Detection of DCIS and IDC in Whole-Slide H&E Stained Breast Histopathology Images Using Stacked Convolutional Neural Networks

Guido C. A. Zuidhof*, Babak Ehteshami Bejnordi, Geert Litjens, and Jason Farquhar

***Abstract*—This paper presents and evaluates a method for detection and localization of breast malignant lesions in histopathological breast tissue images. The goal of this method is to best classify digitized whole-slide hematoxylin and eosin (H&E) stained whole-slide images (WSIs) into three classes: benign, ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC). A convolutional neural network (CNN) was trained on small patches of these images, labeled with the class of the center pixel. To distinguish between the DCIS and IDC class the information available in these small patches is often not sufficient. A second CNN was trained on the output of the aforementioned network with a larger patch size to capture a larger context. Sliding this classifier across a whole slide image yields a probability map. From this probability map structural and statistical features were extracted, which were used to train a final classifier to predict the label for the whole-slide image. The system is evaluated on a dataset containing X WSIs of breast tissue using FROC analysis. The results show a...**

***Index Terms*—Computer-aided diagnosis, DCIS and IDC detection, deep learning, whole-slide imaging.**

I. Introduction

BREAST cancer is the most common cancer in women [1]. Cancers of the breast kill more women than any other form of cancer in all parts of the developing world [2]. An important tool for the detection and management of breast cancer is analysis of tissue samples under the microscope by a pathologist.

Looking at cancer cells under the microscope, the pathologist searches for certain features that can help predict how likely the cancer is to grow and spread. These features include the spatial arrangement of the cells, morphometric characteristics of the nuclei, whether they form tubules, and how many of the cancer cells are in the process of dividing (mitotic count). These features taken together determine the extent or spread of cancer at the time of diagnosis.

Increasingly, slides of human tissue are digitized instead of being analysed only under a microscope. This has spawned the relatively new field of *digital pathology*. Digital images as opposed to glass slides allow for easier consultations between pathology experts. An additional advantage of digital images is the opportunity to analyze these images automatically using computer algorithms.

Visual microscopic interpretation of tissue sections is laborious and prone to subjectivity. *Computer-aided diagnosis (CAD)* has a huge potential in alleviating shortcomings of human interpretation and will reduce the workload of the pathologists. As a result, more accurate diagnostic information may be extracted, helping clinicians in selecting the most optimal treatment for individual patients. CAD can facilitate diagnosis by sieving out obviously benign slides and providing quantitative characterization of suspicious areas.

A. Patch-based classification

Convolutional neural networks, as well as other statistical machine learning methods, are bound by computational and memory constraints. The tissue slides are scanned at a high magnification (up to 40X), resulting in very large images. A typical image can be 100,000 by 200,000 pixels, and have a file size of around 20GB uncompressed.

To create a classification for the whole-slide image a patch based method will be employed. A patch is a small sub-image of the whole image. By applying the classifier in a sliding window approach over the original image we can create a prediction map for the whole image.

A trade-off between input size, mini-batch size and model complexity has to be made. If the input size increases, the amount of activations in the feed-forward step of training also increases. These are required for computing the gradient, and thus have to be stored in memory, leading to a larger memory footprint. The amount of images used to determine the error gradient for one update step is called the batch size. A larger batch size generally leads to a less noisy gradient as it is averaged over more data points. This does not necessarily lead to a faster convergence, or to a better local optimum, as the noise can be beneficial, but in the general sense a larger batch size is beneficial.

B. Incorporating more context

As a patch is limited in size, it can only contain so much information. One could downsample a larger image to fit a

This paper was written as part of the master's thesis in Artificial Intelligence of G. C. A. Zuidhof. This project was supervised by B. Ehteshami Bejnordi, G. Litjens and J. Farquhar. *Asterisk indicates corresponding author.*

*G. C. A. Zuidhof is a student in the Artificial Intelligence Master's Programme at the Radboud University, 6525HP Nijmegen, The Netherlands (e-mail: guido.zuidhof@student.ru.nl).

B. Ehteshami Bejnordi is with the Diagnostic Image Analysis Group, Radboud University Medical Center, 6500HB Nijmegen, The Netherlands.

G. Litjens is with the Department of Pathology, Radboud University Medical Center, 6500HB Nijmegen, The Netherlands.

J. Farquhar is with the Department of Artificial Intelligence, Radboud University, 6525HP Nijmegen, The Netherlands and the Donders Institute for Brain, Cognition and Behaviou, 6525EN Nijmegen, The Netherlands.

