A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images

Babak Ehteshami Bejnordi*[a], Geert Litjens[b], Meyke Hermsen[b],
Nico Karssemeijer[a], and Jeroen AWM van der Laak[b]

[a] Diagnostic Image Analysis Group, Department of Radiology, Radboud University Medical Centre, Nijmegen, The Netherlands;
[b] Department of Pathology, Radboud University Medical Centre, Nijmegen, The Netherlands;

ABSTRACT

This paper presents a new algorithm for automatic detection of regions of interest in whole slide histopathological images. The proposed algorithm generates and classifies superpixels at multiple resolutions to detect regions of interest. The algorithm emulates the way the pathologist examines the whole slide histopathology image by processing the image at low magnifications and performing more sophisticated analysis only on areas requiring more detailed information. However, instead of the traditional usage of fixed sized rectangular patches for the identification of relevant areas, we use superpixels as the visual primitives to detect regions of interest. Rectangular patches can span multiple distinct structures, thus degrade the classification performance. The proposed multi-scale superpixel classification approach yields superior performance for the identification of the regions of interest. For the evaluation, a set of 10 whole slide histopathology images of breast tissue were used. Empirical evaluation of the performance of our proposed algorithm relative to expert manual annotations shows that the algorithm achieves an area under the Receiver operating characteristic (ROC) curve of 0.958, demonstrating its efficacy for the detection of regions of interest.

Keywords: histopathology, whole-slide imaging, multi-scale superpixels, region of interest, breast cancer

1. INTRODUCTION

Automated detection of clinically meaningful Regions of Interest (RoIs) in whole slide histopathological images is an important initial step in the development of an automated computer-aided diagnosis system. Accurate extraction of these RoIs would allow to perform complex image analysis tasks only on specific relevant areas within the whole slide image (WSI). This is in particular of utmost importance for efficient analysis of large histopathological images. Two major approaches have been utilized in the literature for the development of automated CAD systems to detect cancer in whole slide histopathological images. The first is to perform image analysis operations at a single specific image resolution to classify different tissue structures[1, 2]. The second utilizes a multi-resolution scheme to classify high-resolution WSI[3-6]. Contrary to the first approach which does not correspond to the multi-scale approach used by the pathologists, the second approach emulates the way pathologist examines a histology slide. The multi-resolution approach significantly reduces the computational time required to analyze the whole slide by processing the image tiles at low magnifications with the least computational burden and performing more sophisticated analysis of the corresponding tiles in higher magnification only when the decision for the classification requires more detailed information[6]. To achieve this, these algorithms make use of small fixed-size rectangular patches and try to classify them into different tissue classes. Fixed sized patches, however, can span multiple distinct tissue structures, thus degrading the classification performance.

Superpixels are alternative visual primitives which can compensate for the shortcomings of pixels and patches. A superpixel is a perceptually meaningful atomic region that aggregates visually homogeneous pixels while respecting object boundaries. Superpixels are obtained from the over-segmentation of the image. As boundary information is respected during partitioning the image into superpixels, more accurate segmentation results can be obtained by allocating superpixels to the appropriate target class. Superpixels have been increasingly used in medical imaging applications and

Further author information: (Send correspondence to B. Ehteshami Bejnordi)
B. Ehteshami Bejnordi: (✉) Babak.EhteshamiBejnordi@Radboudumc.nl

Medical Imaging 2015: Digital Pathology, edited by Metin N. Gurcan, Anant Madabhushi
Proc. of SPIE Vol. 9420, 94200H · © 2015 SPIE · CCC code: 1605-7422/15/$18
doi: 10.1117/12.2081768

greatly reduce the complexity of image processing tasks. Superpixel classification approaches have been successfully applied in several applications such as segmentation of brain MRI images[7], and prostate cancer detection and classification[8].

