Radboud University Nijmegen
Radboudumc

MSc. Thesis

NWI-IMC029

Prostate Cancer Classification and Label Scarcity

Author:
Koen Dercksen

Advisors:
Wouter Bulten
Geert Litjens

Supervisor:
Arjen de Vries

2nd reader:
Francesco Ciompi

December 19, 2019

Contents

1 Introduction

Prostate cancer is one of the leading causes of death among males in developed countries, growing in frequency with 1.1 million new cases estimated in 2012 and almost 1.3 million new cases estimated in 2018 [1, 2]. While deaths from prostate cancer have decreased due to improved screening techniques, diagnosis and treatments, the number of new cases increases largely as a result of increased life expectancy, as the probability of developing prostate cancer increases with age [3]. If cancer is suspected, there are several options available to reach a diagnosis. Ultrasound and magnetic resonance imaging are the two main non-invasive methods used for prostate cancer detection. Biopsy and radical prostatectomy are invasive options, where tissue samples are collected from the prostate for further histopathological examination. During a biopsy, hollow-core needles are used to extract small quantities of prostate tissue via the rectum or urethra. A radical prostatectomy constitutes the complete removal of the prostate, and is typically performed when malignant cancer is already diagnosed, for example through a biopsy.

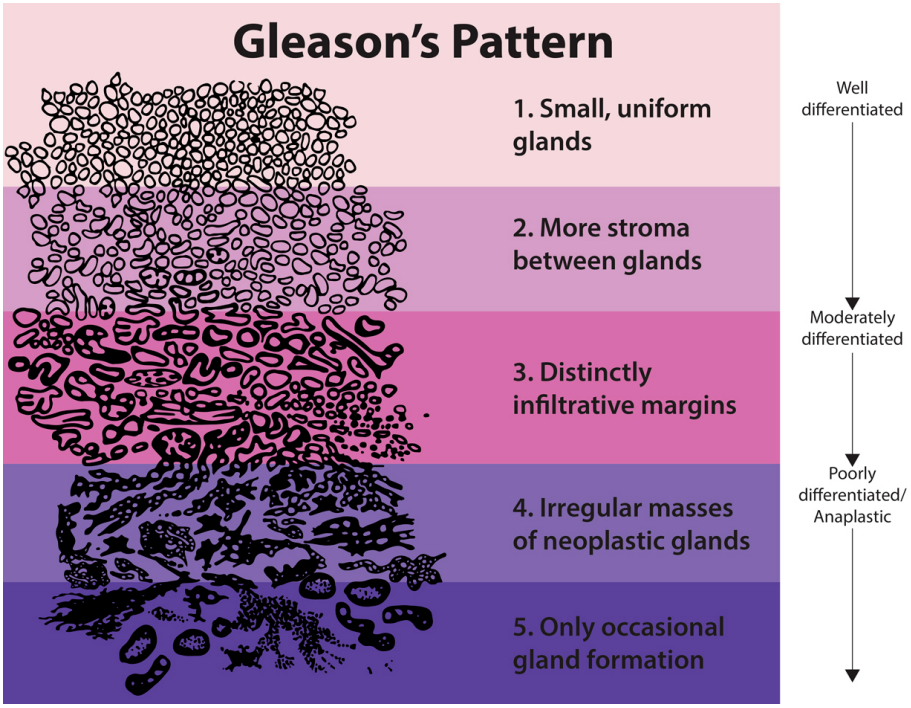

Figure 1: Histologic growth patterns consolidated into five grades, taken from National Institutes of Health. Grades 1–2 are considered benign, 3–5 are considered malignant.

Prostate cancer is graded using the Gleason grading system [4, 5, 6]. The Gleason score is based on the morphological appearance of prostate tissue. Biopsy or prostatectomy tissue is prepared for histological examination by creating microscope slides stained using hematoxylin and eosin (H&E). H&E is one of the principal and most widely used stains in histology and medical diagnosis,

and the Gleason scoring system is based on H&E stained tissue. Additionally, immunohistochemistry (IHC) staining may be used to provide an additional reference that highlights specific molecular markers useful in determining the degree of cancerous tissue in samples. The morphological features associated with certain Gleason scores can be found in Fig. 1. Alongside these guidelines, other correlations in tissue morphology and cancer have been proven; for example, the disappearance of the basal cell layer surrounding prostate epithelium indicates cancer as well [7, 8]. Uropathologists look for these morphological patterns when grading prostate cancer, a labour intensive and manual task that is subject to high inter-observer variability [9, 10, 11].

