Breast Cancer Histology Image Classification Using Deep Neural Networks

Tuhin Das
Dept. of Computer Science
Techno India
College of Technology
Kolkata, India
tuhindas221b@gmail.com

Shrutina Agarwal
Dept. of Computer Science
Techno India
College of Technology
Kolkata, India
shrutina.agarwal10@gmail.com

Gitesh Jain
Dept. of Computer Science
Institute of Engineering
& Management
Kolkata, India
giteshjain844@gmail.com

Amitrajit Bose
Dept. of Computer Science
University of Engineering
& Management
Kolkata, India
amitrajitbose9@gmail.com

Sivangi Tandon
Dept. of Computer Science
University of Engineering
& Management
Kolkata, India
sivangitandon@gmail.com

Shivam
Dept. of Information Technology
Indian Institute of
& Information Technology
Allahabad, India
shivam.1996.in@ieee.org

***Abstract*—Breast Cancer is ranked the number one cancer among Indian females with a rate of 25.8 per 100,000 women and a mortality rate of 12.7 per 100,000 women. It is also one of the most common causes of cancer worldwide. There have been many biological and non-biological research in the past and present to be able to prematurely detect breast cancer. In this paper, we have taken an approach using Deep Learning to try to predict whether a breast cancer tumor is non-cancerous or is benign, in situ or invasive stage from high quality histopathological images. The experimental results shows higher test accuracy than the most state of the art methods in this field.**

***Index Terms*—Breast Cancer Detection, Densenet, Histopathological Images, Computer Vision**

I. Introduction

The cell of the body maintains a cycle of regeneration processes. The balanced growth and death rate of the cells normally maintain the natural working mechanism of the body, but this is not always the case. Sometimes an abnormal situation occurs, where a few cells may start growing aberrantly. This abnormal growth of cells creates cancer, which can start from any part of the body and be distributed to any other part. Different types of cancer can be formed in human body, among them breast cancer creates a serious health concern. Due to the anatomy of the human body, women are more vulnerable to breast cancer than men. Among the different reasons for breast cancer, age, family history, breast density, obesity, and alcohol intake are reasons for breast cancer.

Statistics reveal that in the recent past the situation has become worse. As a case study, shows the breast cancer situation in Australia for the last 12 years [2]. In 2007, the number of new cases for breast cancer was 12775, while the expected number of new cancer patients in 2018 will be 18235 [2]. Statistics show that, in the last decade, the number of new cancer disease patients increased every year at an alarming rate.

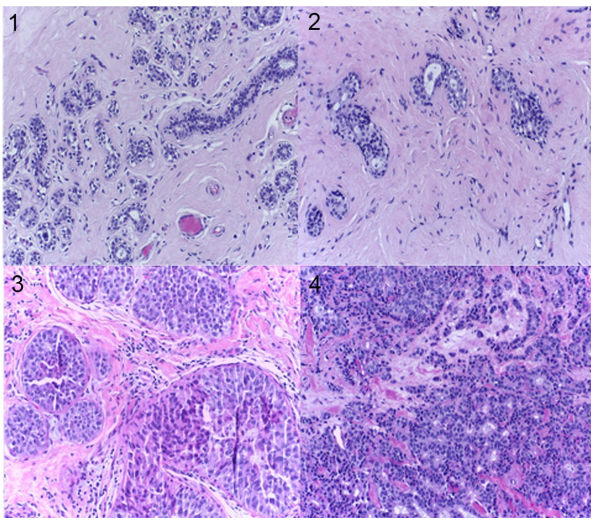

Fig. 1: Examples of Histopathological images. Clockwise from left - (1) Normal (2) Benign (3) In Situ (4) Invasive Carcinoma

Breast cancer tumors can be categorized into two broad scenarios. (i) Benign (Noncancerous) (ii) Malignant (Cancerous)

Identification of the normal, benign, and malignant tissues is a very important step for further treatment of cancer. Based on the penetration of the skin and

