R Gayathri, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 10, Issue 7, (Series-V) July 2020, pp. 35-40

RESEARCH ARTICLE OPEN ACCESS

Characterization Breast Cancer Histology Images using Deep Learning

R Gayathri, R Latha
IV Semester M.Tech., Department of Computer Science and Engineering, The Oxford College of Engineering, Bangalore-560068
Assistant Professor, Department of Computer Science and Engineering, The Oxford College of Engineering, Bangalore-560068,India.

ABSTRACT:
The paper employs deep learning to classify breast cancer histopathological image into normal, benign and malignant subclasses *in situ* carcinoma and invasivecarcinoma categories. The classification is mainly based on cells' density, variability, and organization along with overall tissue structure and morphology. Smaller and larger patches of histological images are extracted that includes cell-level and tissue-level features. Here, Patches are screened by Clustering algorithm and CNN is used to select the discriminative patches. The proposed approach is applied to the multi-class classification of breast cancer histology images.It achieves initial test achieves of 95% accuracy and on the overall test,88.89% accuracy.

Keywords: Breast cancer histology images, CNN, image classification.

Date of Submission: 06-07-2020 Date of Acceptance: 21-07-2020

I. INTRODUCTION:

Breast cancer is the most common cancer and the secondmain cause of cancer for death in women, after lung cancer. Thechance of any woman dying from breast cancer is around 1 in37, or 2.7 percent [1]. The diagnosis from a histology image is the gold standardin diagnosing considerable types of cancer. Pathologistsanalyse the regularities of cell shapes, density, andtissue structures by examining a thin slice of tissue underan optical microscope and determine cancerous regionsand malignancy degree. Due to the complexity and diversityof histology images, the manual examination requiresabundant knowledge and experience of the pathologistsand is fairly time-consuming and error-prone [2]_[4].

II. RELATED WORK:

In the 2012 ImageNet image classification competition, the deep learning model AlexNet won the champion [5]. In study, Spanhol*et al.* [6], [7] constructed a datasetof 7909 breast cancer histology images named BreakHisacquired on 82 patients.Spanhol et al. used six different feature extractors to extract features from the image, and provided four classifiers for each feature extractor for final classification. The final correct rate was 80% to 85%. Bayramoglu et al. put together four different magnifications for uniform training and tested them separately at a single magnification [8]. They trained AlexNet based on the extraction of patches obtained randomly or by a sliding window mechanism from breast cancer images with multiplemagnifications and combined the patch probabilities withthree fusion rules for final classification.Wang et al. [9] used sampling patches to train a CNN to make patch-level predictions, then aggregated the results to create tumour probability heatmaps and made slide-level predictions. The methodology was tested on theCamelyon16 dataset including 400 WSIs [10]. In [11], context-aware stacked convolutionalneural networks for 3-class classification of breast WSIs werepresented.Bejnordi et al. used a CNN trained by high pixel resolution patches to extract cell-level features primarily, followed by a second CNN. Then, large input patches were used to train the stacked CNNs to learn both cellular information and global tissue structures.

III. PROPOSED SOLUTION:

In this paper, deep learning is employed to construct a CAD model, and the pathological images of breast cancer are divided into benign and malignant. In the work herein described, histology image classificationwas performed by processing several patches with fixedsize. Microscopically, cancer cells have distinguishing histologicalfeatures. Therefore, referring to the pathologists' diagnostic process, features related to cells and global tissue structures extractedfrom two kinds of patches with

different sizes will improvethe performance of the classification of breast cancer histologyimages into one of the 4 target classes. The labels of histologyimages for the classification task given by the pathologistsare based on the whole images. Larger size patchessampled from a histology image contain sufficient informationso that the image label can be used for the patches. However, cell-level patches extracted from high resolutionhistology images, especially ultra-high resolution WSIs, maynot contain sufficient diagnostic information. There existsome patches with large areas of fat cells and stroma, sparsebreast cells, and normal patches extracted from malignanthistology images. CNNs trained by these patches can't extractdiscriminative features. Consequently, we present a methodologyto automatically screen more discriminative patchesbased on clustering algorithm and convolutional neural network. Based on the above two aspects, the main objective ofthis paper is to propose a comprehensive and effective schemefor the multi-classification of breast histology images in orderto improve the diagnostic performance. To achieve this, the main contributions of our work canbe summarized as follow: (i) We propose a patch sampling strategy to extract two kinds of patches with different sizesto preserve essential information and contain cell-level andtissue-level features respectively., (ii) We design a patchselecting method to select more discriminative patches basedon CNN and K-means., (iii)We design a classification frameworkwhich extracts features from the patches using thefeature extractors and compute the final feature of each wholeimage for classification through a classifier.