Deep learning techniques are an excellent candidate to alleviate some of this manual work on what is essentially a pattern recognition and image classification challenge. Convolutional networks in particular have been shown to surpass human performance on several visual recognition tasks [12, 13, 14]. Machine learning approaches to analyse radical prostatectomy tissue images already seem to have the potential to aid in clinical studies [15], and work on prostate epithelium segmentation in part acted as inspiration for this thesis [16, 17]. Classical supervised learning algorithms however require large amounts of labeled data. While the raw data is available, it would have to be annotated by uropathologists to create these datasets, and the required expertise and time constraints make this challenging [18]. A solution would be to improve the algorithms to make better use of scarcely labeled data. Learning with limited labeled data can be addressed in multiple ways. For example, it is possible to leverage non-expert labels through crowd-sourcing [19, 20]. There have also been efforts on using sparse annotations [21, 22]. Finally, unsupervised methods have great potential since they do not require labels at all during training.

One of the big issues with deep learning is generalization. While models often perform well on the specific task or data they are trained on, performance often lacks when trying to transfer the model to a different setting [23, 24]. It is challenging to accurately evaluate and compare deep learning systems based on a performance within a limited context, and it is important to take this into account. It is advisable to use multiple datasets, clear lower and/or upper bound numbers using e.g. a non-deep learning technique, and fair evaluation practices like averaging, k-fold validation and bootstrapping.

1.1 Problem statement

In this thesis, various degrees of supervision in deep learning are compared on a prostate cancer classification task. The aim is to show that unsupervised learning using unannotated data can extract meaningful features that can be used in classification or segmentation tasks, in order to eventually reduce manual labour requirements in clinical prostate cancer grading.

Methods from three main classes of algorithms (unsupervised, semi-supervised and supervised) are trained on radical prostatectomy data and evaluated in their ability to classify benign/malignant tissue under the condition of limited availability of labeled data. In the unsupervised approach, autoencoder embeddings are used to create a classifier using k-means. The semi-supervised approach replaces k-means with a single layer neural network trained on a limited set of labeled data. The supervised approach acts as an indication of the full capacity of the network architecture when all available labeled data is included; however,

its performance should decline with limited amounts of labeled data available, showing the advantage of semi- and unsupervised methods in these situations. Additionally, the possible benefit of including IHC stained data is investigated [16, 17]. The generalisation of each method is tested by evaluating classification performance on an external dataset of biopsy tissue images. Finally, adversarial finetuning of the semi-/unsupervised autoencoders with a learned perceptual measure is incorporated in order to find out if additional perceptual detail improves the informational value of the autoencoder latent space.

1.2 Research questions

This thesis aims to answer the following questions:

RQ 1 Can unsupervised techniques offer an advantage over supervised techniques in prostate cancer classification when dealing with label scarcity?

RQ 2 Does incorporating IHC stained data increase performance in prostate cancer classification?

RQ 3 Does introducing adversarial/perceptual loss optimisation increase performance in prostate cancer classification?

RQ 4 How well does the performance of each approach transfer to an external biopsy dataset?

2 Related work & background

As mentioned in the introduction, prostate cancer is graded based on morphological features found in tissue samples acquired through biopsy or prostatectomy. Specifically, the architectural patterns in epithelium (brown in IHC) are a big prognostic factor. Some of these patterns showcased in an actual prostatectomy can be seen in Fig. 2. This particular area is considered benign, as the glands are fairly well organised and there is a clear basal cell layer visible around them (cells looking like small black dots). When this layer is not visible and the tissue patterns are more chaotic like in the last three columns of Figs. 3 and 4, the tissue is typically considered tumour. The tissue between the epithelium (light blue in IHC, pink in H&E) is called stroma. Stroma also holds clues as to the grade of cancer that is visible, but the epithelium is the main giveaway.