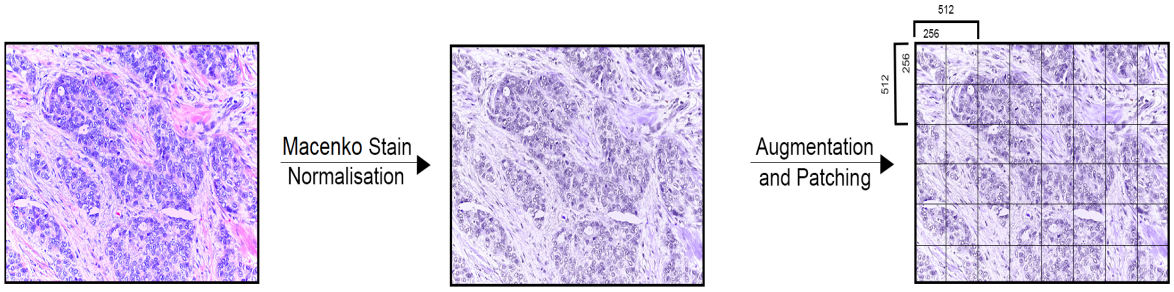

Fig. 2: Pre Processing Steps: Stain Normalization and Patching of Histopathological Images

damage of the tissue medical photography techniques can be classified into two groups. (i) Noninvasive. (ii) Invasive.

Histopathology refers to the microscopic examination of tissue in order to study the manifestations of disease. Specifically, in clinical medicine, histopathology refers to the examination of a biopsy or surgical specimen by a pathologist, after the specimen has been processed and histological sections have been placed onto glass slides [11]. Histopathology slides, on the other hand, provide a more comprehensive view of disease and its effect on tissues, since the preparation process preserves the underlying tissue architecture. As such, some disease characteristics, e.g., lymphocytic infiltration of cancer, may be deduced only from a histopathology image. The diagnosis derived from a histopathology image remains the gold standard in diagnosing a substantial number of diseases including almost all types of cancer [10].

The convolutional neural network architecture used to train the datasets is Dense Neural Network (DNN) [1]. DenseNet is being used instead of other convolutional networks because of many advantages of it:

- DenseNet combines features by concatenating them instead of summing them. [1]
- DenseNet requires fewer parameters to make accurate predictions. [1]
- The layers of DenseNet architecture are very narrow. [1]
- Besides better parameter efficiency, one big advantage of DenseNets is that they are easy to train due to their improved flow of gradients and information throughout the network. [1]
- Each layer has direct access to the gradients from the original input signal and the loss function, which leads to an implicit deep supervision. This is massively helpful in training very deep network architectures. [1]
- Further, we also observe that dense connections have a regularizing effect, which reduces overfitting on tasks with smaller training set sizes. [1]

II. Dataset

The image dataset contains high-resolution (2048 x 1536 pixels), uncompressed, and annotated images from the Bioimaging 2015 breast histology classification challenge [4]. The images in the dataset are then digitized by using the same acquisition conditions, with magnification of 200 and pixel size of (0:42 microm x 0:42 microm). Each image is labeled with one of four classes: i) in situ carcinoma, ii) invasive carcinoma, iii) normal tissue and iv) benign lesion. According to the dataset website, the labeling of the dataset was performed by two pathologists. Without specifying the area of interest for the classification, they only provided a diagnostic from the image contents. In the cases where the pathologists disagreed with each other, those images where discarded. The aim of the project is to output an automatic classification for any histopathological image provided as input. The dataset is made of an extended training set of 249 images, and a separate test set of 20 images. In these datasets, the four classes are balanced. The images were selected so that the pathology classification can be objectively determined from the image contents. We mark as ”extended” dataset a set of 16 additional images provided which have increased ambiguity. The training and test datasets are publicly available at [4].

III. Methodology

For the preprocessing, we used Macenko Stain Normalization [3] technique for getting normalized histopathological images. The normalized images in the training set are used to create an augmented dataset after the stain normalization process. The network may suffer from the problem of overfitting as the dataset has a low number of samples when compared to other convolutional neural network classification problems. The complexity and dimension of the dataset is increased by dividing the images into patches.The dataset is further improved by data augmentation through mirroring and patch rotation. Physicians can study the histopathological images for breast cancer from different orientation without changing the diagnosis. This is what makes the patch rotation possible. This also helps increase the size