Stain inconsistency of histology images, due to differences in color responses of slide digital scanners, will affect the performance of image analysis. As can be seenfrom Fig. 1, the images in the dataset have large stain variation.

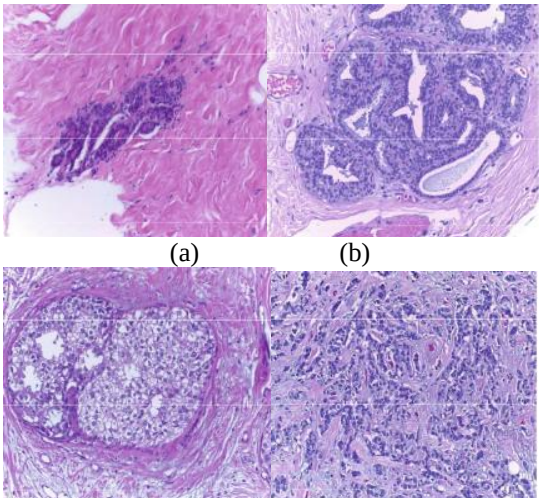

FIGURE 1. H&E stained images from each type, (a): normal tissue, (b): benign abnormality, (c): in situ carcinoma, and (d): invasive carcinoma.

To thisend, stain normalization is essential prior to other processes. Thereare various research for stain normalization in histology images[12], [13]. In this paper,we use a method proposed by Reinhard*et al.* [14], which transforms the RGB images to the decorrelated$l\alpha\beta$ color space, followed by computing themeans and standard deviations for each channel separately in $l\alpha\beta$ space and a set of linear transforms in order to match the color distribution of the source and target images, finally, converts the results back to RGB.

IV. DATASET:

This section is dedicated to introducing the dataset used inour work and pre-processing of images. The dataset is fromthe bioimaging 2015 breast histology classification challenge[15], composed of high-resolution (2048×1536 pixels)and H&E stained breast cancer histology images. The imageswere digitized with a magnification of 200x and pixel size of0:42μm×0:42μm . Two pathologists labelled images as normal, benign, *in situ* carcinoma or *invasive* carcinoma accordingto the predominant cancer type in each image, withoutspecifying the area of interest. Fig. 1 illustrates images fromeach class mentioned in the dataset.

This dataset composed of a training set of 249 images, an initial test set of 20 images and an extended testset of 16 images with increased

ambiguity is public clyavailable at https://rdm.inesctec.pt/dataset/nis-2017-003.The main goal of this paper is to propose an effectivescheme for the multi-class breast histology images
classification.

V. PROPOSED ARCHITECTURE:

The multi-classification scheme of breast histology images ispresented in this section. We introduce the overall frameworkat first, and then describe each process in detail.Fig. 2 illustrates the framework of our approach used for multi-class classification of breast histology images.

A. Framework:

Themain processes can be summarized as follow: (i) We extracttwo kinds of patches with different sizes by a sliding windowmechanism from breast cancer histology images to preserveessential information and contain cell-level and tissue-level
features, and then train two CNNs as feature extractorsrespectively. (ii) We split the small patches into multipleclusters using k-means clustering algorithm and select morediscriminative patches based on the network trained by small
patches to retrain the network. (iii) We extract features from the select smaller patches and larger patches using the featureextractors and compute the final feature of each whole imageto train a classifier for classification.

