology Biology Physics* 68, 771–778 (2007).
6. Weaver, D. L. et al. Comparison of pathologist-detected and automated computer-assisted image analysis detected sentinel lymph node micrometastases in breast cancer. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc* 16, 1159–63 (2003).
7. Madabhushi, A. Digital pathology image analysis: opportunities and challenges. *Imaging In Medicine* 1, 7–10 (2009).
8. Basavanhally, A. et al. Multi-Field-of-View Framework for Distinguishing Tumor Grade in ER+ Breast Cancer From Entire Histopathology Slides. *IEEE transactions on biomedical engineering* 60, 2089–2099 (2013).
9. Donahue, J. et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference in Machine Learning (ICML)* (2014).
10. Hinton, G. & Srivastava, N. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).
11. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* 25, 1106–1114 (2012).
12. Le, Q. et al. Building high-level features using large scale unsupervised learning. In *International Conference in Machine Learning* (2012).
13. Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1798–1828 (2013).
14. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 2278–2324 (1998).
15. LeCun, Y. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* 253–256 (2010).