Classification of breast cancer histology images using transfer learning

Sulaiman Vesal[1](✉), Nishant Ravikumar[1], AmirAbbas Davari[1], Stephan Ellmann[2], Andreas Maier[1]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
sulaiman.vesal@fau.de
[2]Radiologisches Institut, Universitätsklinikum Erlangen, Germany

Abstract. Breast cancer is one of the leading causes of mortality in women. Early detection and treatment are imperative for improving survival rates, which have steadily increased in recent years as a result of more sophisticated computer-aided-diagnosis (CAD) systems. CAD systems are essential to reduce subjectivity and supplement the analyses conducted by specialists. We propose a transfer learning based approach, for the task of breast histology image classification into four tissue subtypes, namely, normal, benign, *in situ* carcinoma and invasive carcinoma. The histology images, provided as part of the BACH 2018 grand challenge, were first normalized to correct for color variations induced during slide preparation. Subsequently, image patches were extracted and used to fine-tune Google's Inception-V3 and ResNet50 convolutional neural networks (CNNs), both pre-trained on the ImageNet database, enabling them to learn domain-specific features, necessary to classify the histology images. Classification accuracy was evaluated using 3-fold cross validation. The Inception-V3 network achieved an average test accuracy of 97.08% for four classes, marginally outperforming the ResNet50 network, which achieved an average accuracy of 96.66%.

1 Introduction

According to a recent report published by the American Cancer Society, breast cancer is the most prevalent form of cancer in women, in the USA. In 2017 alone, studies indicate that approximately 252,000 new cases of invasive breast cancer and 63,000 cases of *in situ* breast cancer are expected to be diagnosed, with 40,000 breast cancer-related deaths expected to occur [1]. Consequently, there is a real need for early diagnosis and treatment, in order to reduce morbidity rates and improve patients' quality of life. Histopathology remains crucial to the diagnostic process and the gold standard for differentiating between benign and malignant tissue, and distinguishing between patients suffering from *in situ* and invasive carcinoma [2]. Diagnosis and identification of breast cancer sub-types typically involve collection of tissue biopsies from masses identified using mammography or ultrasound imaging, followed by histological analysis. Tissue samples are usually stained with Hematoxylin and Eosin (H&E) and subsequently,

visually assessed by pathologists using light microscopy. Visual assessment of tissue microstructure and the overall organization of nuclei in histology images is time-consuming and can be highly subjective, due to the complex nature of the visible structures. Consequently, automatic computer-aided-diagnosis systems are essential to reduce the workload of specialists by improving diagnostic efficiency, and to reduce subjectivity in disease classification.

Classification of histology images into cancer sub-types and metastases detection in whole-slide images are challenging tasks. Numerous studies have proposed automated approaches to address the same in recent years. Kothari et al. [3] examined the utility of biologically interpretable shape-based features for classification of histological renal tumor images. They extracted shape-based features that captured the distribution of tissue structures in each image and employed these features within a multi-class classification model. Doyle et al. [4] proposed an automated framework for distinguishing between low and high grades of breast cancer, from H&E-stained histology images. They employed a large number of image-derived features together with spectral clustering to reduce the dimensionality of the feature space. The reduced feature set was subsequently used to train a support vector machine classifier to distinguish between cancerous and non-cancerous images, and low and high grades of breast cancer. Wang et al. [5] proposed an award-winning (at the International Symposium on Biomedical Imaging) deep learning framework for whole-slide classification and cancer metastases detection in breast sentinel lymph node images. In a recent study [6], the authors proposed a convolutional neural network (CNN) based approach to classifying H&E-stained breast histology images into four tissue classes, namely, healthy, benign, *in situ* carcinoma and invasive carcinoma, with a limited number of training samples. The features extracted by the CNN were used for training a Support Vector Machine classifier. Accuracies of 77.8% for four class classification and 83.3% for carcinoma/non-carcinoma classification were achieved. In this study, we investigate the efficacy of transfer-learning for the task of image-wise classification of H&E-stained breast cancer histology images and examine the classification performance of the pre-trained Inception-V3 [7] and ResNet50 [8] networks, on the BACH 2018 challenge data set.

