Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis

Alexander Rakhlin[1], Alexey Shvets[2], Vladimir Iglovikov[3], and Alexandr A. Kalinin[4]

[1] National Research University of Electronic Technology, Moscow, Russia
rakhlin@gmx.net
[2] Massachusetts Institute of Technology, Cambridge, MA 02142, USA
shvets@mit.edu
[3] Lyft Inc., San Francisco, CA 94107, USA
iglovikov@gmail.com
[4] University of Michigan, Ann Arbor, MI 48109, USA
akalinin@umich.edu

Abstract. Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. Computer-aided diagnosis systems showed potential for improving the diagnostic accuracy. In this work, we develop the computational approach based on deep convolution neural networks for breast cancer histology image classification. Hematoxylin and eosin stained breast histology microscopy image dataset is provided as a part of the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. Our approach utilizes several deep neural network architectures and gradient boosted trees classifier. For 4-class classification task, we report 87.2% accuracy. For 2-class classification task to detect carcinomas we report 93.8% accuracy, AUC 97.3%, and sensitivity/specificity 96.5/88.0% at the high-sensitivity operating point. To our knowledge, this approach outperforms other common methods in automated histopathological image classification. The source code for our approach is made publicly available at https://github.com/alexander-rakhlin/ICIAR2018

Keywords: Medical imaging, Computer-aided diagnosis (CAD), Computer vision, Image recognition, Deep learning

1 Introduction

Breast cancer is the most common cancer diagnosed among US women (excluding skin cancers), accounting for 30% of all new cancer diagnoses in women in the United States [23]. Breast tissue biopsies allow the pathologists to histologically assess the microscopic structure and elements of the tissue. Histopathology aims to distinguish between normal tissue, non-malignant (benign) and malignant lesions (carcinomas) and to perform a prognostic evaluation [8]. A combination of

hematoxylin and eosin (H&E) is the principal stain of tissue specimens for routine histopathological diagnostics. There are multiple types of breast carcinomas that embody characteristic tissue morphology, see Fig. 1. Breast carcinomas arise from the mammary epithelium and cause a pre-malignant epithelial proliferation within the ducts, called ductal carcinoma *in situ*. Invasive carcinoma is characterized by the cancer cells gaining the capacity to break through the basal membrane of the duct walls and infiltrate into surrounding tissues [21].

Morphology of tissue, cells, and subcellular compartments is regulated by complex biological mechanisms related to cell differentiation, development, and cancer [14]. Traditionally, morphological assessment and tumor grading were visually performed by the pathologist, however, this process is tedious and subjective, causing inter-observer variations even among senior pathologists [7,17]. The subjectivity of the application of morphological criteria in visual classification motivates the use of computer-aided diagnosis (CAD) systems to improve the diagnosis accuracy, reduce human error, increase the level of inter-observer agreement, and increased reproducibility [21].

There are many methods developed for the digital pathology image analysis, from rule-based to applications of machine learning [21]. Recently, deep learning based approaches were shown to outperform conventional machine learning methods in many image analysis task, automating end-to-end processing [4,11,13]. In the domain of medical imaging, convolutional neural networks (CNN) have been successfully used for diabetic retinopathy screening [20], bone disease prediction [26] and age assessment [12], and other problems [4]. Previous deep learning-based applications in histological microscopic image analysis have demonstrated their potential to provide utility in diagnosing breast cancer [1, 2, 21, 24].

In this paper, we present an approach for histology microscopy image analysis for breast cancer type classification. Our approach utilizes deep CNNs for feature extraction and gradient boosted trees for classification and, to our knowledge, outperforms other similar solutions.

2 Methods

2.1 Dataset

The image dataset is an extension of the dataset from [1] and consists of 400 H&E stain images (2048×1536 pixels). All the images are digitized with the same acquisition conditions, with a magnification of $200\times$ and pixel size of $0.42\mu m \times 0.42\mu m$. Each image is labeled with one of the four balanced classes: normal, benign, *in situ* carcinoma, and invasive carcinoma, where class is defined as a predominant cancer type in the image, see Fig. 1. The image-wise annotation was performed by two medical experts [10]. The goal of the challenge is to provide an automatic classification of each input image.