In this paper, we propose a multi-resolution superpixel classification approach to detect RoIs in whole slide histopathology images. The proposed system, initially partitions the image, in the lowest magnification, into a set of non-overlapping superpixels. At the lowest magnification, superpixels are classified into regions containing tissue and regions belonging to background. New superpixels are constructed at the intermediate magnification within the areas containing tissue and are classified into a particular tissue component (e.g. stroma, background, epithelial nuclei). Finally, a new set of high-resolution superpixels are built at the highest magnification only in the areas where the classifier, at the lower level, yielded a low confidence in assigning the output label. A second stage classifier is then employed to classify those superpixels more accurately. We present empirical evaluation of the performance of our algorithm on H&E stained WSIs of breast tissue and present a comparison with the traditional tile analysis algorithm for finding ROIs.

2. METHODS

Our algorithm for the detection of regions of interest has three main components. The first is identification of areas containing tissue by classifying superpixels built on the lowest magnification. The second constructs new superpixels at the intermediate magnification on the areas containing tissue and classifies them into different tissue components. The third classifies newly built superpixels at the highest magnification for the regions requiring more detailed information for accurate classification. Detailed description of different steps of the proposed algorithm are discussed below.

2.1 Tissue identification in low magnification

In this paper, the Simple linear iterative clustering (SLIC)[9] algorithm was used to generate superpixels. SLIC algorithm offers strong performance in terms of adherence to edges and segmentation speed, hence very well suited to histopathological image analysis. The proposed implementation of the SLIC algorithm performs image clustering in the CIELAB color space. However, we performed a transformation into the hue-saturation-density (HSD) color model, which was specifically designed for absorption light microscopy[10]. The HSD model transforms RGB data into two chromatic components (c_x and c_y; which are independent of the amount of stain) and a density component (D; linearly related to the amount of stain). Tissue identification is achieved by first partitioning the image into superpixels at the lowest magnification. This is followed by a classification phase to distinguish between foreground objects (tissue) and background. A pixel inside a superpixel is classified into the background class if its overall density is lower than 0.2 and the density of its r, g, and b channels is lower than 0.25. Superpixels containing more than 90% of background pixels are classified as background objects. At the end of this stage a whole slide mask is generated for areas comprising of tissue.

2.2 Tissue component classification at the intermediate magnification

For computer aided diagnosis of breast cancer the epithelial regions of the tissue are of major clinical importance[11]. Although the importance of the stromal features for prognosis of breast cancer has been recognized [12], focusing on the detection of epithelial regions does not limit the applicability of such features. Stromal features can still be computed from the stromal areas surrounding the suspicious epithelial tissue. Therefore, automated diagnosis of cancer requires identification of epithelial regions as its initial step. For this reason, the tissue was classified into three components: epithelium, stroma, and background. To classify the entire WSI into these tissue components, superpixels were generated at the intermediate magnification over the entire regions which contain tissue. In practice computational resource requirements do not allow to generate superpixels on the entire WSI at once. Therefore we need to generate superpixels separately on small image tiles containing tissue. However, this will lead to undesirable superpixel structures at the borders of the image. Figure 1 illustrates the result of generating superpixels on two consecutive tiles. As can be seen, the shape of the superpixels in the transition area between the two tiles is affected by the tile boundary. In the following section we illustrate our proposed method for generating continuous superpixels over the entire WSI.

2.2.1 Generation of continuous superpixels over the WSI

To address the problem associated with undesirable superpixel boundaries at the edge of each tile, superpixels are generated on overlapping tiles. The size of the overlap area is determined in such a way that covers at least 2 layers of superpixels from the previous tile. Figure 2 illustrates how the generation of continuous superpixels on overlapping tiles is performed. First the original image shown in Figure 1a is extended by addition of the overlapping area from the next tile. Superpixels are then generated on this image yielding the image shown in Figure 2a. To build superpixels on the next overlapping tile, we replace the overlapping area (on the left side) of the second tile image using the mask shown in Figure 2b. This mask is extracted from the overlapping area from Figure 2a, in which the values of the superpixels attached to the image boundary (on the right side) are set to one and the rest to zero. By multiplying this mask with the corresponding overlapping area of the second tile image and building new superpixels on the image the result in Figure 2c is obtained. As shown in this Figure, the black area creates a strong transition of pixel intensity values in this image which will consequently force the superpixels to adhere to the strong artificially created boundaries hence yielding a continuous superpixel arrangement in the transition area of the two tiles. The final result after stitching the tiles is presented in Figure 2d. In practice, the same technique is applied to the other sides of each patch, to preserve the superpixel continuity from all sides.

