Semi-supervised Tissue Segmentation of Histological Images

Ove Nicolai Dalheim[1]*, Rune Wetteland[1]*, Vebjørn Kvikstad[2,3], Emiel A.M. Janssen[2,3], and Kjersti Engan[1]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway
ove.nicolai@dalheim.as, rune.wetteland@uis.no, kjersti.engan@uis.no
[2] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Norway
vebjorn.kvikstad@sus.no, emilius.adrianus.maria.janssen@sus.no
[3] Department of Pathology, Stavanger University Hospital, Norway

Abstract. Supervised learning of convolutional neural networks (CNN) used for image classification and segmentation has produced state-of-the-art results, including in many medical image applications. In the medical field, making ground truth labels would typically require an expert opinion, and a common problem is the lack of labeled data. Consequently, the models might not be general enough. Digitized histological microscopy images of tissue biopsies are very large, and detailed truth markings for tissue-type segmentation are scarce or non-existing. However, in many cases, large amounts of unlabeled data that could be exploited are readily accessible. Methods for semi-supervised learning exists, but are hardly explored in the context of computational pathology. This paper deals with semi-supervised learning on the application of tissue-type classification in histological whole-slide images of urinary bladder cancer. Two semi-supervised approaches utilizing the unlabeled data in combination with a small set of labeled data is presented. A multiscale, tile-based segmentation technique is used to classify tissue into six different classes by the use of three individual CNNs. Each CNN is presented tissue at different magnification levels in order to detect different feature types, later fused in a fully-connected neural network. The two self-training approaches are: using probabilities and using a clustering technique. The clustering method performed best and increased the overall accuracy of the tissue tile classification model from 94.6% to 96% compared to using supervised learning with labeled data. In addition, the clustering method generated visually better segmentation images.

Keywords: CNN · semi-supervised learning · bladder cancer · histological images · tissue segmentation

[*] These authors contributed equally to the work

Colour and Visual Computing Symposium 2020, Gjøvik, Norway, September 16-17, 2020.

1 Introduction

In Norway, 1 748 patients were diagnosed, and 319 people died from bladder cancer in 2018. The majority of these, at 73%, were male, while the remaining 27% were female [1]. Worldwide in 2018, 199 922 people of both sexes died of bladder cancer [2], and 549 393 new patients were diagnosed, placing bladder cancer as the 10th most common cancer type in the world. Since 2001, bladder cancer (including the urinary tract) has been the fourth most common cancer diagnosis for men in Norway [3][4][5][6]. In addition, bladder cancer is known as one of the most recurring cancer types, with the probability of recurrence for high-risk patients after one year at 61% [7].

An important step in determining the cancer stage and correct treatment plan for bladder cancer patients is to examine the tissue samples that are extracted during transurethral resection. The tissue samples contain large amounts of information from individual cell characteristics, to specific cell quantities in large tissue clusters. Scanning and digitalization of the histological stains produce whole slide images (WSI), uncovering the field of computational pathology. A significant increase is seen in the number of tissue samples sent to pathologist labs, affecting the waiting time for patients [8]. The increase in amount of specimens is unfortunately not seen in the number of pathologists. Another aspect is that since the WSI is studied manually, pathologists staging and grading of bladder cancer may differ in relation to the same tissue as pathologists have a different set of subjective expectations and experiences. With computational pathology, computerized tools can aid the pathologist in diagnostic predictions, localization of interesting regions, and segmentation, to name a few applications.

During the last decade, convolutional neural networks (CNN) have proven very useful in image processing and image classification tasks [9][10]. CNNs are gaining popularity also in medical image processing and in computational pathology. The most common way to train neural networks (NN) is by supervised learning (SL) and backpropagation. This requires a large training set where all samples have associated relevant ground truth labels. Labeled data within medicine is often limited, and producing it is a time-consuming process that requires annotations made by experts. A way around the lack of labels is clustering or unsupervised learning. One method is the use of autoencoders, where a compression-decompression setup is used, making the network try to reconstruct the original input [11]. The learned features are found at the most compressed state, and might ultimately be connected to a classification network. The drawback here is that they rarely perform as well as models trained with a supervised method.

CNNs are referred to as shift-invariant, meaning that a particular feature can be detected wherever it may be located in the image. Intuitively, the initial layers of a CNN can be viewed as feature extraction, while the last layers can be viewed as the most task-specific object detection or classification layers. There are many parameters to go about when setting up a new CNN, and normally large quantities of labeled data are needed to do so. Therefore, the first layers can be inherited from a pre-trained network, and the last layers are trained from scratch, a process known as transfer learning [12].

A consolidation of the above methods is semi-supervised learning (SSL), where both labeled and unlabeled data is used to train a network. This can be beneficial in cases where there are small amounts of labeled data, but large quantities of unlabeled data. Different semi-supervised methods exist, like graph-based learning methods that often implement clustering algorithms to locate and distinguish inputs in feature space [13]. One other semi-supervised method called self-training aims to first train a NN on labeled data in a supervised manner. Thereafter, predictions are found for new unlabeled data using the first model, and finally, a new model can be trained on both the ground truth labels from annotations and the weak labels from the predictions [14].

In very recent years, we find some works on semi-supervised learning within computational pathology. In Dercksen et al. [15], a method based on autoencoders and k-means clustering of features is presented. A combination of contrastive predictive coding and multiple instance learning on breast cancer data is presented in Lu et al. [16]. In Peikari et al. [17], a cluster-then-label approach is taken using SVM classifiers. Our group presented a method for multiclass tissue classification of urothelial carcinoma in [18][19][20]. Encouraged by the results, but challenged by the lack of labeled data to generalize the model further and utilize larger amounts of unlabeled data, we propose to combine the TRI-CNN transfer-learning based architecture with semi-supervised learning.