2 Methods

The data set used in this study was provided as part of BACH 2018 grand challenge[1], comprising H&E-stained breast histology microscopy images. The images are high-resolution (2040×1536 pixels), uncompressed, and annotated as normal, benign, *in situ* carcinoma or invasive carcinoma, as per the predominant tissue type visible in each image. The annotation was performed by two medical experts and images with disagreements were discarded. All images were digitized using the same acquisition conditions, with a magnification of $200\times$. The data set comprises 400 images (100 samples per class), with a pixel scale of 0.42 μm

[1] https://iciar2018-challenge.grand-challenge.org/home/

$\times$ 0.42 μm. It was partitioned into training, validation (80 samples) and test (20 samples) sets, by selecting images at random for each class independently.

2.1 Stain Normalization

A common problem with histological image analysis is substantial variation in color between images due to differences in color responses of slide scanners, raw materials and manufacturing techniques of stain vendors, and staining protocols. Consequently, stain normalization is essential as a pre-processing step, prior to conducting any analyses using histology images. Various strategies have been proposed for stain normalization in histological images. In this paper, we used the approach proposed by Reinhard et al. [9] which matches the statistics of color histograms of a source and target image, following transformation of the RGB images to the de-correlated LAB color space. Here, the mean and standard deviation of each channel in the source image is matched to that of the target by means of a set of linear transforms in the LAB color space. Histogram matching techniques assume that the proportions of stained tissue components for each staining agent are similar across the images being normalized. Fig. 2 illustrates the effect of stain normalization on a few samples from the breast cancer histology image data set using the method proposed in [9].

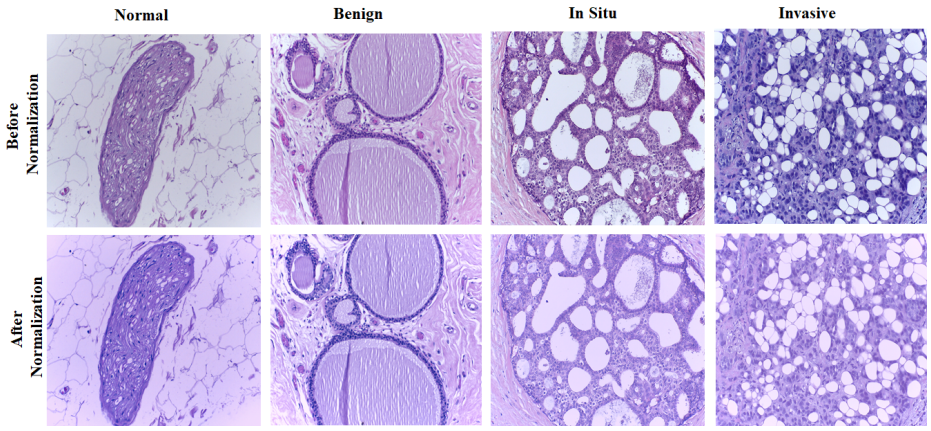

Fig. 1: Examples of histology images from each class before (top row) and after (bottom row) stain normalization.

2.2 Pre-processing

Deep learning approaches are heavily dependent on the volume of training data available, with models of higher complexity requiring more data to generalize well and avoid over-fitting to the training samples. A common challenge in the

medical domain is a lack of sufficient data, as was the case with the BACH 2018 challenge. Additionally, the breast histology images provided in the challenge data set are very large in size, spanning 2040 $\times$ 1536 pixels. In order to address the issues of limited data and large image sizes, we extracted patches from each image and augmented the data set using a variety of rigid transformations, thereby increasing the number of training samples. Image-wise classification into tissue/cancer sub-types requires learning features describing overall tissue architecture and localized organization of nuclei. Consequently, we chose to extract patches of size 512 $\times$ 512 pixels from each image, while ensuring 50% overlap between patches (similar to [6]), as there was no guarantee that smaller patches would contain information relevant to the class assigned to the whole image. This resulted in the extraction of 35 patches from each image and a final data set comprising 11,200 patches.