2.2 Approach overview

The limited size of the dataset (400 images of 4 classes) poses a significant challenge for the training of a deep learning model [4]. Very deep CNN architectures

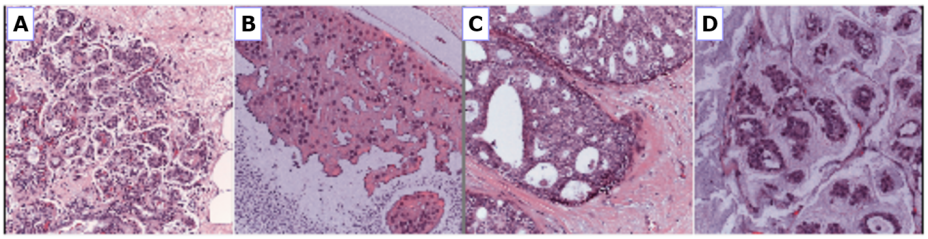

Fig. 1: Examples of microscopic biopsy images in the dataset: (A) normal; (B) benign; (C) *in situ* carcinoma; and (D) invasive carcinoma

that contain millions of parameters such as VGG, Inception and ResNet have achieved the state-of-the-art results in many computer vision tasks [25]. However, training these neural networks from scratch requires a large number of images, as training on a small dataset leads to overfitting i.e. inability to generalize knowledge. A typical remedy in these circumstances is called fine-tuning when only a part of the pre-trained neural network is being fitted to a new dataset. However, in our experiments, fine-tuning approach did not demonstrate good performance on this task. Therefore, we employed a different approach known as deep convolutional feature representation [9]. To this end, deep CNNs, trained on large and general datasets like ImageNet (10M images, 20K classes) [6], are used for unsupervised feature representation extraction. In this study, breast histology images are encoded with the state-of-the-art, general purpose networks to obtain sparse descriptors of low dimensionality (1408 or 2048). This unsupervised dimensionality reduction step significantly reduces the risk of overfitting on the next stage of supervised learning.

We use LightGBM as a fast, distributed, high performance implementation of gradient boosted trees for supervised classification [15]. Gradient boosting models are being extensively used in machine learning due to their speed, accuracy, and robustness against overfitting [18].

2.3 Data pre-processing and augmentation

To bring the microscopy images into a common space to enable improved quantitative analysis, we normalize the amount of H&E stained on the tissue as described in [16]. For each image, we perform 50 random color augmentations. Following [22] the amount of H&E is adjusted by decomposing the RGB color of the tissue into H&E color space, followed by multiplying the magnitude of H&E of every pixel by two random uniform variables from the range $[0.7, 1.3]$. Furthermore, in our initial experiments, we used different image scales, the original 2048×1536 pixels and downscaled in half to 1024×768 pixels. From the images of the original size we extract random crops of two sizes 800×800 and 1300×1300. From the downscaled images we extract crops of 400×400 pixels and 650×650 pixels. Lately, we found downscaled images is enough. Thereby, each image is represented by 20 crops. The crops are then encoded into 20 descriptors.

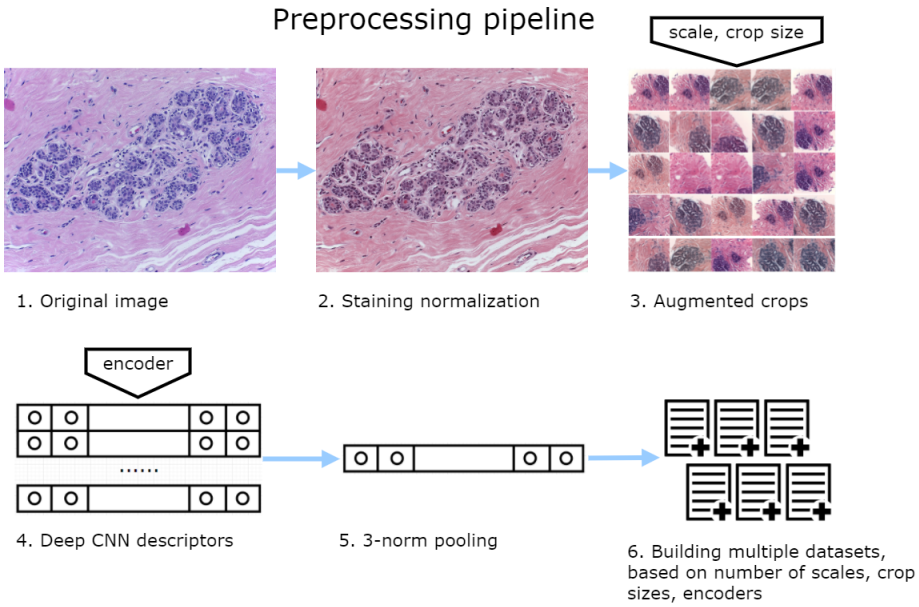

Fig. 2: An overview of the pre-processing pipeline.