This paper presents two methods within self-training applied to tissue segmentation of WSIs of urothelial carcinoma. The first method is a probability-based method based on predicted probabilities from an initial model. The second method is a cluster-based self-training method based on both predicted probability from the initial model and local neighborhood in the predictions.

2 Material and Methods

2.1 Data Material

The material used in this paper consists of tissue samples from tumors of patients with bladder cancer in the form of urothelial carcinoma. The tumor is removed from the patient through Transurethral Resection of Bladder Tumor (TURBT) by the use of a resectoscope. The resectoscope holds a heated wire loop for removing the tumors, and the resulting tissue will often bear marks with burnt or torn tissue. After the tumor is removed, it is fixed in formalin before being embedded into paraffin. When the paraffin is solidified, it has a similar consistency to tissue and can more easily be sliced into 4 μm thick slides with a microtome. Variation in slice thickness can occur, in turn sourcing problems like color variation and tissue folds in the resulting image, opposing and extra challenge to the classifier. The slices are then stained with Hematoxylin Eosin Saffron [21] and further scanned with the digital slide scanner system, Leica SCN400, to produce the WSI. This, as well as previous work done on the same dataset, leads to the six classes which can be seen in Fig. 1.

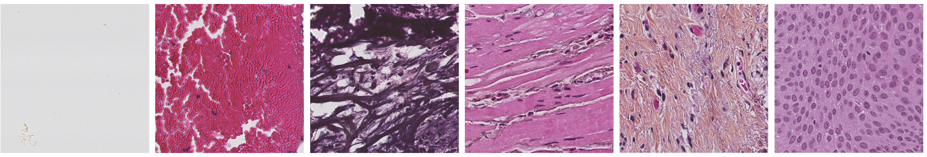


Fig. 1: Tissue representing the different classes used in the AI models.
From left to right: Background, Blood, Damaged, Muscle, Stroma, Urothelium.

The manually marked ground truth dataset, D_{gt}, consists of 37 patients, from which 125,020 tiles have been extracted. The labels originate from annotations made at 400x magnification by a pathologist at Stavanger University Hospital, (VK), illustrated in Fig. 2. It is a private dataset, however, reasonable requests may be made to the corresponding author. The three extracted tiles have the same size of 128x128 pixels, but are extracted at different magnification levels. The lower magnification tiles (25x, 100x) have a larger field-of-view than the high magnification tile (400x), allowing the multiscale model to capture both details and context of the input images. The coordinates are then saved with the three magnification levels, accompanied by its corresponding ground truth label. The dataset was divided into $D_{gt}\{train\}$ consisting of 103,650 tiles from 29 patients, and $D_{gt}\{test\}$ consisting of 21,370 tiles from 8 patients.

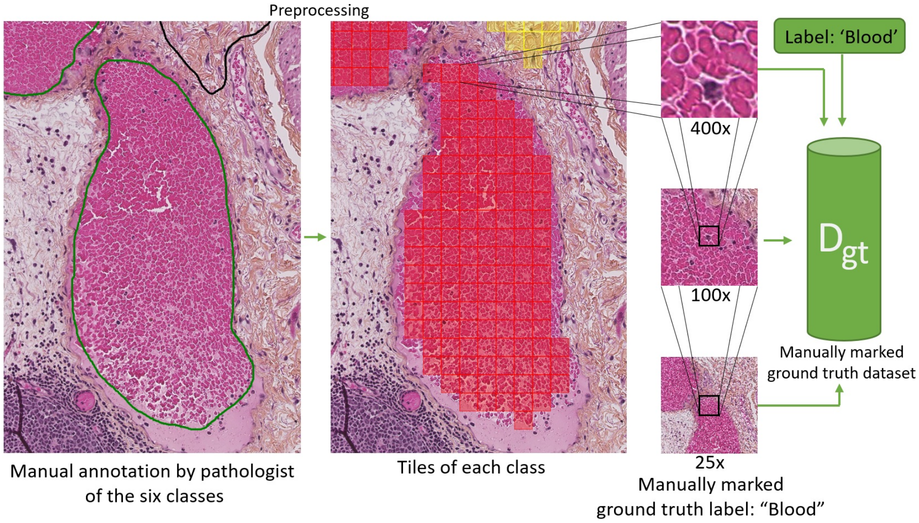

Preprocessing
Label: 'Blood'
400x
100x
25x
Dgt
Manually marked ground truth dataset
Manual annotation by pathologist of the six classes
Tiles of each class
Manually marked ground truth label: "Blood"

Fig. 2: Origin of manually marked ground truth dataset, D_{gt}.

46 new patients from the unlabeled dataset were chosen to extract tiles from, with the two self-training methods. For the probability-based method, a total of

121,239 tiles were extracted from all 46 patients and formed the probability-weak dataset, D_{pw}. For the cluster-based method, a total of 221,612 tiles were collected from 44 patients and formed the cluster-weak dataset, D_{cw}. An overview is presented in Table 1.

Table 1: Overview of labels used during training of the different models.
Ba = Background tiles, Bl = Blood tiles, Da = Damaged tissue tiles,
Mu = Muscle tissue tiles, St = Stroma tissue tiles, Ur = Urothelium tiles.