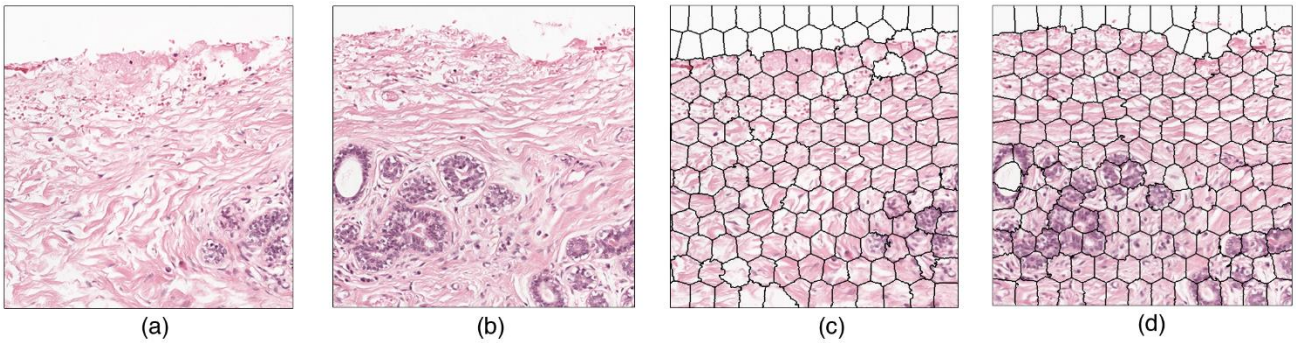

Figure 1: Illustration of generating superpixels on consecutive image tiles. (a), (b) Original images of the first and second tile. (c), (d) the output of superpixel generation on *(a)*, and *(b)*.

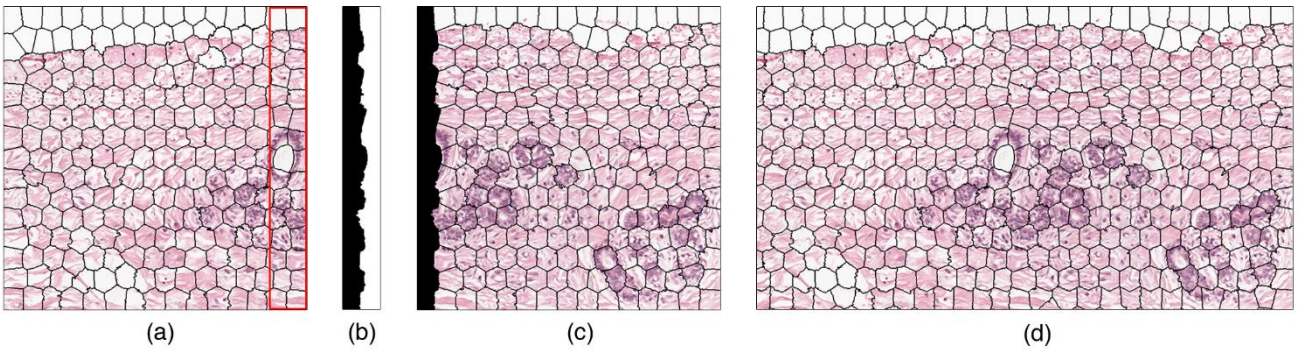

Figure 2: Illustration of generating continuous superpixels on overlapping tiles. (a) Building superpixels on the tile shown in *Figure 1a* by inclusion of overlapping area (rectangle in red). (b) The mask extracted from the overlapping area in *(a)* is achieved by setting the value of the superpixels attached to the right border of the image to 1 and the rest of pixels to 0. (c) Building superpixels on the next overlapping tile. The overlapping area of the tile is replaced using the mask generated in *(b)*. (d) The result of stitching the tiles yielding continuous superpixels over the entire WSI. Note that the last layer of superpixels attached to the right side of the image in *(a)* and the first layer of superpixels attached to the left side of the image in *(c)* are removed for stitching the two tiles.