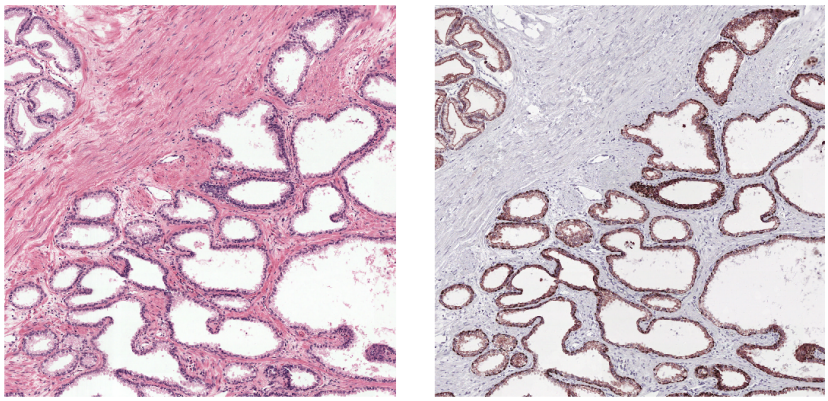

Figure 2: Example of a larger area of benign prostate tissue in both stainings. The glands are well-formed and closely packed. There is a clear basal cell layer surrounding the glandular epithelium, particularly visible in the IHC staining.

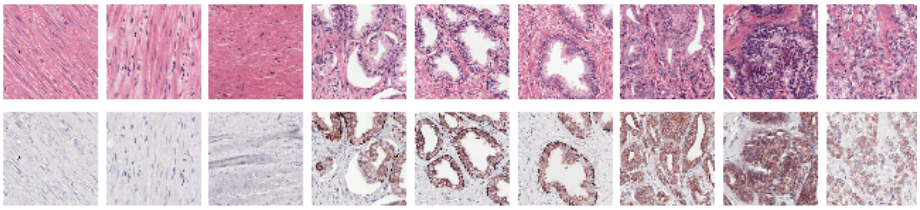

Figure 3: Examples of patches extracted from the prostatectomy WSIs. Columns 1 through 3 are stroma tissue, 4 through 6 are benign tissue and 7 through 9 are tumour tissue.

In the introduction, some publications using deep learning in medical image classification and segmentation were already named. In this thesis, the focus lies mainly with autoencoder networks. Multiple studies have been conducted on using autoencoders to learn *deep features*, that can then be used for classification using a limited amount of labeled data [25, 26, 27, 28, 29, 16]. Autoencoders are neural networks that compress their input into a *latent vector* (embedding) that contains the information to reconstruct the original input as accurately

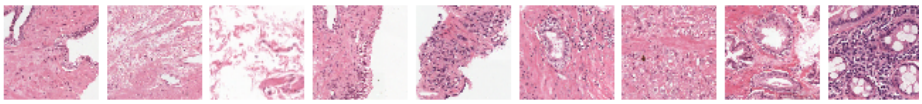

Figure 4: Examples of patches extracted from the prostate biopsy WSIs. Columns 1 through 3 are stroma, 4 through 6 are benign tissue and 7 through 9 are tumour tissue.

as possible. The latent vector then inherently contains information specific to the input that it represents, and can potentially be used as a feature vector for classification problems. Research on structuring this latent feature space into predetermined distributions is prevalent [30, 31, 32] since it allows for reasoning about the internal representation of the models.
As mentioned, this thesis project is partly inspired by [16]. In this paper, Bulten et al. proposed a clustering adversarial autoencoder to classify prostate tissue. This network was trained to automatically cluster prostate tissue, while adversarially constraining its latent space distribution (a variation on the idea of variational autoencoders [30]). Similarly, [28] used an autoencoder style network that simultaneously learns cluster assignments and feature representations.
Autoencoders have been applied in many other medical imaging problems as well. For example, [25] used an autoencoder to learn features from lung computed tomography (CT) images, which were then fed into a decision tree for classification. In [26], a sparse convolutional autoencoder was used to achieve state of the art results on unsupervised nucleus detection in histopathological images. [27] used unsupervised feature learning as part of a larger framework developed for basal cell carcinoma detection, achieving state of the art results. Kallenberg et al. [29] used unsupervised feature learning at multiple scales to replace handcrafted features for use in breast density segmentation and mammographic risk scoring, achieving state of the art results in both tasks.
In problems concerning image data, autoencoders are typically trained to minimise element-wise measures like the squared error, but this type of metric is not very suitable as it does not accurately model errors as perceived by the human visual system. Additionally, reconstructions often end up looking blurred and unrealistic. Various studies have proposed methods to overcome this, for example by using adversarial training or perceptual measures [33, 34].