Label type	**Dataset**	**Ba**	**Bl**	**Da**	**Mu**	**St**	**Ur**	**Total**
Ground truth	$D_{gt}\{train\}$	21 423	16 949	28 452	8 061	3 614	25 151	103 650
Ground truth	$D_{gt}\{test\}$	5 589	2 883	5 155	1 905	1 261	4 577	21 370
Probability-weak	D_{pw}	20 300	20 036	20 176	20 416	20 229	20 082	121 239
Cluster-weak	D_{cw}	21 281	42 630	24 817	48 359	52 794	31 731	221 612

2.2 Methods

This section presents the original model, which originates from the framework developed by Wetteland et al. [20]. Afterwards, the methods behind the two self-training approaches within semi-supervised learning are explained.

Initial supervised approach The original model arises from a traditional supervised learning method, using the ground truth labels presented in Table 1. The dataset, D_{gt}, is split between a train/test ratio of approximately 83/17, taking into account that the same patient does not exist in both sets regardless of class. The individual per-class train/test split varies from a 86/14 ratio for blood to a 74/26 ratio for stroma. All models trained using a SL approach are referred to as TRI-SL.

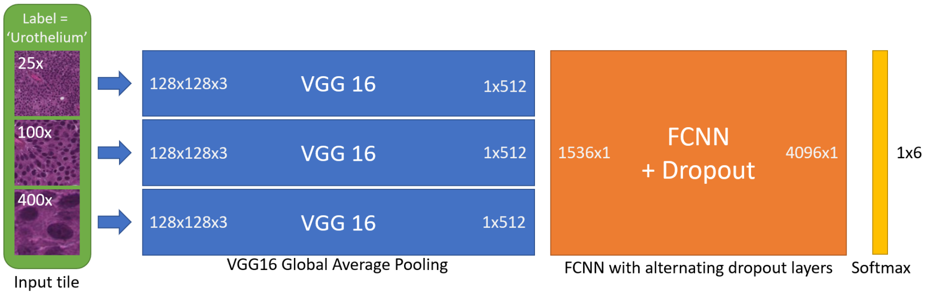


Fig. 3: Illustration of the multiscale TRI-architecture used in all models.

As illustrated in Fig. 3, the architecture of the TRI-CNN model utilizes transfer learning by implementing three VGG16 models [22] in parallel that operate individually. The VGG16 network converts a 128x128x3 input RGB image into a feature vector with dimension 1x512. This is done by a sequence of five CNN blocks that each consist of two or three CNN layers followed by a rectified linear unit (ReLU) layer and finally an average pooling layer. The three 1x512 outputs from the VGG16 models are then merged into a single 1x1536 layer followed by a fully-connected neural network (FCNN). The FCNN consists of two layers with 4096 neurons each, with one dropout layer between them. Thereafter, another dropout layer before the final output layer classifies the tissue with a Softmax activation function. Each VGG16 network is fed the input tiles at the three different magnifications 25x, 100x and 400x, to allow for different features to be detected at each level. The multiscale model is therefore abbreviated with the name TRI-CNN, which originates from the nomenclature in Wetteland et al. [20].

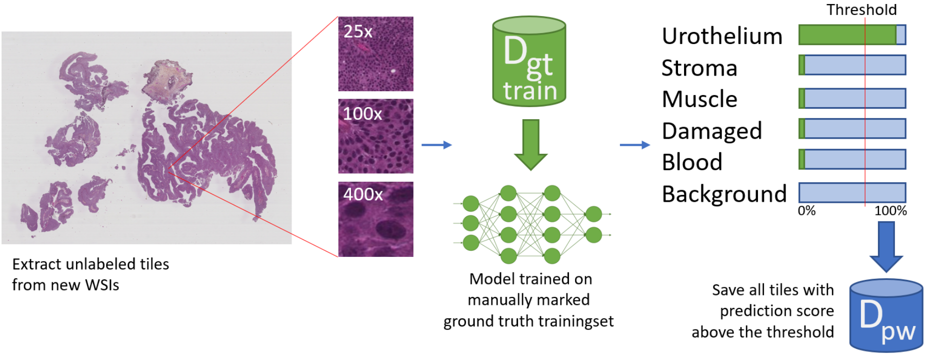

25x
100x
400x
Extract unlabeled tiles from new WSIs
D_{gt} train
Model trained on manually marked ground truth trainingset
Threshold
Urothelium
Stroma
Muscle
Damaged
Blood
Background
0%
100%
Save all tiles with prediction score above the threshold
D_{pw}

Fig. 4: Origin of probability-weak dataset, D_{pw}.

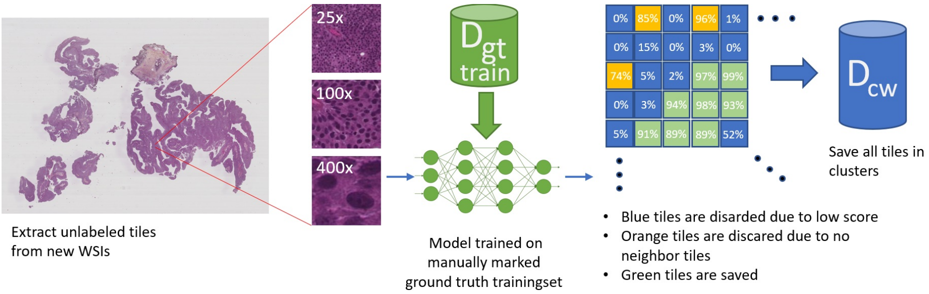
Probability-based self-training The probability-based self-training method is the most straight forward approach within self-training. Each of the 46 images is split up into tiles of size 128x128 pixels, and each tile is classified by the original model, TRI-SL-AF, which is trained on the ground truth labels. Every tile that is classified with a minimum probability threshold of 60% is saved, while tiles classified with lower probability are discarded. The 60% threshold is a trade-off between acquiring enough tiles while having a large enough probability. As illustrated in Fig. 4, the saved tiles are then selected based on several criteria given in Table 2. All models trained using the probability-based self-training method are referred to as TRI-P-SSL.