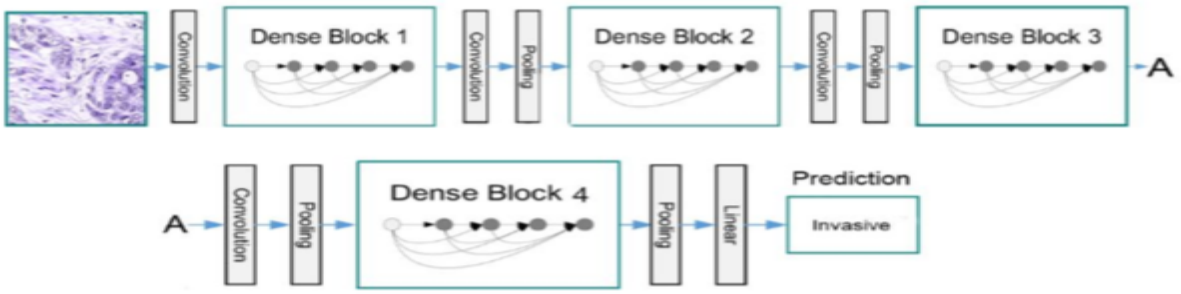

Fig. 3: The architecture pipeline after the image patches are generated. In the figure, the convolution layer and the pooling layer together the Transition layer as mentioned in Section III.A

of the dataset for the model without harming its quality.

First, the image is divided in patches of 512 x 512 pixels size, with 50 percent overlap. The average value is subtracted from the red, green and blue value individually for the Patch Normalization process. Eight different patches are formed from each patch by combining k*90 degrees rotations, with k values being 0, 1, 2, 3, and vertical reflections. The original 250 images is thus converted into 70000 images with each of the patches having the same class label as the original image.

A. *DenseNet*

We are using the DenseNet-121 architecture [1] for our prediction model. We resize our 512 x 512 pixel images to 224 x 224 after which we pass them to a convolutional layer which gives it a size of 112 x 112 pixels. The images are then transferred through the pooling layer which resizes the images to a size of 56 x 56. This architecture has four Dense blocks and subsequently three transition layers. To further improve model compactness, we reduce the number of feature-maps at transition layers. The Dense blocks applies the filters of different sizes in each block to the image for training and reduces the output image size by half each time. The final dense block outputs a 7 x 7 pixel image which is vectorized into a 1 x 1 pixel output by the classification layer. The weights are assigned to the nodes by using Glorot [6] method and Batch Normalization [7], to prevent overfitting, Dropout [5]technique was used by dropping out units in random with a dropout rate of 0.8.

IV. Training

A virtual machine with the following specifications was created on Google Cloud Platform and used to train the model:

i 9GB RAM
ii 220GB SSD
iii 8 Core Processor
iv NVIDIA Tesla K80 Graphics Card

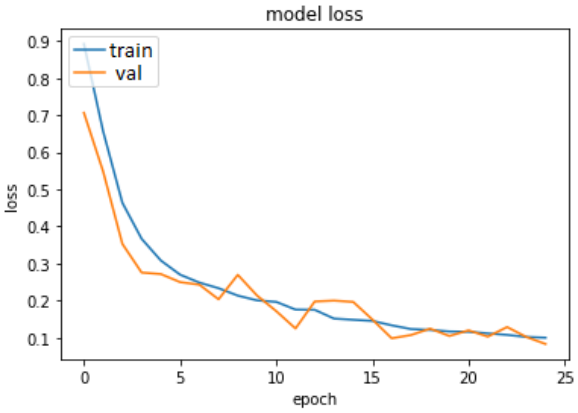

Fig. 4: Model Loss

The dataset was divided into training, validation and test set. 68880 images were used as training set, 5600 images as validation set and 5600 images as test set.A batch size of 32 was used. We trained a densenet model with 121 layers as mentioned in [1] using Adams Optimizer [12] for 25 epochs. Total params: 4,226,224 Trainable params: 4,180,330 Non-trainable params: 45,894 A dropout rate of 0.8 was used to avoid overfitting. The hyper-parameters were tuned according to [6] and [7]. As we see in Figure 4, both training and validation losses are continuously decreasing with each epoch. After 25 epochs we get a train loss of 0.0852 and validation loss of 0.0822. The model accuracy, shown in Figure 5, also increases for both training and validation set with every passing epoch and reaches a maximum of 0.974 and 0.973 respectively on the 25th epoch.

V. Results

We assess the performance of the model on the test set comprising of 1400 images from each class. Our model achieves a remarkable accuracy of 89.50%. From Table-1 we see that our model achieves an AUROC score of almost 1 in the three classes : Invasive , In Situ and Normal and very high value of .98 in the class Benign.

