Breast Cancer Detection Using Histopathological Images

Project Report submitted at
Computer Society India Symposium 2018, Bodh Gaya.

Group Members

Tuhin Das Shrutina Agarwal

Gitesh Jain Amitrajit Bose

Sivangi Tandon

Contents

1 Abstract

Breast cancer is one of the largest causes of women's death in the world today. Advance engineering of natural image classification techniques and Artificial Intelligence methods has largely been used for the breast-image classification task. The involvement of digital image classification allows the doctor and the physicians a second opinion, and it saves the doctors' and physicians' time. Despite the various publications on breast image classification, very few review papers are available which provide a detailed description of breast cancer image classification techniques, feature extraction and selection procedures, classification measuring parameterizations, and image classification findings. We have put a special emphasis on the Convolutional Neural Network (CNN) method for breast image classification. Along with the CNN method we have also described the involvement of the conventional Neural Network (NN), Logic Based classifiers such as the Random Forest (RF) algorithm, Support Vector Machines (SVM), Bayesian methods, and a few of the semisupervised and unsupervised methods which have been used for breast image classification.

2 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. 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. Statistics show that, in the last decade, the number of new cancer disease patients increased every year at an alarming rate. Breast cancer tumors can be categorized into two broad scenarios.

(i) Benign (Noncancerous). Benign cases are considered as noncancerous, that is, non-lifethreatening. But on a few occasions it could turn into a cancer status. An immune system known as "sac" normally segregates benign tumors from other cells and can be easily removed from the body.
(ii) Malignant (Cancerous). Malignant cancer starts from an abnormal cell growth and might rapidly spread or invade nearby tissue. Normally the nuclei of the malignant tissue are much bigger than in normal tissue, which can be life-threatening in future stages.

Cancer is always a life-threatening disease. Proper treatment of cancer saves people's lives. Identification of the normal, benign, and malignant tissues is a very important step for further treatment of cancer. For the identification of benign and malignant conditions, imaging of the targeted area of the body helps the doctor and the physician in further diagnosis. With the advanced modern photography techniques, the image of the targeted part of the body can be

captured more reliably. Based on the penetration of the skin and damage of the tissue medical photography techniques can be classified into two groups.

(i) Noninvasive. (a) Ultrasound: this photography technique uses similar techniques to SOund Navigation And Ranging (SONAR) which operates in the very-high-frequency domain and records the echos of that frequency, invented by Karl Theodore Dussik. An ultrasound image machine contains a Central Processing Unit (CPU), transducer, a display unit, and a few other peripheral devices. This device is capable of capturing both 2D and 3D images. Ultrasound techniques do not have any side-effects, with some exceptions like production of heat bubbles around the targeted tissue. (b) X-ray: X-rays utilize electromagnetic radiation, invented by Wilhelm Conrad Roentgen in 1895. The mammogram is a special kind of X-ray (low-dose) imaging technique which is used to capture a detailed image of the breast. X-rays sometimes increase the hydrogen peroxide level of the blood, which may cause cell damage. Sometimes X-rays may change the base of DNA. (c) Computer Aided Tomography (CAT): CAT, or in short CT imaging, is advanced engineering of X-ray imaging techniques, where the X-ray images are taken at different angles. The CT imaging technique was invented in 1970 and has been mostly used for three-dimensional imaging. (d) Magnetic Resonance Imaging (MRI): MRI is a noninvasive imaging technique which produces a 3D image of the body, invented by Professor Sir Peter Marsfield, and this method utilizes both a magnetic field as well as radio waves to capture the images. MRI techniques take longer to capture images, which may create discomfort for the user. Extra cautions need to be addressed to patients who may have implanted extra metal.