The method used to select tiles from the 46 patients is designed to select the tiles only based on its probability score across all WSIs. First, a scan runs through all the patients and counts the number of tiles per patient. Patients with

Table 2: Tile criteria for probability-weak dataset D_{pw}.
Ba = Background tiles, Bl = Blood tiles, Da = Damaged tissue tiles,
Mu = Muscle tissue tiles, St = Stroma tissue tiles, Ur = Urothelium tiles.

Criteria	**Ba**	**Bl**	**Da**	**Mu**	**St**	**Ur**
Min. tile probability	95%	80%	95%	95%	95%	95%
Max. tiles per WSI	1 900	8 000	710	5 000	5 000	480
Min. tiles per WSI	707	53	277	707	5 688	32 916
Max. tiles tot.	20 500	20 500	20 500	20 500	20 500	20 500

an insufficient number of tiles according to the minimum number of tiles per WSI are discarded, and tiles are collected from the remaining patients. For each patient, tiles with the highest probability are collected first, until the maximum number of tiles per WSI has been collected, or no more sufficient tiles remain. All tiles from all WSIs are then appended to an array and sorted based on probability. The tiles with the highest probability are then selected from this array according to the maximum total number of tiles. This is done for each class and later saved to the probability-weak dataset D_{pw}, see Table 1.


25x
100x
400x
Extract unlabeled tiles from new WSIs
D gt train
Model trained on manually marked ground truth trainingset
0% 85% 0% 96% 1%
0% 15% 0% 3% 0%
74% 5% 2% 97% 99%
0% 3% 94% 98% 93%
5% 91% 89% 89% 52%
D cw
Save all tiles in clusters

- Blue tiles are disarded due to low score
- Orange tiles are discared due to no neighbor tiles
- Green tiles are saved

Fig. 5: Origin of cluster-weak dataset, D_{cw}.

Cluster-based self-training Similar to the probability-based method, the cluster-based method uses model TRI-SL-AF to classify the WSIs. The tiles are classified with a minimum of 60% probability, and tiles with a lower probability are discarded. The classified tiles are then selected based on several criteria listed in Table 3. A visual representation of this is illustrated in Fig. 5. All models trained using the cluster-based self-training method are referred to as TRI-C-SSL.

An algorithm searches through the tiles and groups them into clusters. If, at any point in the search, the maximum number of tiles per cluster is not reached, the difference is appended to the limit of the next cluster in line. The average cluster probability is calculated per cluster, and the clusters are sorted after

Table 3: Tile criteria for cluster-weak dataset D_{cw}.
Ba = Background tiles, Bl = Blood tiles, Da = Damaged tissue tiles,
Mu = Muscle tissue tiles, St = Stroma tissue tiles, Ur = Urothelium tiles.

Criteria	**Ba**	**Bl**	**Da**	**Mu**	**St**	**Ur**
Min. tiles per WSI	50	20	50	20	50	50
Max. tiles per WSI	20 000	20 000	798	4 815	1 440	1 235
Max. clusters per WSI	100	100	100	100	100	100
Min. cluster size	50	20	50	20	50	50
Max. tiles per cluster	20 500	20 500	20 500	20 500	20 500	20 500
Min. avg. cluster probability	60%	60%	60%	60%	60%	60%

the highest probability. Each cluster originating in the WSI is then sorted into an array, and the program selects the clusters based on the highest probability according to the maximum number of clusters. The labels are then saved to the cluster-weak dataset D_{cw}, see Table 1.

3 Experimental setup

Six multiscale models are presented in this paper, and the following letters are used to describe them: SL is short for supervised learning, and SSL for semi-supervised learning. P indicates that the models are trained through the probability-based self-training method, and C implies that the cluster-based self-training method is used. A refers to that augmentation by rotation of tiles is involved. F and U refer to the weights in the VGG16 models being frozen or unfrozen during training, respectively. An overview is given in Table 4.

Table 4: Overview of the six models. st = stroma tiles, mu = muscle tiles

Model	**Magnification**	**Method**	**Augm.**	**VGG16**
TRI-SL-AF	25x,100x,400x	Supervised learning	2x st, mu	Frozen
TRI-P-SSL-F	25x,100x,400x	Probability-based SSL	No	Frozen
TRI-C-SSL-F	25x,100x,400x	Cluster-based SSL	No	Frozen
TRI-SL-AU	25x,100x,400x	Supervised learning	3x	Unfrozen
TRI-P-SSL-AU	25x,100x,400x	Probability-based SSL	3x	Unfrozen
TRI-C-SSL-AU	25x,100x,400x	Cluster-based SSL	3x	Unfrozen

Models TRI-SL and TRI-SL-AU were trained through supervised learning on dataset $D_{gt}\{train\}$ and tested on $D_{gt}\{test\}$, see Table 1. The models based on the probability-based self-training method, TRI-P-SSL and TRI-P-SSL-AU, were trained on the labels in both $D_{gt}\{train\}$ and D_{pw}. TRI-C-SSL and TRI-C-SSL-AU were trained with the cluster-based self-training method on labels from both datasets $D_{gt}\{train\}$ and D_{cw}. The models TRI-SL-F, TRI-P-SSL-F, and TRI-C-SSL-F were trained with VGG16 frozen, meaning only the FCNN and output layer was trained. For models TRI-SL-AU, TRI-P-SSL-AU, and TRI-C-SSL-AU, the VGG16 model was unfrozen during training, and weight in the whole network was adjusted.