First draft, work in progress; October 1, 2016.

patch, but fine-grained detail would be lost. To distinguish between benign and cancerous tissue, this fine detail is especially important. To tell apart DCIS from IDC, larger scale architectural features play a more important role.

One approach to incorporate both fine-grained detail as well as larger scale structural information is the use of *multi-scale convolutional neural networks* [3], [4]. With this approach an image is offered to one network at different magnifications. Often, there are multiple sub-networks for the different magnification levels. The outputs of these sub-networks is then fused, generally using one or more fully connected layers. This allows for training the system as a whole using ordinary backpropagation.

In this paper we will be exploring a different method for incorporating more context, whilst keeping the same level of detail, that of *stacked convolutional neural networks*.

C. Stacked convolutional neural networks

Convolutional neural networks generally consist of a stack of different types of layers. Some types of layers retain the spatial information of the image, such as convolution layers. In these layers, different filters are convolved over the image. These learned filters are not dependent on the input size and this layer can thus be fed an image of a different size without problems. This characteristic is exploited in *fully convolutional neural networks*, which are the state of the art in segmentation [5].

In such a framework the network the output segmentation size is only dependent on the input image size, it can thus be used to segment variable sized images. We will be exploiting this here as well. First, we train a network to classify patches of 224 by 224 pixels. This network will especially learn feature representations that are in the fine detail of the tissue.

Then, we strip this network of its last few layers that do not satisy this fully convolutional requirement of having spatial information present and are input size dependent. We increase the input size of the network to 768 by 768 pixels, and do a forward pass through this network. The output is a volume, a feature volume that resembles an image with many feature maps (or color channels).

We then train another convolutional neural network using this feature volume as its input. The learned weights in the 224 by 224 input-sized network are no longer updated (i.e. they are frozen). The idea is that this stacked network has access to the learned fine detail features, and has very large receptive fields. It can thus learn to exploit both the fine detail, as well as the spatial arrangement of where specific patterns are.

By training these networks in this tiered manner, the computational requirements can be kept in check, whilst enabling both a large patch size, a decent batch size while training and a complex model with large receptive fields.

TODO more about domain, dataset

II. Dense prediction

TODO a lot

A. Architectures

We will apply two different architectures to the problem of patch classification, with patches of size 224x224 pixels.

1) Wide Residual Networks: Residual network was the winner of the 2015 ImageNet challenge [6]. Other than previous networks that entered this competition, it features no fully connected layer at the end of the network. It was the deepest succesful architecture to date featuring up to 152 layers. It can get away with this depth by using residual learning blocks, where the layers learn residual functions with relation to the input instead of learning unreferenced functions. To the extreme, it makes it easier to learn the identity function if that is optimal for that layer.

In this study we will apply an adaptation of this architecture, the wide residual network as proposed by Zagoruyko et al. [7]. They showed that wider and less deep residual networks outperform their deeper and thinner counterparts both in accuracy and efficiency.

The architecture is a recipe that has three main design choices, the N value that determines the depth, the k value that determines the width (the amount of filters) and the type of ResNet block. Here, we use $N = 4$, $k = 2$, which was empirically evaluated to fit a decent batch size and train fast. The ResNet block type is detailed in figure 2b, TODO more about this block. The resulting architecture is shown in figure 2a.

2) VGG-16: VGG-16 is a popular architecture because of its simplicity [8]. It features small 3x3 convolution filters and 2x2 pooling throughout the network and has a depth of 16 layers. The architecture is detailed in figure 2c. Unlike the ResNet architecture it does not feature skip connections, also it features a stack of two dense fully connected layers at the end.

B. Preprocessing

As a preprocessing step, the images are normalized by dividing their pixel value by 255. Then, the mean value of each color channel is subtracted to zero center the data.

C. Learning rate decay

The learning rate reduction policy is as follows. The learning rate is multiplied by 0.2 after no better validation accuracy has been observed for 8 epochs, this value is called the *patience*. This patience is increased by 20% after every reduction in learning rate (rounded up).

D. Data augmentation

Artificially increasing the amount of data by adding variations of the original data can further help train a model that generalizes well. Augmentation consists of applying (random) pertubations to the samples in ways that do not change the label of the samples. It has a regularizing effect which helps prevent overfitting. Especially effective are augmentations that are also realistic examples of real world data.