Additionally, to enrich the training set we augmented the data by applying varying degrees of rotation and flipping the extracted patches. This mode of data augmentation emulates a real-world scenario as there is no fixed orientation adopted by pathologists when analyzing histology slides/images. Such a patch extraction and dataset augmentation approach have been used previously for an identical classification problem [6]. The training data was augmented by flipping the extracted patches along their horizontal and vertical edges. Thus, each patch was transformed to create 2 additional, unique patches resulting in a total of 33,600 training and validation patches from the original 320 training images. During the training, we also applied real-time augmentation to rotate the patches randomly by 90, 180, 270 degrees. The label for each patch was inherited from the class assigned to the original image. The remaining 'unseen' 80 images were used as test data, to evaluate the classification accuracy of the methods investigated.

2.3 Pre-trained CNN Architectures

The application of CNNs pre-trained on large annotated image databases, such as ImageNet for example, to images from different modalities/domains, for various classification tasks, is referred to as transfer learning. Pre-trained CNNs can be fine-tuned on medical image data sets, enabling large networks to converge quicker and learn domain-/task-specific features. Fine-tuning pre-trained CNNs is crucial for their re-usability [10]. With such an approach, the original network architecture is maintained and the pre-trained weights are used to initialize the network. The initialized weights are subsequently updated during the fine-tuning process, enabling the network to learn features specific to the task of interest. Recently, numerous studies have demonstrated that fine-tuning is effective and efficient for a variety of classification tasks in the medical domain [11]. In this study, we investigate two well known pre-trained CNN architectures, namely, Google's Inception-V3 [7] and deep residual convolutional (ResNet50) network [8], which are fine-tuned to learn domain and modality specific features for classifying breast histology images. ResNet50 is based on a residual learning framework where, layers within a network are reformulated to learn a residual

mapping rather than the desired unknown mapping between the inputs and outputs. Such a network is easier to optimize and consequently, enables training of deeper networks, which correspondingly leads to an overall improvement in network capacity and performance. The Inception-V3 network employs factorized inception modules, allowing the network to choose suitable kernel sizes for the convolution layers. This enables the network to learn both low-level features with small convolutions and high-level features with larger ones.