For the original model, TRI-SL-F, stroma and muscle tissue tiles were augmented by rotation to produce two times as many tiles in an effort to equalize the dataset with respect to tiles per class. For models TRI-P-SSL-F and TRI-C-SSL-F, no augmentation was used. Models TRI-SL-AU, TRI-P-SSL-AU, and TRI-C-SSL-AU were all trained with 3x augmentation of tiles in all classes except background, as the background is filtered out such that only the foreground is processed by the models. This is done to save processing time, however, background tiles containing debris were not filtered out, and needs to be processed.

During training of all six models the learning rate was set to 1.5e-4 at a batch-size of 128. The stochastic gradient descent (SGD) backpropagation algorithm was used as optimizer, and the dropout rate was set to 20%. An early-stopping criterion was set to end training when the change in validation loss was smaller than 1e-6 for six consecutive epochs. No weighting of the different labels in the datasets was used during training. All methods were implemented in Python 3.5, with TensorFlow 1.13 [23] and Keras 2.3 [24]. Scikit-learn [25] was used for evaluation, and PyVips [26] was used to process the images.

4 Results

All six multiscale models were tested on dataset $D_{gt}\{test\}$, yielding the results in Table 5. To further investigate the individual model performance with regards to segmentation, a new WSI was segmented by all six models by tile-wise classifying all foreground regions without overlap. The WSI has been annotated by a pathologist and has not been used during training before. This WSI is referred to as WSI_segment_test, and the predictions of the WSI is compared to the ground truth annotations in it. Fig. 6 shows the close-up 400x image of an area in WSI_segment_test, where the whole foreground is labeled as blood, with the corresponding prediction by all six models. A visual comparison of an area in WSI_segment_test with multiple tissue classes is presented at a lower magnification in Fig. 7a. Predictions of the corresponding area made by both the models with the lowest and highest accuracy are compared in Fig. 7b and 7c.

5 Discussion and limitations

The most accurate model is the SSL based model TRI-C-SSL-AU, which improved the accuracy by 1.38% compared to the model from a pure supervised approach, TRI-SL-AF. Through a comparison of the predictions of TRI-C-SSL-AU with the other models, it also appears superior with regards to segmentation, being the model with the least faulty predictions in the annotated regions in Fig. 7. In addition, the prediction map in Fig. 7c, produced with model C-SSL-AU, appears to have less noise when compared to the others for WSI_segment_test.

Table 5: F1-Scores for each of the classes, and overall accuracy for the six models. Green cells indicate the best result within each category.

Model	Ba	Bl	Da	Mu	St	Ur	Total
TRI-SL-AF	100.00%	98.59%	89.14%	79.42%	96.44%	98.01%	94.61%
TRI-P-SSL-F	100.00%	98.64%	90.01%	82.68%	96.14%	98.29%	95.19%
TRI-C-SSL-F	99.99%	96.66%	90.55%	82.54%	95.93%	98.59%	95.12%
TRI-SL-AU	100.00%	99.88%	87.86%	78.10%	98.10%	99.09%	94.57%
TRI-P-SSL-AU	100.00%	97.36%	88.21%	82.18%	96.79%	99.45%	94.85%
TRI-C-SSL-AU	100.00%	98.70%	91.88%	84.71%	95.92%	98.96%	95.99%

Comparing the results in Table 5 with the different predictions in Fig. 6, it would be reasonable to assume that the model with the highest F1-Score for blood, TRI-SL-AU, would produce the most accurate prediction. TRI-SL-AU is trained through a traditional supervised approach on dataset $D_{gt}\{train\}$ that contains a relatively large amount of urothelium tiles, and achieves the 2nd highest F1-Score for urothelium. This is, however, quite the opposite of the situation, as it is the model that predicted the most urothelium tiles in the blood area in Fig. 6. This is most likely an outcome with several underlying factors: The labeled training set $D_{gt}\{train\}$ is quite small, with an even smaller test set $D_{gt}\{test\}$. It is also possible that the area in Fig. 6 contains features not present in the ground truth dataset.

Each WSI will typically produce hundreds of thousands of tiles, opposing a challenge when selecting tiles through a probability-based self-training method. A large number of tiles will have a high probability if the specific class is trained with many labels in the original model, i.e., more features have been learned for that class. To counter this, a minimum tile per patient threshold was set to discard WSI containing a small number of tiles, as they are most likely misclassified. This does, however, not prevent over-representation of the top-left portion of the WSIs, which will occur when a WSI contains large amounts of sufficient tiles of a certain class. One might also argue that the model will not learn that many new features from tiles it already is 100% certain about and that the method becomes more of an alternative to augmentation.

By using the cluster-based approach, it is safer to include tiles of lower probability, as it is safe to assume that tiles closer to each other are more likely to hold the same label. Also, the method ensures that tiles are distributed more evenly across the WSIs in comparison to the probability-based self-training method. This can also be seen as augmentation, and an unfrozen VGG16 model has a significant improvement when comparing the two cluster-based models TRI-C-SSL-F and TRI-C-SSL-AU, where accuracy increases from 95.12% to 95.99% respectively. The opposite effect is observed for augmenting and unfreezing with the probability-based models, decreasing the accuracy from 95.19% for TRI-P-SSL-F to 94.85% for TRI-P-SSL-AU. The SSL models without augmentation, TRI-P-SSL-F and TRI-C-SSL-F, performed relatively equal with regards to classification, however, TRI-C-SSL-F performs best with regards to segmentation.