(ii) Invasive. (a) Histopathological images (biopsy imaging): histopathology is the microscopic investigation of a tissue. For histopathological investigation, a patient needs to go through a number of surgical steps. The photographs taken from the histopathological tissue provide histopathological images

2.1 Contribution

In our work, a densely connected CNN designed for the analysis of breast cancer Macenkov stained histology images is proposed. Unlike previous approaches we perform image-wise classification in four classes of medical relevance: i) normal tissue, ii) benign lesion, iii) in situ carcinoma and iv) invasive carcinoma.
For this, a new breast cancer image dataset is presented taken from BioImaging website. In addition, the proposed CNN architecture is designed to integrate information from multiple histological scales, including nuclei, nuclei organization and overall structure organization. By considering scale information, the CNN can also be used for patch-wise classification of whole-slide histology images. A data augmentation method is adopted to increase the number of cases in the training set.

3 Dataset

The image dataset is composed of high-resolution (2040 1536) pixels, uncompressed, and annotated images from the Bioimaging 2015 breast histology classification challenge. All the images are digitized with the same acquisition conditions, with magnification of 200 and pixel size of (0:42microm0:42microm). Each image is labeled with one of four classes: i) normal tissue, ii) benign lesion, iii) in situ carcinoma and iv) invasive carcinoma The labeling was performed by two pathologists, who only provided a diagnostic from the image contents, without specifying the area of interest for the classification. Cases of disagreement between specialists were discarded. The goal of the challenge is to provide an automatic classification of each input image. The dataset is composed 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. An additional test set of 16 images is provided with images of increased ambiguity, which we denote as "extended" dataset. The training and test datasets are publicly available at https://rdm.inesctec.pt/dataset/nis- 2017-003.

4 Preprocessing

We used Macenkov stain normalization technique for staining the images.

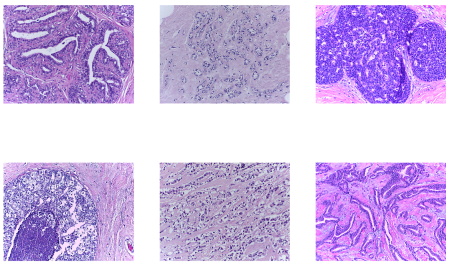


Figure 1: This is how the original image looks like

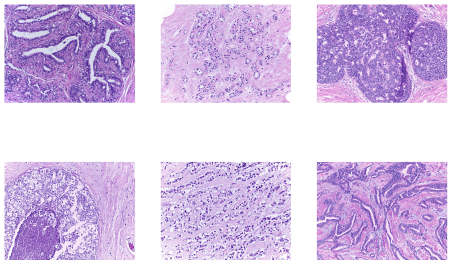


Figure 2: This is how the preprocessed image looks like

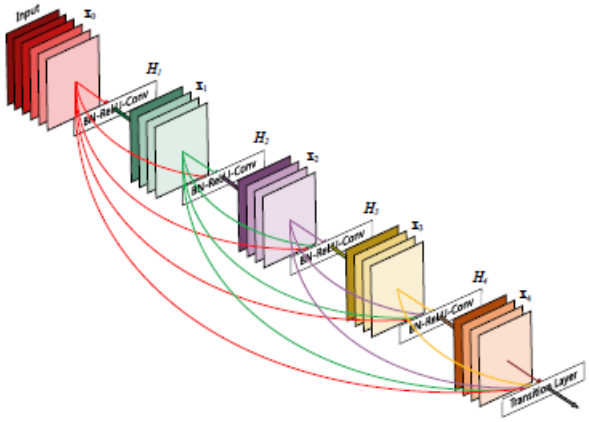
5 Augment Patch Classification

An augmented patch dataset is created from the normalized images in the training set. The used dataset has a low number of samples when compared to other CNN classification problems [18]. The network might thus be prone to overfit. Dividing images into patches allows to increase the dataset complexity and dimension. Data augmentation through patch rotation and mirroring further improves the dataset. This is possible because the studied problem is rotation invariant, i.e., physicians can study breast cancer histological images from different orientations without altering the diagnosis. Consequently, rotations and mirroring allow to increase the size of the dataset without deteriorating its quality. Patching and dataset augmentation have already been used successfully on

similar histological classification problems [19]. However, they have not been used for carcinoma classification.
First, the image is divided in patches of 512 512 pixels size, with 50 percent overlap. Some example patches are shown in Fig 1. Patch normalization is performed by subtracting the average value to the red, green and blue channels separately. Each patch is then transformed into eight different patches by combining k 90 degrees rotations, with k = 0, 1, 2, 3, and vertical reflections. This results in a total of 70000 different patches from the original 250 training images. Each of the patches is considered to have the same class label as the original image.