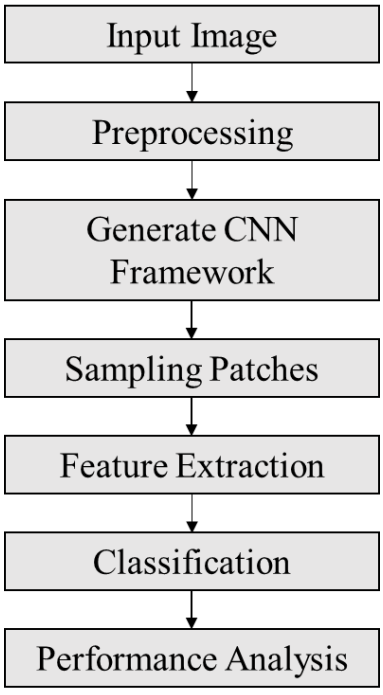

Figure 2. System Architecture

B.CNNTraining:

Our goal is to classify the breast histology image intofour classes: normal tissue, benign tissue, *in situ* carcinomaand *invasive* carcinoma. The performance of classificationis highly dependent on the information extracted from theimages. We use features related to breast cells and globaltissue structures to represent each whole image. Firstly, because the arrangement of cancer cells is extremely disorderedand the cancerous cells have atypia such as larger nucleiand inconsistent morphology, therefore, cell-level featuresincluding the nuclei information, such as shape and variability, as well as cells organization features like density and morphology, are used to diagnose whether cells are cancerous. The pixel size of the breast histology images in the dataset is0:42μm×0:42μm, and the radius of cells is between 3 and 11pixels approximately. Consequently, we extract small patchesof 128 ×128 pixels to contain cell-level features. Secondly,the structure of the diseased tissue may be atypical. *In situ*carcinoma is a carcinoma growth of minor grade precancerous,with no invasion of the surrounding tissue within a particular tissue compartment in themammary duct. Interestingly, *invasive* carcinoma does not confine itself to theinitial tissue compartment [16]. Therefore, tissue structuresinformation is essential to differentiate between *in situ* and*invasive* carcinomas. It is unpractical for CNNs to extractfeatures from a histology image with a large size directly. According to the size of images in the provided dataset, we extract patches of 128×128 pixels to contain the globaltissue structures information.We extract patches by a sliding window mechanism frombreast cancer histology images. The patches of 128×128 pixelsare small and focus on cell-related characteristics, therefore, we extract contiguous non-overlapping patches from thebreast histology images. In addition, we extract overlapping128×128 pixels patches with a 50% overlap to containcontinuous tissue morphology and structures information. Allextracted patches are given the same label as the correspondinghistology image.

C. FEATURE EXTRACTOR:

The histology images have different cell morphology, texture, tissue structures, and so on. The representation of complexfeatures is significant for the classification task. The handcraftedfeature extraction method needs abundant expertdomain knowledge, and it is labour-intensive and difficult toextract discriminative features. CNNs can directly extractrepresentative features from images, and have achievedremarkable results in various fields. ResNet50 [17] is used asfeature extractor in this

paper because it is a classical CNNand easy to train compared to other deeper models under thepremise of ensuring the extraction of usability features. The*deep residual learning* framework (ResNet) is proposedby He and Sun [18] to address the *degradation* ofdeep networks. Formally, the desired underlying mappingis denoted as $H(x)$, then the stacked nonlinear layers arefitted to another mapping of $F(x) := H(x) - x$ and the original mapping is rewritten as $F(x) + x$. The formula of$F(x) + x$ is implemented by feed-forward neural networkswith ``shortcut connections" which perform the *identity* mapping. For deeper nets, a *bottleneck* design whichuses a stack of 3 layers instead of 2 for each residual functionis proposed. The ResNet50 consists of 16 ``bottleneck"building blocks and takes as input a {3, 224, 224} RGB image.

The training of ResNet50 from scratch requires a largenumber of training images to avoid overfitting. However,because of the paucity of histology images in ourdataset, we adapt a transfer learning strategy[20], [21] anduse ResNet50 pre-trained on the ImageNet dataset[22].Weremove the top layer of the network and add a softmaxclassifier with 4 neurons, then, we resize the patches of128×128 pixels to 224×224 pixels forfine-tuning two modified networks as original feature extractorsand the trained networks are denoted as ResNet50-512and ResNet50-128 respectively. 2048-dimensional featuresof patches can be obtained from the g*lobalAveragePooling*layer of ResNet50.

For the multi-class classification of breast cancer histology images, the sampling strategy of two kinds of patches, the screening method of 128×128 pixels patches and feature extractors based on ResNet50 have been introduced above.