As the models are fed three levels of magnification, where the ground truth marking is based on the 400x magnification, the corresponding 100x and 25x images contain very little of the same tissue type in some cases. This causes problems for the models, especially if the 100x and 25x images are both of a different tissue class than the 400x image. An example of this is how several tiles of ground truth label blood are predicted as background in Fig. 6, as this area is rather isolated from nearby tissue.

A limiting factor of this study is the small size of muscle and stroma compared to the other classes in the ground truth dataset. Augmentation techniques are implemented to try and mitigate this issue, but still, the accuracy of muscle is not as high as the other classes.

6 Conclusion and future work

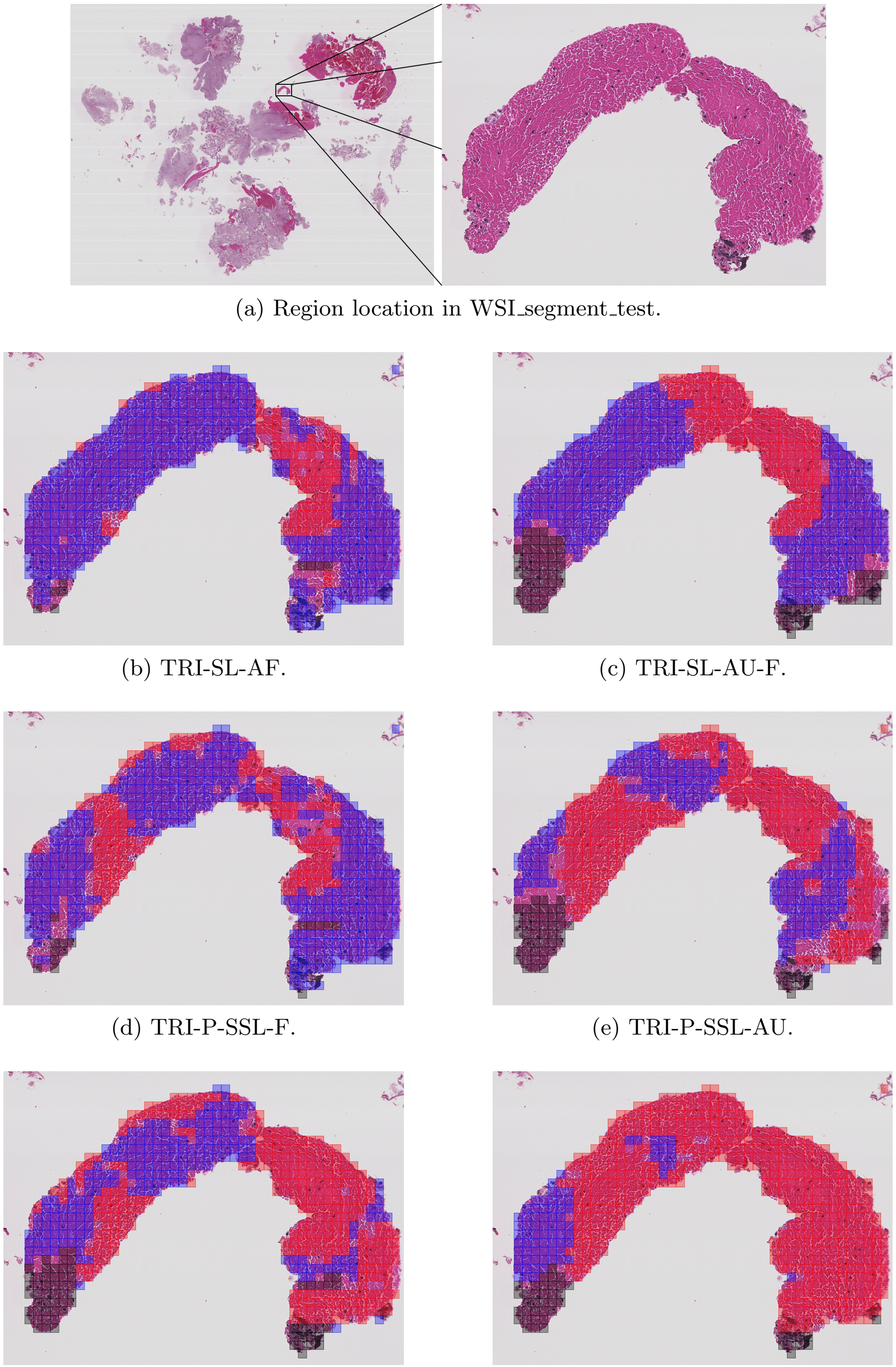
The supervised model, TRI-SL-AF, trained only on the ground truth dataset, $D_{gt}\{train\}$, achieved an accuracy of 94.61%, with 2x augmentation of the two classes with the lowest representation. By including the cluster-weak dataset, D_{cw}, the model TRI-C-SSL-AU improved the accuracy by 1.38%. Furthermore, F1-Score stayed the same or increased for every single class, and a distinct improvement is seen when comparing the prediction maps in Fig. 6 and 7.