2.2.2 Superpixel classification at the intermediate magnification

In the next stage, a classifier is constructed which operates on the regions defined by the superpixels at the intermediate magnification. A total of 54 features were extracted for the classification task including local binary patterns and statistics

derived from the histogram of the three channels of the HSD color model. Training data was acquired from a set of superpixels which were annotated as epithelium, stroma, and background. Identifying superpixels belonging to the background class was done by setting a threshold on the median density of the superpixels. The remaining superpixels were classified as epithelium or stroma using a random forest classifier. To defy the curse of dimensionality and to reduce the feature computation time, a feature selection experiment was carried out. Two feature selection methods were utilized:

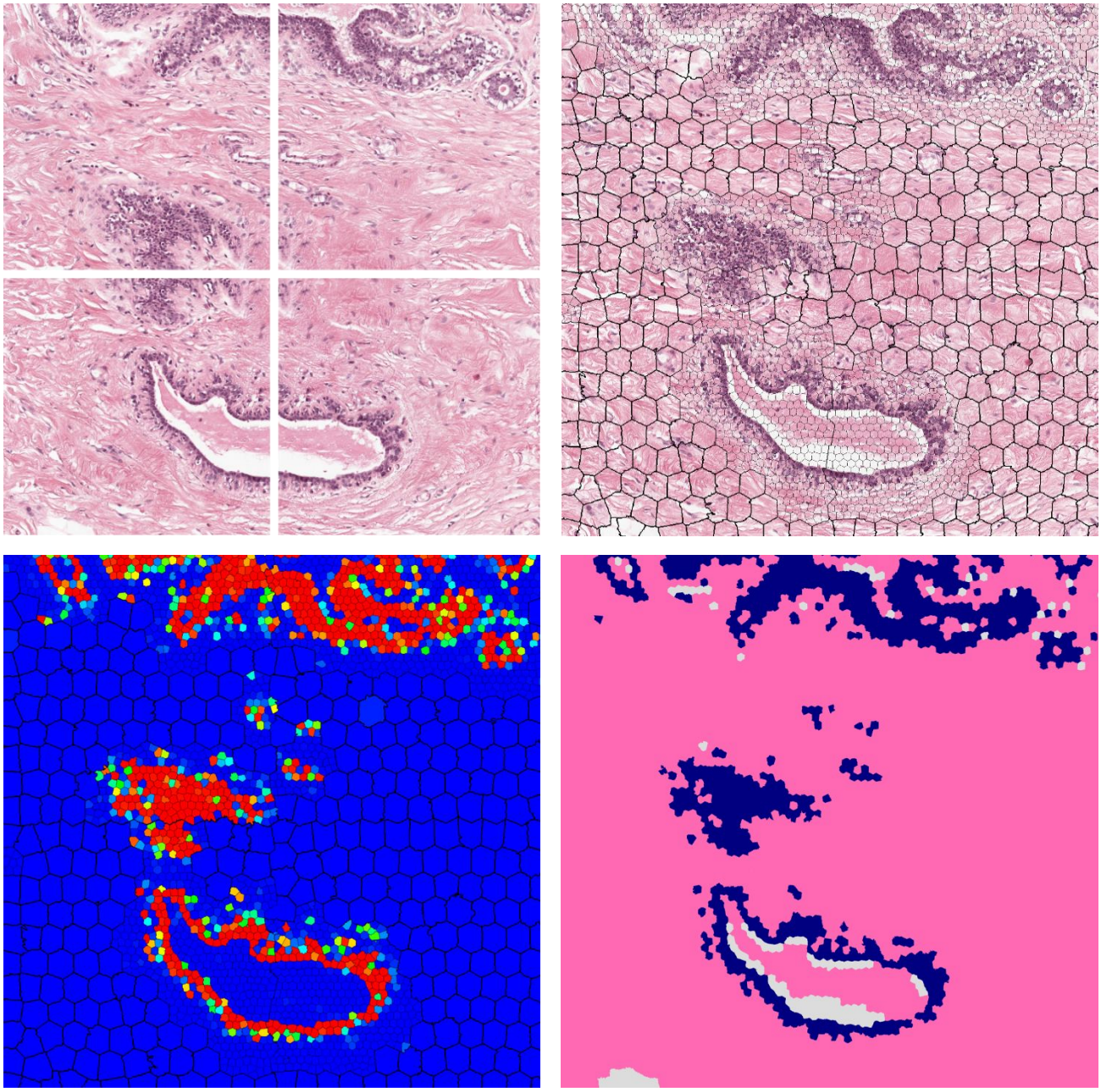

Figure 3: Illustration of multi-scale superpixel classification. (a) The original tiles. (b) Generation of superpixels in multiple levels. (c) Likelihood map showing the probability of a superpixel to belong to the epithelium class. Larger superpixels were classified with high confidence in low magnification. Smaller superpixels are generated at highest magnification and classified using the second stage classifier. (d) The result of hard classification by incorporating context information for superpixel having low confidence in their classification output.

multiple support vector machines with recursive feature elimination (MSVM-RFE)[13] and guided regularized random forest (GRRF)[14]. Feature selection by these two methods were achieved by 100 iterations of 5-fold cross validation, each iteration with random combinations of samples in the training and the test set. The output from both methods showed that Local binary patterns of the three channels do not contribute significantly to the classification accuracy, hence excluded from the classification task in the intermediate level. The random forest classifier was therefore built on the training data using the selected features. Non-background superpixels were classified using the trained model. Based on the confidence of the classifier for assigning a label to the superpixel, we decide if more detailed information is needed to classify the region. If the probability of the superpixel belonging to a specific tissue class exceeds 90%, no further analysis is required. We perform more detailed analysis not only when the classifier has a low confidence (lower than 90%) but also when the classifier assigns the epithelium label to a superpixel. The reason for this is that we want to classify epithelium regions with more accurate boundaries which is often achievable at higher magnifications.