Then, the set of 20 descriptors is combined through 3-norm pooling [3] into a single descriptor:

$$\mathbf{d}_{pool} = \left(\frac{1}{N} \sum_{i=1}^{N} (\mathbf{d}_i)^p \right)^{\frac{1}{p}}, \tag{1}$$

where the hyperparameter $p = 3$ as suggested in [3, 27], N is the number of crops, $\mathbf{d}_i$ is descriptor of a crop and $\mathbf{d}_{pool}$ is pooled descriptor of the image. The p-norm of a vector gives the average for $p = 1$ and the max for $p \to \infty$. As a result, for each original image, we obtain 50 (number of color augmentations) $\times 2$ (crop sizes) $\times 3$ (CNN encoders) $= 300$ descriptors.

2.4 Feature extraction

Overall pre-processing pipeline is depicted in Fig. 2. For features extraction, we use standard pre-trained ResNet-50, InceptionV3 and VGG-16 networks from Keras distribution [5]. We remove fully connected layers from each model to allow the networks to consume images of an arbitrary size. In ResNet-50 and InceptionV3, we convert the last convolutional layer consisting of 2048 channels via `GlobalAveragePooling` into a one-dimensional feature vector with a length of 2048. With VGG-16 we apply the `GlobalAveragePooling` operation to the four internal convolutional layers: `block2`, `block3`, `block4`, `block5` with 128, 256, 512, 512 channels respectively. We concatenate them into one vector with a length of 1408, see Fig. 3.

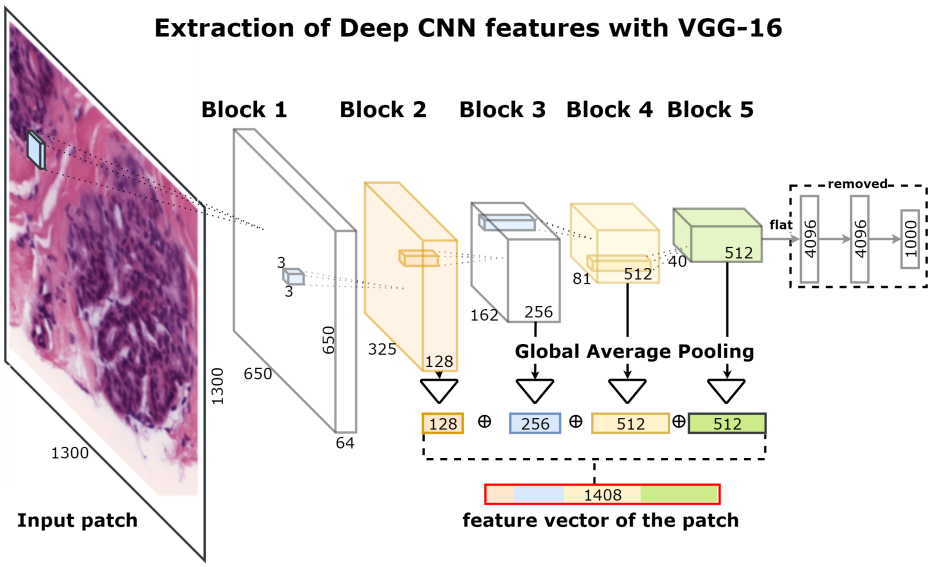

Fig. 3: Schematic overview of the network architecture for deep feature extraction.