The probability-based model TRI-P-SSL-AU saw a significant improvement in classifying urothelium, with an increase of 1.44% in F1-Score, from an initial 98.08%. The accuracy was, however, only increased by 0.24%, as the model had a large reduction in F1-Score for blood.

The two different semi-supervised methods tested, both outperformed the supervised methods with regards to classification and segmentation. This shows that the combination of clusters and probability is better than only probability. The lack of labeled data makes both methods well suited to increase the training data, however, our experiments conclude that no augmentation and frozen VGG16 weights are preferred to using augmentation and unfrozen weights in a pure probability-based approach.

For the probability-based self-training method, better distribution of tiles in the WSI is needed for this method to be improved. This can be achieved by implementing linear spacing between all tiles of a sufficient probability score per WSI. For the cluster-based self-training method, several things can be considered for future work: a) implementing a random selection of clusters with sufficient average probability, b) selecting clusters more evenly spaced, or c) increase criteria for stroma and muscle tissue classes. Implementing mixup [27] to generate more training data of under-represented classes could be a viable method for improving segmentation capabilities with regards to tiles of several tissue types.

A viable segmentation method for histological images can assist pathologists in faster evaluation speeds, as pre-segmented images can immediately point out regions of interest. In addition, the system could contribute to computer-aided diagnosis systems, which can improve the rate of grading and staging of cancer and result in a more unison and objective diagnosis.



(a) Region location in WSI_segment_test.

(b) TRI-SL-AF.

(c) TRI-SL-AU-F.

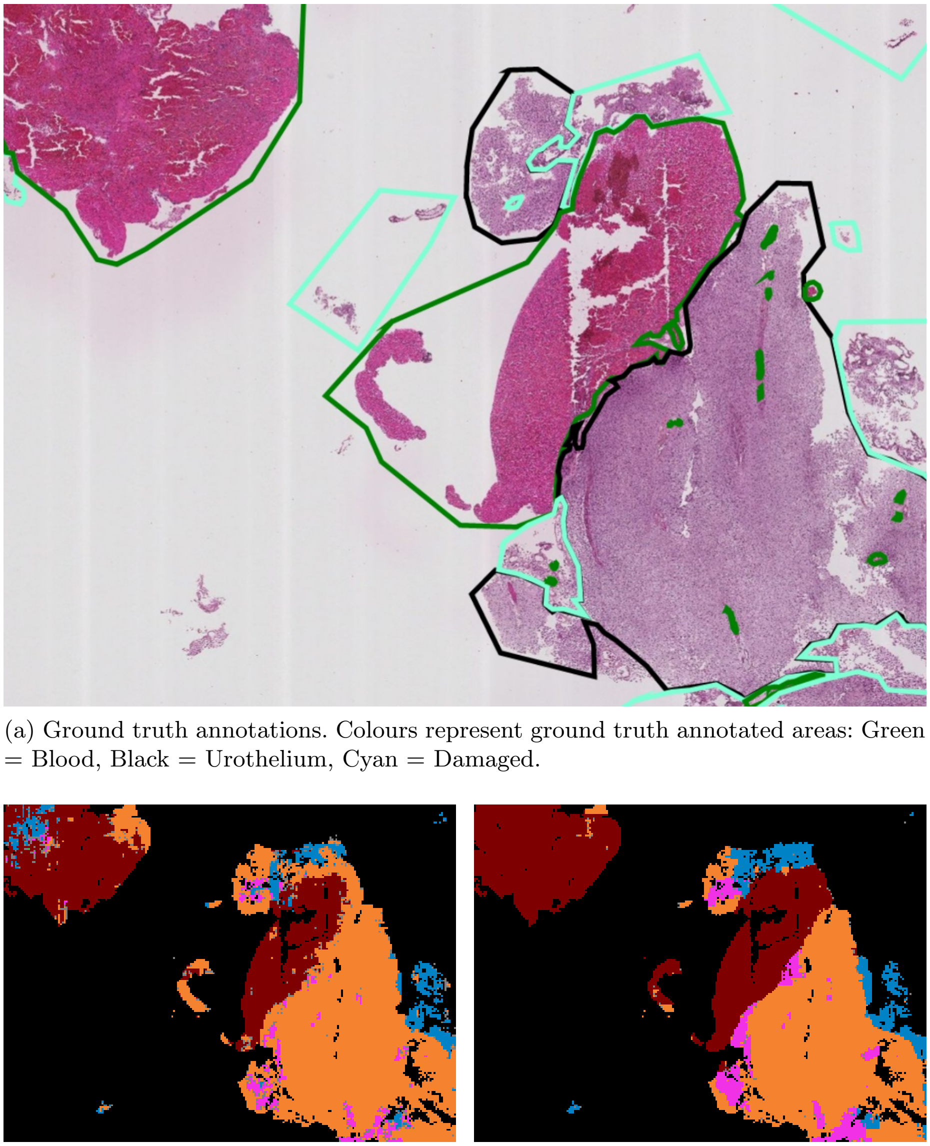
(d) TRI-P-SSL-F.

(e) TRI-P-SSL-AU.

(f) TRI-C-SSL-F.

(g) TRI-C-SSL-AU.

Fig. 6: Predictions for a region in WSI_segment_test with ground truth label blood. Color specifies predicted tile class: Blue = Urothelium tissue, Red = Blood cells, Black = Background.



(a) Ground truth annotations. Colours represent ground truth annotated areas: Green = Blood, Black = Urothelium, Cyan = Damaged.

(b) TRI-SL-AF.

(c) TRI-C-SSL-AU.