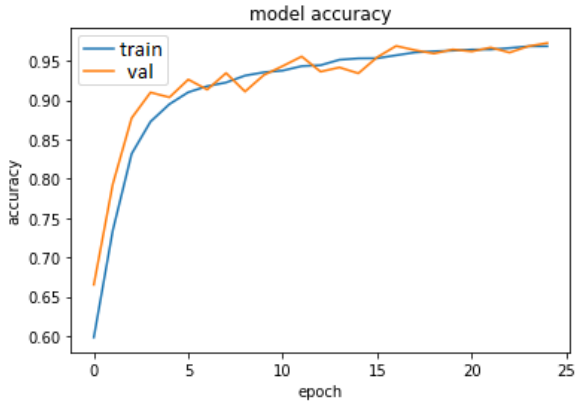

Fig. 5: Model Loss

	Precision	Recall	F1-Score	Support	AUROC
Benign	0.98	0.65	0.78	1400	0.9845
In Situ	0.84	0.99	0.91	1400	0.9982
Invasive	0.98	1.00	0.99	1400	1.00
Normal	0.87	1.00	0.93	1400	0.9989

Table 1. The precision, recall, F1 and AUROC scores

VI. Comparison with Other Work

In the work of Araujo, Teresa et al [9], a CNN and SVM combination was used to classify breast cancer histology images. The accuracy achieved was 84% by the CNN+SVM model [9]. The dataset used is the same dataset as the one in the current work. In the current work, we get a 89.5% accuracy in classifying the images by using DenseNet, which is a significant improvement.

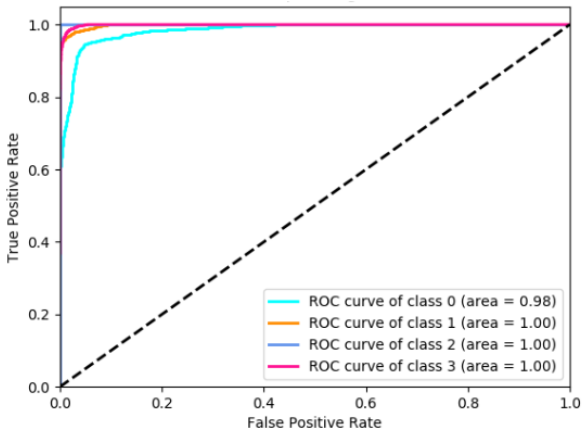

Fig. 6: The ROC Curves of the different classes. Class 0 - Benign, Class 1 - In Situ, Class 2 - Invasive, Class 3 - Normal

In the work of Rakhlin et al [8], a Deep CNN approach was used to classify breast cancer histology images. The accuracy achieved was 87.22.6% by this process [8]. In the current work, we get a 89.5% accuracy on the classification, which is again an improvement.

VII. Conclusion

Here we aimed to predict the class of breast cancer from Macenko stain normalized images using a type of DNN (DenseNet). The major problem while using DNNs for classification is feature loss with increase in number of layers, and even the smallest section of the histopathological images contains immense information. Thus it gets difficult and at times meaningless to apply DNNs on medical imaging. But, here with the use of DenseNet we were able to overcome this problem.

Further patching and augmentation not only helped us to increase the size of the training set which was and an important issue to resolve if we wanted to avoid overfitting, but also let us focus on every minute details in the image, as we only focus on a very small part of the tissue imaging in every instance. Our results are better than all the state of the art approaches in this field, giving us a higher accuracy. In future we aim to apply this approach to various other types of medical images as well. If achieved the desired results, it can change our way of approaching medical reports.

References

[1] Gao Huang, Zhuang, Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, 2016. Densely Connected Convolutional Networks
[2] Breast Cancer Statistics in Australia.
[3] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, 2009, pp. 1107-1110. doi: 10.1109/ISBI.2009.5193250
[4] ICIAR 2015 Grand Challenge on Breast Cancer Histology Images.
[5] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), pp.1929-1958.
[6] Glorot, X. and Bengio, Y., 2010, March. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256).
[7] Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167
[8] Rakhlin, A., Shvets, A., Iglovikov, V. and Kalinin, A.A., 2018. Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. arXiv preprint arXiv:1802.00752
[9] Arajo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using Convolutional Neural Networks. Sapino A, ed. PLoS ONE. 2017;12(6):e0177544. doi:10.1371/journal.pone.0177544
[10] Gurcan MN, Boucheron L, Can A, Madabhushi A, Rajpoot N, Yener B. Histopathological Image Analysis: A Review. IEEE reviews in biomedical engineering. 2009;2:147-171. doi:10.1109/RBME.2009.2034865
[11] Wikipedia page on Histopathology
[12] Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]