Then, we rescale the extracted patches of 128×128 pixels and selected patches of 128 × 128 pixels corresponding to each image in the training set, and feed them into the fine-tuned ResNet50-512 and ResNet50-cluster respectively to obtain the 2048-dimensional features group, which can represent the cells and tissue structures information of the image. In order to obtain the final feature of an image, we employ the P-norm pooling fusion method [19] and the formulation is as follows:

$$f_p(v)=\left(\frac{1}{N}\sum_{i=1}^{N} v_i^p\right)^{\frac{1}{p}}$$

Here, N represents the number of patches, vi denotes the 2048-dimensional feature of the i-th patch and P D 3 is used in our paper. At last, the image-wise features of histology images in the training set are used to train the SVM classifier for multi-class breast cancer histology classification.

VI. PERFORMANCE EVALUATION:

IMAGE-WISE CLASSIFICATION:

We use the normalized breast histology images in the test set to verify the approach proposed in this paper. The procedure of the experiment is as follows:

a) The sampling strategy introduced in Section V.B is used to extract contiguous non-overlapping patches of 128×128 pixels and patches of 128×128 pixels with 50% overlap from the test images.

b) The ResNet50-cluster fine-tuned by patches of 128×128 pixels in the selected clusters is sensitive to morediscriminative patches, therefore, we use the networkto predict the smaller patches and select patches withclassification probability higher than a set threshold.

c) We rescale the selected patches of 128 × 128 pixels correspondingto each test image to 224 × 224 pixels, and feed them into the fine-tuned ResNet50-512 and ResNet50-cluster respectively to obtain the 2048-dimensional featuresgroup.

d) We employ the 3-norm pooling method to compute the final feature of each image and make final classification by using SVM.

The patches of 128 × 128 pixels are predicted using the ResNet50-cluster,and the patches with classifier probability higher than

90% are retained. Four test images are classified into wrong categories, three of which belong to the extended test set, and the remaining one labelled as normal is classified as benign.

The calculation formulas are as follows:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots(2)$$

$$Precision=\frac{TP}{TP+FP}, Recall=\frac{TP}{TP+FN} \ldots\ldots(3)$$

$$F=\frac{2\times Precison\times Recall}{Precision+Recall} \ldots\ldots\ldots\ldots\ldots(4)$$

$$Macro\text{-}F=\frac{1}{N}\sum_{i=1}^{n} Fi \ldots\ldots\ldots\ldots\ldots\ldots(5)$$

Here, TP (true positive) is the number of positives cases that are classifies as positive. Analogously, TN, FN and FP represent the numbers of true negatives, false negatives and false

positives respectively. The recall represents the percentage of positive samples that are correctly classified, which is more clinically relevant.Macro-F, also known as macro-averaging, is used to evaluate the performance of multi-classification

globally and is computed by first computing the F-scores for the ncategories then averaging these per-category scores to compute the global means.

The image-wise accuracy of the initial test set and overall test set is 95% and 88.89% respectively. According to the confusion matrix, precision, recall and F-score of each class can be obtained respectively, as shown in Table 2. The value of

macro-F calculated according to formula (5) is 89.14%.

Results	Normal	Benign	InSitu	Invasive
Precision	0.875	0.75	1.0	1.0
recall	0.78	1.0	0.89	0.89
F-Score	0.825	0.857	0.942	0.942

Table 2: The Performance of the proposed model.

VII. RESULTS AND DISCUSSION:

In our work, we extract smaller patches of 128 × 128 pixels from the breast histology images to contain cell-level and tissue-level features, then, we screen discriminative 128 × 128 pixels patches based on clustering algorithm and CNN. Through comparative experiments, it is proved that the method proposed in this paper can effectively improve the performance of multi-classification of breast histology images.We contrast the aftereffects of our methodology and the benchmark strategy proposed in [23] (CNNCSVM) and the near result is appeared in Table 11. Araújo et al. utilized the equivalent dataset as us and extricated patches of 512 × 512 pixels. Theyutilized their very own CNN planned and accomplished a best exactness of 77.8% of multi-classification with enlarged dataset. It very well may be seen that our methodology has a considerable improvement in exactness and review contrasted and the benchmark conspire, particularly in the classification of benevolent what's more, in situ carcinoma pictures.