Fig. 7: Low magnification region in WSI_segment_test.
(b,c) Colours represent predicted labels: Red = Blood, Black = Background, Orange = Urothelium, Blue = Damaged, Pink = Stroma, Green = Muscle, Grey = Undefined.

Bibliography

[1] Kreftregisteret. BLÆREKREFT. URL https://www.kreftregisteret.no/Temasider/kreftformer/blarekreft/. Last accessed 29.04.2020.

[2] The Global Cancer Observatory, World Health Organization. Bladder, source: Globocan 2018. URL https://gco.iarc.fr/today/data/factsheets/cancers/30-Bladder-fact-sheet.pdf. Last accessed 15.04.2020.

[3] Cancer Registry of Norway. Cancer in norway 2005. *Cancer incidence, mortality, survival and prevalence in Norway*, page 18, 2006. ISSN 0332-9631. URL https://www.kreftregisteret.no/globalassets/publikasjoner-og-rapporter/cin2005_del1_web.pdf.

[4] Cancer Registry of Norway. Cancer in norway 2010. *Cancer incidence, mortality, survival and prevalence in Norway*, page 26, 2012. ISSN 0332-9631. URL https://www.kreftregisteret.no/globalassets/cin/2010.pdf.

[5] Cancer Registry of Norway. Cancer in norway 2015. *Cancer incidence, mortality, survival and prevalence in Norway*, page 28, 2016. ISSN 0332-9631. URL https://www.kreftregisteret.no/globalassets/cancer-in-norway/2015/cin-2015.pdf.

[6] Cancer Registry of Norway. Cancer in norway 2018. *Cancer incidence, mortality, survival and prevalence in Norway*, page 20, 2019. ISSN 0806-3621. URL https://www.kreftregisteret.no/globalassets/cancer-in-norway/2018/cin2018.pdf.

[7] A. Anastasiadis, T. M. de Reijke. Best practice in the treatment of non-muscle invasive bladder cancer. 4Therapeutic Advances in Urology:13–32, 2012. https://doi.org/10.1177/1756287211431976 3.

[8] Stavanger Aftenblad. Pasienter må vente åtte uker på prøvesvar. 2020. URL https://www.aftenbladet.no/lokalt/i/Wk332/pasienter-ma-vente-atte-uker-pa-prvesvar.

[9] G. Litjens, T. Kooi, B. E. Bejnordi et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005.

[10] Z. Shi, L. He, K. Suzuki, T. Nakamura, H. Itoh. Survey on neural networks used for medical image processing. *International journal of computational science*, 3:86–100, 2009. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4699299/.

[11] C. Tao, H. Pan, Y. Li, Z. Zou. Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geoscience and Remote Sensing Letters*, 12(12):2438–2442, 2015. https://doi.org/10.1109/LGRS.2015.2482520.

[12] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. https://doi.org/10.1109/TKDE.2009.191.

[13] Y. Song, C. Zhang, J. Lee et al. Semi-supervised discriminative classification with application to tumorous tissues segmentation of mr

brain images. *Pattern Analysis and Applications*, 12:99–115, 2009. https://doi.org/10.1007/s10044-008-0104-3.

[14] V. Cheplygina, M. de Bruijne, J. P.W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, pages 280–297, 2019. ISSN 1361-8415. https://doi.org/10.1016/j.media.2019.03.009.

[15] K. Dercksen, W. Bulten, G. Litjens. Dealing with label scarcity in computational pathology: A use case in prostate cancer classification. *Proceedings of Machine Learning Research – Accepted :1–4, 2019, Extended Abstract – MIDL 2019 submission*, 2019. URL https://arxiv.org/pdf/1905.06820.pdf.

[16] M. Y. Lu, R. J. Chen, J. Wang, D. Dillon, F. Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. 2019. URL https://arxiv.org/abs/1910.10825.

[17] M. Peikari, S. Salama, S. Nofech-Mozes et al. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific Reports*, (7193), 2018. https://doi.org/10.1038/s41598-018-24876-0.

[18] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. A. M. Janssen. Multiclass tissue classification of whole-slide histological images using convolutional neural networks. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM*, pages 320–327. INSTICC, SciTePress, 2019. ISBN 978-989-758-351-3. https://doi.org/10.5220/0007253603200327.

[19] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. A. M. Janssen. Multiscale Deep Neural Networks for Multiclass Tissue Classification of Histological Whole-Slide Images. *Medical Imaging with Deep Learning: MIDL 2019 – Extended Abstract Track*, May 2019. URL https://arxiv.org/abs/1909.01178.

[20] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad and E. A. M. Janssen. Multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. https://doi.org/10.1177/1533033820946787. Accepted for publication in Journal of Technology in Cancer Research and Treatment (TCRT) on 19 June 2020. (*in press*).

[21] E. Edston, L. Gröntoft. Saffron—a connective tissue counterstain in routine pathology. *Journal of Histotechnology*, 20:123–125, 1997. https://doi.org/10.1179/his.1997.20.2.123.

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

[23] M. Abadi et al. Tensorflow: Large-scale machine learning on heterogeneous systems. 2015. URL https://www.tensorflow.org/.

[24] F. Chollet et al. Keras. 2015. URL https://github.com/keras-team/keras.

[25] F. Pedregosa et al. A survey on transfer learning. *Scikit-learn: Machine Learning in Python*, 22(10):1345–1359, 2010. URL http://jmlr.org/papers/v12/pedregosa11a.html.

[26] J. Cupitt et al. pyvips. 2017. URL https://github.com/libvips/pyvips.

[27] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. URL http://arxiv.org/abs/1710.09412.