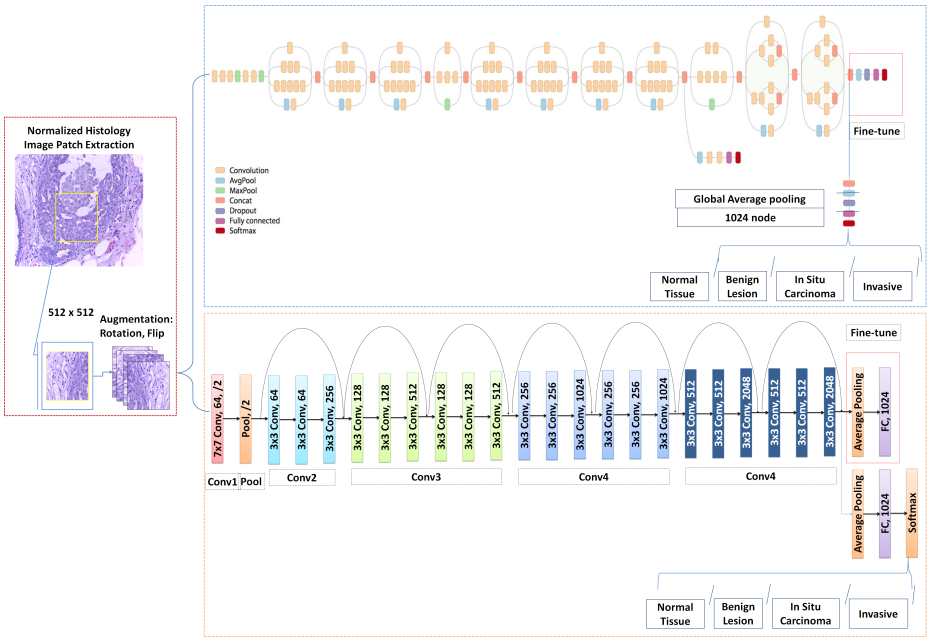

Fig. 2: Breast histology image classification workflow by fine-tuning Google's Inception-V3 and ResNet50 network architectures. The block on the left represents the pre-processing steps and the blocks on the right depict the Inception-V3 (top) and ResNet50 (bottom) network architectures.

The dataset was pre-processed as described in the previous section and used to fine-tune Google's Inception-V3 and ResNet50 networks. While such a transfer learning approach has been adopted for a variety of classification and detection tasks in medical images, few studies have employed the same for breast cancer histology image classification. Fig. 2 describes our proposed workflow for the Inception-V3 and ResNet50 network architectures. The original Inception network is modified by replacing the last 5 layers with an average global pooling layer, 1 fully connected layer, and a softmax classifier. The latter outputs probabilities for each of the four classes of interest, for each patch, fed as input to the network during the fine-tuning process. The stochastic gradient descent

optimizer with momentum was employed to train the Inception-V3 network, with a batch size of 32 for both training and validation. A learning rate and Nesterov momentum of 0.0001 and 0.9, respectively, were found to be suitable. The network stopped learning after 100 epochs. The same fine-tuning approach was applied to the ResNet50 network with identical optimization parameters. Model performance was measured by first classifying several patches extracted from each unseen test image, and then combining the classification results of all patches through a majority voting process, to obtain the final class label for each image. We trained and evaluated classification accuracy of both networks with the same configuration using 3-fold cross-validation to ensure consistency.

3 Results And Discussion

We conducted several experiments on the challenge data set to evaluate the classification performance of the networks investigated. First, the overall prediction accuracy of the networks was assessed as the ratio between the number of images classified correctly and the total number of images evaluated in the cross validation experiments. Average patch-wise and image-wise classification accuracy are presented in Table 1 for Inception-V3 and ResNet50. We also implemented the CNN model proposed in [6] to compare the performance of these transfer learning approaches with a CNN trained from scratch (refer to Table 1). Patch-wise classification accuracy of InceptionV3 for the validation and test sets were 93.40% and 92.95%, respectively. The ResNet50 network on the other hand achieved patch-wise classification accuracies of 93.02% and 92.95% for the validation and test sets, respectively. Additionally, the results presented in Table 1 indicate that transfer learning approaches achieve significant improvements in classification accuracy compared to a state-of-the-art CNN model trained from scratch [6]. As discussed previously, whole image classification was achieved using a majority voting process, based on the patch-wise class labels estimated using each network. The InceptionV3 achieved whole-image classification accuracies of 96.66% and 97.08%, for the validation and test sets, respectively. Meanwhile, the ResNet50 network achieved classification accuracies of 95.41% and 96.66% for the validation and test sets, respectively. Overall, Incepvtion-V3 and ResNet50 consistently outperformed the [6] network, achieving higher patch-wise and image-wise classification accuracy, for both the validation and test data.

We also computed the average receiver operating characteristic (ROC) curves (evaluated across the cross validation experiments) for each network, depicted in Fig. 3. ROC curves plot the true positive rate (TPR) versus the false positive rate (FPR) at different threshold settings. TPR also known as sensitivity, represents the proportion of correctly classified samples and FPR, also known as fall-out, represents the proportion of incorrectly classified samples. Thus classification accuracy was measured as the area under the ROC curve (AUC), with an area of 1 representing perfect classification on the test set. We assessed network performance for each class individually by computing their ROCs and calculated their

Table 1: Average patch-wise and image-wise classification accuracy (%) for all three networks.