2.3 Tissue component classification at the highest magnification

A second stage classifier was constructed to classify only the areas which were marked as requiring more detailed analysis. For this purpose, a new set of superpixels were generated at the highest magnification on these areas. Figure 3b shows how the new set of superpixels are generated in areas requiring more detailed analysis. The newly built superpixels were classified into epithelium, stroma, and background class with the same approach illustrated in lower magnification using a second random forest classifier trained on superpixels annotated in higher magnification. A similar feature selection experiment was carried out for the classifier at this magnification. Unlike the intermediate level classification problem, local binary patterns had discriminatory power for the classification. All of the 54 extracted features were therefore used for the second random forest classifier. Finally, we performed a post-processing for the superpixels which were classified with a low confidence on the highest magnification. The new probability for these superpixels were calculated using the average probabilities of their neighboring superpixels.

3. EMPIRICAL EVALUATION

3.1 Histology images

The image data used in this study originate from a set of 10 digitized H&E stained histopathology slides of Breast tissue sampled from 10 patients. Each slide was reviewed by a pathologist and assigned a pathological diagnosis. The dataset contains two samples from each of the following cases: Normal, Ductal carcinoma in situ, Invasive ductal carcinoma, Lobular carcinoma in situ, and Invasive lobular carcinoma. The whole slide histopathology images were acquired using 3DHistech Pannoramic 250 Flash II scanner on 20X magnification.

To generate ground truth data for evaluating the performance of the algorithm, two trained subjects were recruited to delineate epithelial regions within the entire slide using ImageScope viewer tool.

3.2 Experiments and Results

To evaluate the performance of our proposed algorithm, a comparison was made against the traditional tile analysis. The exact same scheme was employed to identify ROIs by this method. The ability of the SLIC algorithm to generate approximately equal sized superpixels enables us to make a fair comparison with the tile analysis method. Consequently, each tile image was divided into rectangular arranged square patches which have the same size as the average superpixel size in SLIC algorithm. Tiles were classified at different magnification using the same classifiers used by the multi-scale superpixel classification algorithm.