Finally, some neural network specific techniques are introduced here. The reader is assumed to have a basic understanding of convolutional neural networks. The models in this thesis utilise leaky rectified linear units (ReLU) as activation functions. The regular ReLU function is listed in Eq. (1). Essentially, it cuts off all values less than zero. While not nicely differentiable like e.g. the sigmoid function, this function has proven to work very well in neural network architectures [35, 13].

$$f(x) = \max(0, x) \tag{1}$$

However, Xu et al. showed that it might be beneficial to retain the negative part of the activations in some part as well. In leaky ReLU, the negative part is retained after multiplication with some leak factor α as per Eq. (2).

$$f_\alpha(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases} \tag{2}$$

Another technique that is used often in this thesis is batch normalisation [36]. In batch normalisation, a running average of activations in the network is maintained in order to make sure that activations can be normalised to a zero mean identity variance distribution. This helps with the stability and convergence of the network during training, and in some situations can improve performance. Another technique that helps with stability as well as combatting overfitting and providing regularization is dropout [37]. Dropout randomly switches off neurons in network layers during training, and effectively forces the network to spread out information over multiple neurons which can greatly reduce overfitting tendencies.

3 Methodology

3.1 Autoencoders

Autoencoders are neural networks that consist of an encoder and decoder part. The encoder compresses the input into a latent vector, and the decoder tries to reconstruct the original input from the information contained in this latent vector. Typically, this latent vector is smaller than the input data in order to force high level feature extraction instead of simply learning the identity function. For image data, this is typically accomplished by training the network to minimise the pixel-wise mean squared error (MSE), listed in Eq. (3). Variants such as the variational autoencoder (VAE) will add an additional KL-divergence loss term to restrict the latent space to a Gaussian distribution [30]. In this work, we augmented the MSE loss term by taking its minimal value over all four possible 90° rotations to allow rotated reconstructions, since the orientation of the tissue is not important for the classification task.

$$\mathcal{L} = \frac{1}{n}\sum_{i=0}^{n} t_i - x_i \tag{3}$$

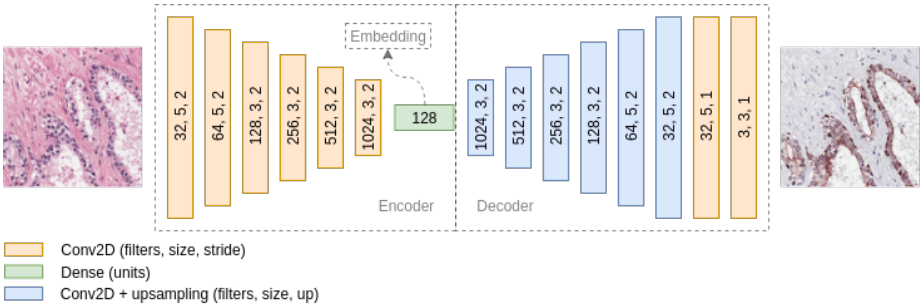

Figure 5: Autoencoder architecture used in this research. All layers use batch normalization and a leaky ReLU activation function, except for the last layer which uses a sigmoid activation.

Several considerations were made to arrive at the final autoencoder architecture, shown in Fig. 5. First, the size of the latent vector should be great enough that it can effectively encode high level tissue features, but small enough that it does not suffer from the curse of dimensionality [38] and can be clustered effectively. Second, the number and size of layers should be great enough that the network has enough expressive capacity, but small enough that it does not impose immense hardware requirements. Finally, the upsampling strategy of the decoder has to be chosen such that it can produce reconstructions of good quality.
In preliminary architecture exploration, various setups were tested starting from networks used in [39, 16] by visually judging their reconstruction fidelity after training. Strided convolution layers were used in the encoder portion of the network to downsample the input image into the latent space. This style of convolutional autoencoder has been shown to work well for feature learning [26, 28, 29]. Once a general setup was established, classification accuracy was tested

by clustering the latent vector sizes $\{128, 256, 512, 1024\}$ on both variational and regular autoencoders. Of these tests, a regular autoencoder with latent vector size 128 reached the best F1 score. There are various choices available for the upsampling strategy, such as fractionally strided convolution (or deconvolution) or regular convolution followed by an interpolated upsampling step. Since deconvolutions often end up producing checkerboard artifacts [40], the latter strategy was chosen. Some extra layers were added after the last upsample step to further smooth out the reconstructions. Batch normalisation [36] was added to stabilise gradients during training.