Model	Patch-Wise		Image-Wise	
	Validation Set(%)	Test Set(%)	Validation Set(%)	Test Set(%)
Inception-V3	93.40	92.95	96.66	97.08
ResNet50	93.02	92.95	95.41	96.66
Arajo *et al.* [6]	88.95	88.15	82.91	93.33

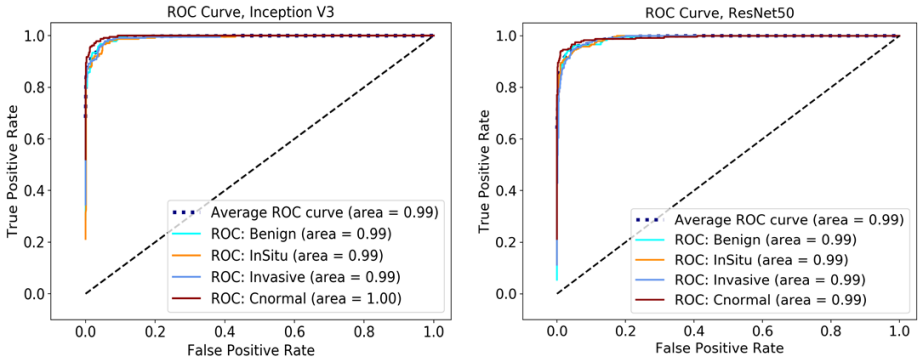

Fig. 3: ROC curves for unseen test set using Google's Inception-V3 and ResNet50 fine-tuned architectures.

corresponding AUCs (presented in Fig. 3). The overall specificity and sensitivity of both InceptionV3 and ResNet50 is approximately 99.0%.

4 Conclusions

A transfer learning-based approach for classification of H&E-stained histological breast cancer images is presented in this study. The network learns features using Google's Inception-V3 and residual network (ResNet50) architectures, which have been pre-trained on ImageNet. The data set of images provided for the BACH 2018 grand challenge are classified into four tissue classes, namely, normal, benign, *in situ* carcinoma and invasive carcinoma. We trained all the networks using 80% of the data set for training and validation, in all 3-fold cross validation experiments, and tested their performance on the remaining 20% of images. The proposed transfer-learning approach is simple, effective and efficient for automatic classification of breast cancer histology images. The investigated networks successfully transferred ImageNet knowledge encoded as convolutional features to the problem of histology image classification, in the presence of limited training data. The residual network (ResNet50) and Google's Inception-V3 outperformed a trained CNN network from scratch consistently, in terms of classification accuracy. The presented work demonstrates the applicability and pow-

erful classification capacity of transfer learning approaches, for the automatic analysis of breast cancer histology images. However, majority voting is a limitation of this study as there is a possibility that cancerous cells are present in only a small part of the image and the rest of the image depicts healthy or benign. Such cases lead to high false negative rates. Future work will look to address this limitation of majority voting by devising a suitable alternative approach.

References

1. DeSantis, C.E., Ma, J., Goding Sauer, A., Newman, L.A., Jemal, A.: Breast cancer statistics, 2017, racial disparity in mortality by state. CA: a cancer journal for clinicians **67**(6) (2017) 439–448
2. Xu, Y., Jia, Z., Wang, L.B., Ai, Y., Zhang, F., Lai, M., Chang, E.I.C.: Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics **18**(1) (May 2017) 281
3. Kothari, S., Phan, J.H., Young, A.N., Wang, M.D.: Histological image classification using biologically interpretable shape-based features. BMC Medical Imaging **13**(1) (Mar 2013) 9
4. Doyle, S., Agner, S., Madabhushi, A., Feldman, M., Tomaszewski, J.: Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, IEEE (2008) 496–499
5. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)
6. Arajo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polnia, A., Campilho, A.: Classification of breast cancer histology images using convolutional neural networks. PLOS ONE **12**(6) (06 2017) 1–14
7. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2818–2826
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
9. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **21**(5) (Sep 2001) 34–41
10. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14, Cambridge, MA, USA, MIT Press (2014) 3320–3328
11. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging **35**(5) (2016) 1285–1298