The performance of the two systems were evaluated in terms of the area under the receiver operating characteristic (ROC) curve. Figure 3 illustrates the selected steps for the classification process with our proposed algorithm for 4 neighboring tiles stitched together and the likelihood map of the classifier probability output. The superpixels classified as the background class have been excluded from the analysis because the classification task for this class is very simple and including them might result in an optimistic measure for the false positive rate. The area under the ROC curve (AUC) reflecting the overall performance of the multi-scale superpixel classification algorithm was 0.958. The AUC for the tile analysis in comparison was 0.932.

4. CONCLUSIONS AND DISCUSSION

This paper presented a novel multi-scale superpixel classification approach to detect regions of interest relevant to the diagnosis of breast cancer in whole slide histopathology images. The multi-resolution whole slide analysis allows identification of areas easy to classify in low magnifications and classifications of areas requiring more detailed analysis in higher magnifications. This approach significantly reduces the computational time required to analyze the whole slide compared to pixel classification methods but comes with an additional computation cost of calculating the superpixels compared to rectangular patch classification approaches. However, compared to traditional rectangular patch based algorithm, the proposed algorithm yields better performance, as boundary information is respected during partitioning the image into superpixels. The empirical evaluation of the multi-scale superpixel classification algorithm shows that it yields very high classification performance in terms of area under ROC curve. Although, the evaluation of the performance of our algorithm has been on a breast tissue dataset, the technique described here can, in essence, be applied to other tissue types as well. Moreover, the multi-resolution superpixel classification approach can potentially be utilized to discriminate between cancerous and normal regions. This will be the subject of future work.

REFERENCES

[1] A. N. Esgiar, R. N. Naguib, B. S. Sharif *et al.*, "Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa," Information Technology in Biomedicine, IEEE Transactions on 2(3), 197-203 (1998).

[2] A. Tabesh, V. P. Kumar, H.-Y. Pang *et al.*, "Automated prostate cancer diagnosis and Gleason grading of tissue microarrays." Proc. SPIE 5747, 58-70 (2005).

[3] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features." Computer Vision and Pattern Recognition 1, I-511-I-518 (2001).

[4] S. Doyle, A. Madabhushi, M. Feldman *et al.*, "A boosting cascade for automated detection of prostate cancer from digitized histology," Medical Image Computing and Computer-Assisted Intervention–MICCAI 4191, 504-511 (2006).

[5] Y. Freund, and R. E. Schapire, "Experiments with a new boosting algorithm." International Conference on Machine Learning 96, 148-156 (1996).

[6] O. Sertel, J. Kong, H. Shimada *et al.*, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," Pattern recognition 42(6), 1093-1103 (2009).

[7] S. Ji, B. Wei, Z. Yu *et al.*, "A New Multistage Medical Segmentation Method Based on Superpixel and Fuzzy Clustering," Computational and Mathematical Methods in Medicine 2014, 13 (2014).

[8] L. Gorelick, O. Veksler, M. Gaed *et al.*, "Prostate Histopathology: Learning Tissue Component Histograms for Cancer Detection and Classification," Medical Imaging, IEEE Transactions on 32(10), 1804-1818 (2013).

[9] R. Achanta, A. Shaji, K. Smith *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(11), 2274-2282 (2012).

[10] J. A. van der Laak, M. M. Pahlplatz, A. G. Hanselaar *et al.*, "Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy," Cytometry 39(4), 275-284 (2000).

[11] M. Veta, J. Pluim, P. J. van Diest *et al.*, "Breast cancer histopathology image analysis: a review," IEEE transactions on bio-medical engineering 61(5), 1400-1411 (2014).

[12] A. H. Beck, A. R. Sangoi, S. Leung *et al.*, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," Science translational medicine 3(108), 108ra113-108ra113 (2011).

[13] D. Kai-Bo, J. C. Rajapakse, W. Haiying *et al.*, "Multiple SVM-RFE for gene selection in cancer classification with expression data," NanoBioscience, IEEE Transactions on 4(3), 228-234 (2005).

[14] H. Deng, and G. Runger, "Gene selection with guided regularized random forest," Pattern Recognition 46(12), 3483-3489 (2013).