The augmentation methods used are as follows:

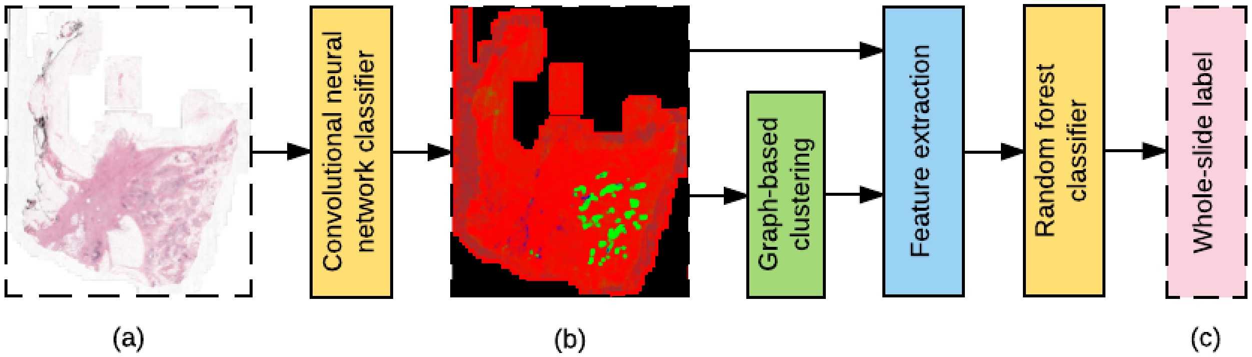
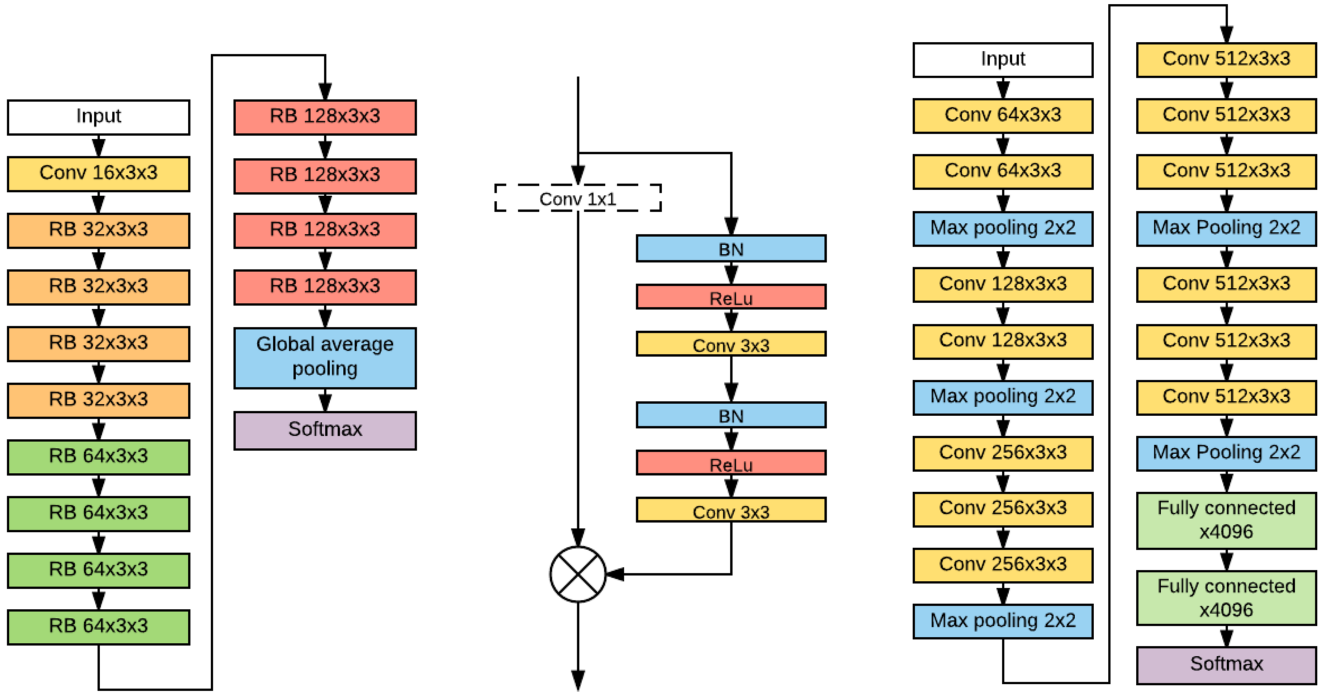


Fig. 1. Overview of the proposed system for labeling whole-slide images. (a) Original WSI of breast tissue. (b) Resulting probability map of applying the CNN in a patch-based fashion. (c) Output of the system; a single label for the whole-slide image (either benign, DCIS or IDC).