VIII. CONCLUSION:

In this paper, we propose an effective method to classify the H&E stained breast histology images into four classes: normal tissue, benign lesion, in-situ carcinoma and invasive carcinoma. Due to the atypia of cancerous cells andthe difference in tissue morphology and structures between in situ carcinoma and invasive carcinoma, we extract patches of 128 × 128 pixels from the histology images to contain different levels features.

REFERENCES:

[1]. N. Christian. (2018). *What You Need to Know About Breast Cancer*. [Online]. Available: https://www.medicalnewstoday.com/articles/37136.php

[2]. M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, ``Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147_171, 2009.

[3]. L. He, L. R. Long, S. Antani, and G. Thoma, ``Computer assisted diagnosis in histolopathology," *Sequence Genome Anal., Methods Appl.*, vol. 3,pp. 271_287, 2010.

[4]. L. He, L. R. Long, S. Antani, and G. R. Thoma, ``Histology image analysis for carcinoma detection and grading," *Comput. Methods Programs Biomed.*, vol. 107, no. 3, pp. 538_556, 2012.

[5]. A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems. 2012.

[6]. F. A. Spanhol, L. S. Oliveira, L. Heutte, and C. Petitjean, ``A dataset for breast cancer histopathological image classi_cation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455_1462, Jul. 2016.

[7]. F. A. Spanhol, L. S. Oliveira, L. Heutte, and C. Petitjean, ``Breast cancer histopathological image classi_cation using convolutional neural networks,"in*Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016,pp. 2560_2567.

[8]. N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," 2016 23rd International conference on pattern recognition (ICPR). IEEE, 2016.

[9]. D.Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. (2016). ``Deep learning for identifying metastatic breast cancer." [Online].Available:https://arxiv.org/abs/1606.05718

[10]. B. E. Bejnordi et al., ``Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,"JAMA, vol. 318, no. 22, pp. 2199_2210, Dec. 2017.

[11]. B. E. Bejnordi et al., ``Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," J. Med. Imag., vol. 4, no. 4, p. 044504, 2017.

[12]. A. Vahadane et al., ``Structure-preserved color normalization for histological images," in Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI),Apr. 2015, pp. 1012_1015.
[13]. M. Macenko et al., ``A method for normalizing histology slides for quantitative analysis," in Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro (ISBI), Jun./Jul. 2009, pp. 1107_1110.
[14]. E. Reinhard, M. Adhikhmin, P. Shirley, and B. Gooch, ``Color transfer between images," IEEE Comput. Graph. Appl., vol. 21, no. 5, pp. 34_41,Sep./Oct. 2001.
[15]. A. Pêgo and P. Aguiar. (2015). Bioimaging. [Online]. Available: http://www.bioimaging2015.ineb.up.pt/dataset.html
[16]. A. K. Mary. (2018). Breast Cancer. [Online]. Available: https:// www.merckmanuals.com/home/women-s-health-issues/breast-disorders/breast-cancer
[17]. K. He, X. Zhang, J. Sun, and S. Ren, ``Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016,pp. 770_778.
[18]. K. He and J. Sun, ``Convolutional neural networks at constrained time cost," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015,pp. 5353_5360.
[19]. Y. L. Boureau, J. Ponce, and Y. LeCun, ``A theoretical analysis of feature pooling in visual recognition," in Proc. 27th Int. Conf. Mach. Learn.(ICML), 2010, pp. 111_118.
[20]. J. Yosinski, J. Clune, H. Lipson, and Y. Bengio, ``How transferable are features in deep neural networks?" in Proc. Adv. Neural Inf. Process. Syst.,2014, pp. 3320_3328.
[21]. N. Tajbakhsh et al., ``Convolutional neural networks for medical image analysis: Full training or _ne tuning?" IEEE Trans. Med. Imag., vol. 35,no. 5, pp. 1299_1312, May 2016.
[22]. J. Deng, W. Dong, L.-J. Li, K. Li, L. Fei-Fei, and R. Socher, ``ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248_255.
[23]. T. Araújo et al., ``Classi_cation of breast cancer histology images using convolutional neural networks," PLoS ONE, vol. 12, no. 6, p. e0177544, 2017.