6 Working

Dense connectivity
To further improve the information flow between layers we propose a different connectivity pattern: we introduce direct connections from any layer to all subsequent layers. Consequently, the ‘th layer receives the feature-maps of all preceding layers, x0; : : : ; x‘1, as input: x‘ = H‘([x0; x1; : : : ; x‘1]); (1) where [x0; x1; : : : ; x‘1] refers to the concatenation of the feature-maps produced in layers 0; : : : ; ‘1. Because of its dense connectivity we refer to this network architecture as Dense Convolutional Network (DenseNet). For ease of implementation, we concatenate the multiple inputs of H‘() in eq. (1) into a single tensor.



A 5-layer dense block with a growth rate of k = 4. Each layer takes all preceding feature-maps as input.

Bottleneck layers
Although each layer only produces k output feature-maps, it typically has many more inputs. It has been noted in [36, 11] that a 11 convolution can be introduced as bottleneck layer before each 33 convolution to reduce the number of input feature-maps, and thus to improve computational efficiency. We find this design especially effective for DenseNet and we refer to our network with such a bottleneck layer, i.e., to the BN-ReLU-Conv(1 1)-BN-ReLU-Conv(33) version of H‘, as DenseNet-B. In our experiments, we let each 11 convolution produce 4k feature-maps.

Compression
To further improve model compactness, we can reduce the number of feature-maps at transition layers. If a dense block contains m feature-maps, we let the following transition layer generate bmc output featuremaps, where 0 ¡ 1 is referred to as the compression factor. When =1, the number of feature-maps across transition layers remains unchanged. We refer the DenseNet with ¡1 as DenseNet-C, and we set = 0:5 in our experiment. When both the bottleneck and transition layers with ¡ 1 are used, we refer to our model as DenseNet-BC.

Layers	Output Size	DenseNet-121
Convolution	112 × 112	
Pooling	56 × 56	
Dense Block (1)	56 × 56	[1 × 1 conv, 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	
	28 × 28	
Dense Block (2)	28 × 28	[1 × 1 conv, 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	
	14 × 14	
Dense Block (3)	14 × 14	[1 × 1 conv, 3 × 3 conv] × 24
Transition Layer (3)	14 × 14	
	7 × 7	
Dense Block (4)	7 × 7	[1 × 1 conv, 3 × 3 conv] × 16
Classification Layer	1 × 1	
	1000D fully-connected, softmax	

DenseNet architecture for ImageNet.Each "conv" layer shown in the table corresponds the sequence BN-ReLU-Conv

7 Training and Results

We trained a densenet model with 121 layers as mentioned in [Densely Connected Convolutional Networks: Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger] which gave a testing accuracy of 89.44 with just 25 epochs of training. Total params: 4,226,224 Trainable params: 4,180,330 Non-trainable params: 45,894 We used a dropout of 0.8 to avoid overfitting.

8 Bibliography

1. http://clipsrules.sourceforge.net/documentation/v630/ug.pd [Clips,Documentation].
2. Densely Connected Convolutional Networks [Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger]
3. M. Macenko et al., 'A method for normalizing histology slides for quantitative analysis', in 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009, pp. 1107–1110.

4. Classification of breast cancer histology images using Convolutional Neural Networks
5. Bioimaging Challenge 2015 Breast Histology Dataset
6. Breast Cancer Histopathological Database (BreakHis)