2.5 Training

We split the data into 10 stratified folds to preserve class distribution. Augmentations increase the size of the dataset ×300 (2 patch sizes x 3 encoders x 50 color/affine augmentations). Nevertheless, the descriptors of a given image remain correlated. To prevent information leakage, all descriptors of the same image must be contained in the same fold. For each combination of the encoder, crop size and scale we train 10 gradient boosting models with 10-fold cross-validation. In addition to obtaining cross-validated results, this allows us to increase the diversity of the models with limited data (bagging). Furthermore, we recycle each dataset 5 times with different random seeds in LightGBM adding augmentation on the model level. As a result, we train 10 (number of folds) ×5 (seeds) ×4 (scale and crop) ×3 (CNN encoders) = 600 gradient boosting models. At the cross-validation stage, we predict every fold only with the models not trained on this fold. For the test data, we similarly extract 300 descriptors for each image and use them with all models trained for particular patch size and encoder. The predictions are averaged over all augmentations and models. Finally, the predicted class is defined by the maximum probability score.

3 Results

To validate the approach we use 10-fold stratified cross-validation.

For 2-class non-carcinomas (normal and benign) vs. carcinomas (*in situ* and invasive) classification accuracy was 93.8±2.3%, the area under the ROC curve

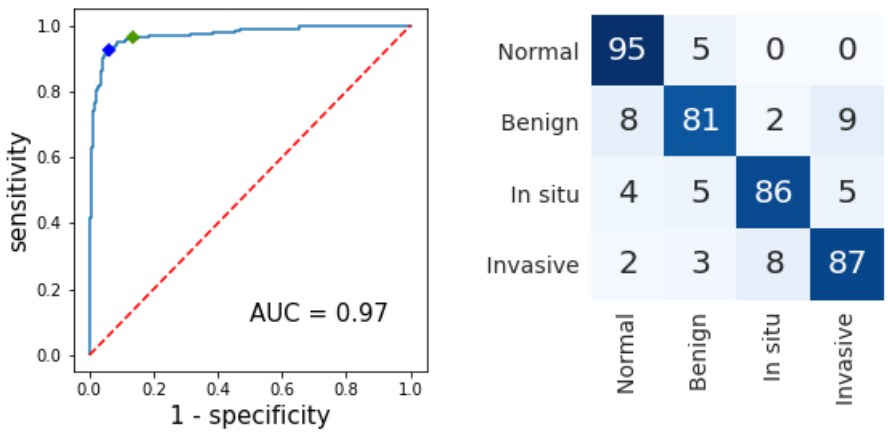

Fig. 4: a) Non-carcinoma vs. carcinoma classification, ROC. High sensitivity setpoint=0.33 (green): 96.5% sensitivity and 88.0% specificity to detect carcinomas. Setpoint=0.50 (blue): 93.0% sensitivity and 94.5% specificity b) Confusion matrix, without normalization. Vertical axis - ground truth, horizontal - predictions.

Table 1: Accuracy (%) and standard deviation for 4-class classification evaluated over 10 folds via cross-validation. Results for the blended model is in the bottom. Model name represented as ⟨CNN⟩-⟨crop size⟩, thereby VGG-650 denotes LightGBM trained on deep features extracted from 650x650 crops with VGG-16 encoder. Each column in the table corresponds to the fold number.

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	**mean**	**std**
ResNet-400	92.0	77.5	86.5	87.5	79.5	84.0	85.0	83.0	84.0	82.5	84.2	4.2
ResNet-650	91.0	77.5	86.0	89.5	81.0	74.0	85.5	83.0	84.5	82.5	83.5	5.2
VGG-400	87.5	83.0	81.5	84.0	84.0	82.5	80.5	82.0	87.5	83.0	83.6	2.9
VGG-650	89.5	85.5	78.5	85.0	81.0	78.0	81.5	85.5	89.0	80.5	83.4	4.4
Inception-400	93.0	86.0	71.5	92.0	85.0	84.5	82.5	79.0	79.5	76.5	83.0	6.5
Inception-650	91.0	84.5	73.5	90.0	84.0	81.0	82.0	84.5	78.0	77.0	82.5	5.5
std	1.8	3.5	5.7	2.8	2.0	3.7	1.8	2.1	3.9	2.7	3.0	-
Model fusion	92.5	82.5	**87.5**	87.5	**87.5**	**90.0**	85.0	**87.5**	87.5	**85.0**	**87.2**	**2.6**

was 0.973, see Fig.4a. At high sensitivity setpoint 0.33 the sensitivity of the model to detect carcinomas was 96.5% and specificity 88.0%. At the setpoint 0.50 the sensitivity of the model was 93.0% and specificity 94.5%, Fig. 4a. Out of 200 carcinomas cases only 9 *in situ* and 5 invasive were missed, Fig.4b.