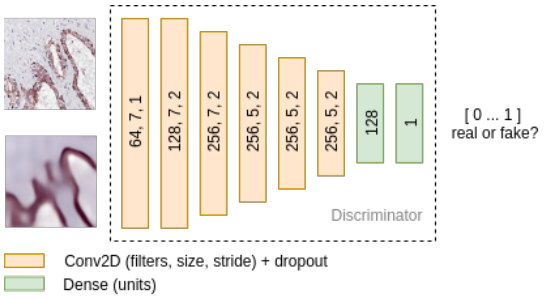

Figure 6: Discriminator architecture used for adversarial finetuning. All layers use batch normalization and a leaky ReLU activation function, except for the last layer which uses a sigmoid activation.

3.2 Adversarial finetuning

Adversarial training effectively tries to find an equilibrium in a min-max game between a *generator* and a *discriminator*. The generator tries to fool the discriminator into classifying generated samples as real samples, while the discriminator learns to discriminate real from fake samples simultaneously. This forces the generator to generate increasingly realistic samples. Intuition dictates that the better the reconstruction quality, the more information is captured in the autoencoder latent space. However, pixel-wise accuracy is not a very good measure for reconstruction quality. It is not robust to translation, and it often results in blurry reconstructions. Especially for a classification task, the reconstructions should ideally capture the same tissue features as the original image, but they do not have to be one hundred percent identical. They should be *perceptually* alike, which ensures that the latent space contains high level perceptual features.

$$\mathcal{L} = -\frac{1}{n}\sum_{i=0}^{n} t_i \dot{\log} x_i + (1 - t_i)\dot{\log}(1 - x_i) \tag{4}$$

In order to implement this idea, an adaptation of the VAE-GAN [34] model was used as inspiration. A discriminator network was chained to an autoencoder pre-trained using MSE as in the previous section. The discriminator takes reconstructions from the autoencoder as well as the original images, and is trained to predict real or fake by optimising a binary cross entropy loss term, listed in Eq. (4). The autoencoder has two loss terms; a perceptual loss and a discriminatory loss. The perceptual loss is calculated by taking the feature output

of the penultimate layer of the discriminator network for both reconstructions and real images, and finding the difference in terms of a Gaussian likelihood estimation. Given real features f_t and reconstruction features f_r, the Gaussian model is defined as $p(f_t) = \mathcal{N}(f_t|f_r, I)$. A complete formal description of the perceptual loss can be found in Sec. 2.3, Eq. 6 and 7 of [34]. The discriminator loss is calculated by taking the binary cross entropy loss of the discriminator classifying the reconstructions as real images.
The network architectures listed in [34] initially led to training instability during preliminary experiments. Overfitting and mode collapse (i.e. only a single identical image is generated, consistently fooling the discriminator) were observed in the discriminator and generator respectively. Both issues can typically be solved through parameter tuning, either within the network architectures or in the training paradigm. To reach stable results, larger convolutional kernels were used to capture more context. This increased field of view in the discriminator helped it deal with the size of images used (four times larger than those in [34]). Additionally, starting off with pre-trained network weights as opposed to training from scratch helped stabilise training as well. The final discriminator architecture is detailed in Fig. 6.