(a) Wide Residual Network architecture. *RB* stands for ResNet Block.

(b) Architecture of the resnet blocks. The ⊗ node indicates an elementwise sum.

(c) VGG-16 model architecture. Dropout with probability of 0.5 was used in the fully connected layers.

Fig. 2. Architectures used for 224x224 patch classification. The 1x1 convolution layer in the resnet blocks is only present in the blocks where the input is downsampled, which is every first block with a higher amount of filters in the Wide ResNet architecture (figure 2a).

1) Flips: The images are randomly mirrored in the X and/or Y direction, both with a 0.5 probability.

2) Rotations: The images are randomly rotated 0, 90, 180, or 270 degrees with equal probability.

3) HSV jittering: HSV is a colorspace that represents a color image in three channels; a *hue* (color), *saturation* (vibrance) and a *value* (brightness) channel. There exists a large variability in the staining of the slides, which shows itself mostly in the hue and saturation channel. We randomly jitter the hue and saturation of an image with a random value between -0.075 and 0.075. The values of all pixels are then clipped between 0 and 1.

Another commonly used data augmentation method is also zooming. This was not used here, as the size of the nuclei plays a role in determining the class label. Elastic distortions were succesfully applied as part of the data augmentation strategy in other applications of convolutional neural networks [9], also in the medical imaging domain (cytology) [10]. The irregularity of the sizes of the nuclei, as well as the architectural features are important biomarkers that can be used to distinguish between benign and cancerous lesions. Applying these elastic distortions would likely contaminate this information, and as such, it was not used here.

III. Whole-slide image labeling

IV. Results

V. Conclusion

The conclusion goes here.

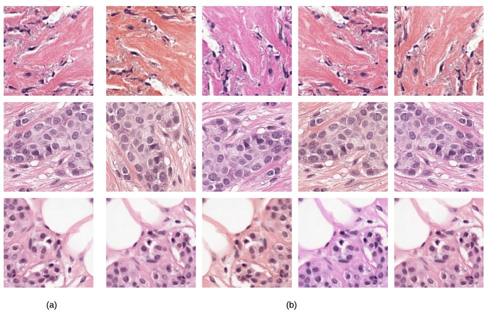


Fig. 3. Examples of augmentations applied to patches of 224 by 224 pixels. (a) The original image. (b) Augmented versions of this image. Note the slight shift in hue and saturation for some of the images.

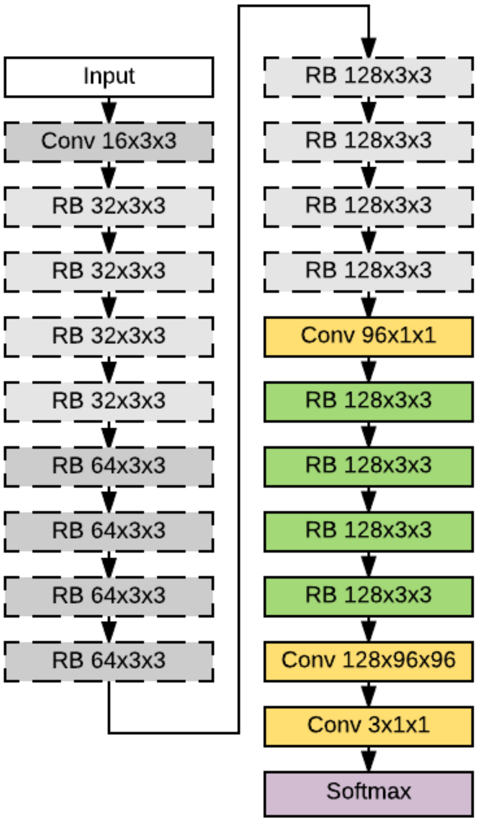


Fig. 4. Architecture of the stacked network. The weights of the components with the dotted outlines are taken from the previously trained 224x224 patch model, and are no longer updated (these are frozen).