Table 1 shows classification accuracy for 4-class classification. Accuracy averaged across all folds was 87.2±2.6%. Finally, the importance of strong augmentation and model fusion we use is particularly evident from the Table 1. The fused model accuracy is by 4-5% higher than any of its individual constituents.

The standard deviation of the ensemble across 10 folds is twice as low than the average standard deviation of the individual models. Moreover, all our results in the Table 1 are slightly improved by averaging across 5 seeded models.

4 Conclusions

In this paper, we propose a simple and effective method for the classification of H&E stained histological breast cancer images in the situation of very small training data (few hundred samples). To increase the robustness of the classifier we use strong data augmentation and deep convolutional features extracted at different scales with publicly available CNNs pretrained on ImageNet. On top of it, we apply highly accurate and prone to overfitting implementation of the gradient boosting algorithm. Unlike some previous works, we purposely avoid training neural networks on this amount of data to prevent suboptimal generalization.

To our knowledge, the reported results are superior to the automated analysis of breast cancer images reported in literature [1, 2, 24].

Acknowledgments The authors thank the Open Data Science community [19] for useful suggestions and other help aiding the development of this work.

References

1. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A.: Classification of breast cancer histology images using convolutional neural networks. PloS one 12(6), e0177544 (2017)
2. Bejnordi, B.E., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318(22), 2199–2210 (2017)
3. Boureau, Y.L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 111–118 (2010)
4. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Xie, W., Rosen, G.L., et al.: Opportunities and obstacles for deep learning in biology and medicine. bioRxiv p. 142760 (2017)
5. Chollet, F., et al.: Keras. `https://github.com/keras-team/keras` (2015)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
7. Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. Jama 313(11), 1122–1132 (2015)
8. Elston, C.W., Ellis, I.O.: Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 19(5), 403–410 (1991)

9. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S.: Deep learning for visual understanding: A review. Neurocomputing 187, 27–48 (2016)
10. ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. https://iciar2018-challenge.grand-challenge.org/, online; accessed January 31, 2018
11. Iglovikov, V., Mushinskiy, S., Osin, V.: Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. arXiv preprint arXiv:1706.06169 (2017)
12. Iglovikov, V., Rakhlin, A., Kalinin, A., Shvets, A.: Pediatric bone age assessment using deep convolutional neural networks. arXiv preprint arXiv:1712.05053 (2017)
13. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)
14. Kalinin, A.A., Allyn-Feuer, A., Ade, A., Fon, G.V., Meixner, W., Dilworth, D., Jeffrey, R., Higgins, G.A., Zheng, G., Creekmore, A., et al.: 3d cell nuclear morphology: microscopy imaging dataset and voxel-based morphometry classification results. bioRxiv p. 208207 (2017)
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. pp. 3149–3157 (2017)
16. Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on. pp. 1107–1110. IEEE (2009)
17. Meyer, J.S., Alvarez, C., Milikowski, C., Olson, N., Russo, I., Russo, J., Glass, A., Zehnbauer, B.A., Lister, K., Parwaresch, R.: Breast carcinoma malignancy grading by bloom–richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Modern pathology 18(8), 1067 (2005)
18. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Frontiers in neurorobotics 7, 21 (2013)
19. Open Data Science (ODS). https://ods.ai, online; accessed January 31, 2018
20. Rakhlin, A.: Diabetic retinopathy detection through integration of deep learning classification framework. bioRxiv p. 225508 (2017)
21. Robertson, S., Azizpour, H., Smith, K., Hartman, J.: Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. Translational Research (2017)
22. Ruifrok, A.C., Johnston, D.A., et al.: Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology 23(4), 291–299 (2001)
23. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2018. CA: A Cancer Journal for Clinicians 68(1), 7–30 (2018), http://dx.doi.org/10.3322/caac.21442
24. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: Breast cancer histopathological image classification using convolutional neural networks. In: Neural Networks (IJCNN), 2016 International Joint Conference on. pp. 2560–2567. IEEE (2016)
25. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. vol. 4, p. 12 (2017)
26. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S.: Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Scientific Reports 8, 1727 (2018)
27. Xu, Y., Jia, Z., Ai, Y., Zhang, F., Lai, M., Eric, I., Chang, C.: Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. pp. 947–951. IEEE (2015)