TABLE I
BEST EPOCH PATCH-LEVEL ACCURACY OF 224X224 NETWORKS

LABELS	ARCHITECTURE	ACCURACY
Benign, Cancer		
	VGG-16	0.9237
	WRN-4-2	0.9241
Benign, DCIS, IDC		
	VGG-16	0.8245
	WRN-4-2	0.7995

ACKNOWLEDGMENT

The authors would like to thank...

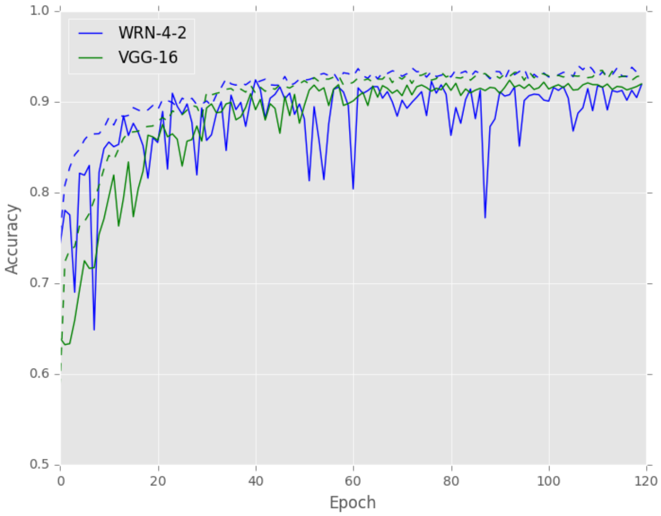


Fig. 5. Performance of networks trained on the two class problem of benign versus cancerous patches. The dashed line shows the performance on the images from the train set it was shown that epoch.

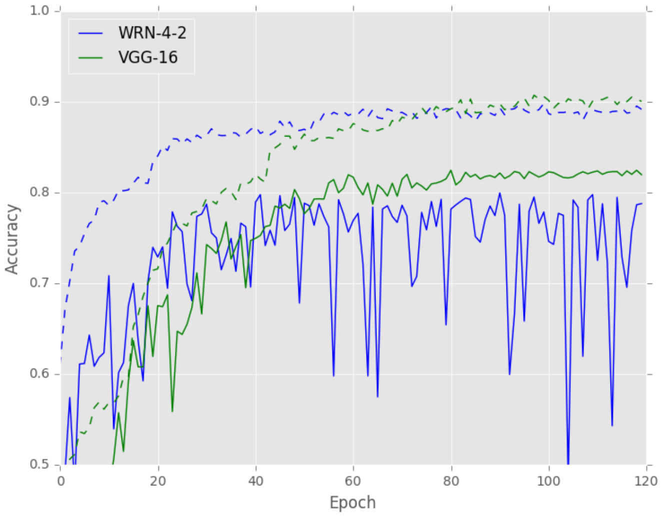


Fig. 6. Performance of networks trained on all three classes. The dashed line shows the performance on the images from the train set it was shown that epoch. The validation accuracy of the wide resnet is especially noisy.

TABLE II
BEST EPOCH PATCH-LEVEL ACCURACY OF 768X768 STACKED NETWORKS

LABELS	STACKED ON	ACCURACY
Benign, Cancer		
	2 Class WRN-4-2	123
	3 Class WRN-4-2	123
Benign, DCIS, IDC		
	2 Class WRN-4-2	123
	3 Class WRN-4-2	123

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2016,” *CA: A Cancer Journal for Clinicians*, vol. 66, no. 1, pp. 7–30, 2016. [Online]. Available: http://dx.doi.org/10.3322/caac.21332

[2] P. Porter, “Westernizing women’s risks? breast cancer in lower-income countries,” *New England Journal of Medicine*, vol. 358,

no. 3, pp. 213–216, 2008, pMID: 18199859. [Online]. Available: http://dx.doi.org/10.1056/NEJMp0708307

[3] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, *Multi-scale Convolutional Neural Networks for Lung Nodule Classification*. Cham: Springer International Publishing, 2015, pp. 588–599. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19992-4_46

[4] P. Buyssens, A. Elmoataz, and O. Lézoray, "Multiscale convolutional neural networks for vision–based classification of cells," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 342–352.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[7] S. Zagoruyko and N. Komodakis, "Wide residual networks," *CoRR*, vol. abs/1605.07146, 2016. [Online]. Available: http://arxiv.org/abs/1605.07146

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] J. P. Patrice Y. Simard, Dave Steinkraus, "Best practices for convolutional neural networks applied to visual document analysis." Institute of Electrical and Electronics Engineers, Inc., August 2003.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597