3.3 Data

3.3.1 Sources

This research uses two different data sources, one set of prostatectomy images and another set of biopsy images. Both sets are clinically interesting because of two different reasons. Biopsies are typically the first line of defence. Cancer prognosis is given by looking at biopsy samples, and thus it would be useful to detect cancer in these images automatically. Prostatectomies are graded on cancer for posterity, as well as to determine if any cancer may have spread from the prostate.
The prostatectomy set consists of 102 registered whole-slide image (WSI) pairs from patients that underwent a radical prostatectomy at the Radboud university medical center (Radboudumc) [17]. While prostatectomies are only carried out if there is a reasonable certainty of cancer, they are an interesting target to use for deep learning. They contain a larger amount of much more varied tissue than biopsies. The slides were digitized using a *3DHistech Pannoramic Flash II scanner* at 20× magnification (pixel resolution $0.24\mu m$). The images are stored as multi-resolution TIFF files, which allow for efficient viewing at multiple levels of magnification. Each pair is made up of a single slide that was stained using H&E and subsequently restained using IHC with an epithelial (CK8/18) and basal cell (P63) marker. The slides were divided into 62 training and 40 test slides. From the training slides, 9 slides were held out for validation. All H&E slides are publicly available as the PESO [41] dataset.
The biopsy dataset consists of 100 prostate biopsy WSIs digitised with the same scanner and pixel resolution as the prostatectomy WSIs. These slides are only available with the H&E staining. In order to address RQ 4, this dataset is used to test the generalisation of models trained on the prostatectomy data.

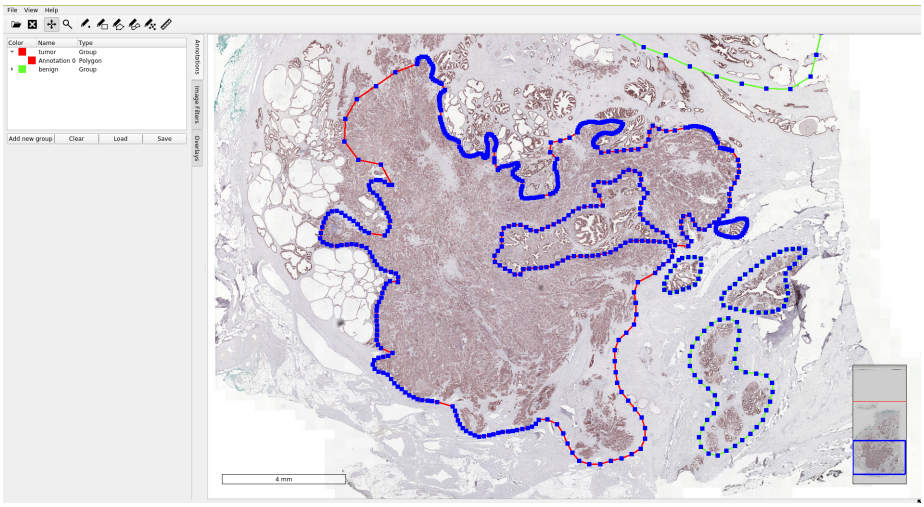

Figure 7: Annotations around malignant (red polygon on the left) and benign (green polygons on the right) tissue in a prostatectomy WSI.

3.3.2 Preparation

To allow sampling from the slides based on tissue class label, each slide was prepared by manually annotating benign and cancerous areas using the Automated Slide Analysis Platform (ASAP) [42] software (Fig. 7). While the test slides were annotated by a pathologist, the training slides were annotated by myself. The annotations are fairly coarse and probably not perfect. These annotations were turned into pixel masks and combined with epithelium segmentation masks [17] to label each slide pixel as stroma, benign epithelium or tumour.

3.3.3 Patch extraction

In order to capture a balance in tissue detail and morphological context while keeping hardware requirements manageable, a patch size of 256×256 at $5\times$ magnification (pixel resolution $0.96\mu m$) was chosen. Two sets of patches were extracted from the prostatectomy WSIs, designated D_r and D_b. D_r is a set of 100.000 randomly sampled patches, used to train the autoencoder networks. D_b is a balanced set of 100.000 patches with a global class ratio of $0.25, 0.25, 0.50$ stroma, benign epithelium and tumour respectively. This ratio was chosen to create a balanced dataset that could still be used to reliably learn the difference between the three classes in question. The class of a patch was chosen by its center-pixel label, e.g. if the pixel in the center of the patch was labeled as tumour in the accompanying segmentation mask, the whole patch was labeled tumour. For the test sets, 10.000 patches from both the biopsy and prostatectomy WSIs are extracted with the same ratio. To sample patches by class, all WSI pixels are grouped based on their label at $1.25\times$ magnification. Random pixels are chosen from each respective group and used to sample patches at the proper magnification.

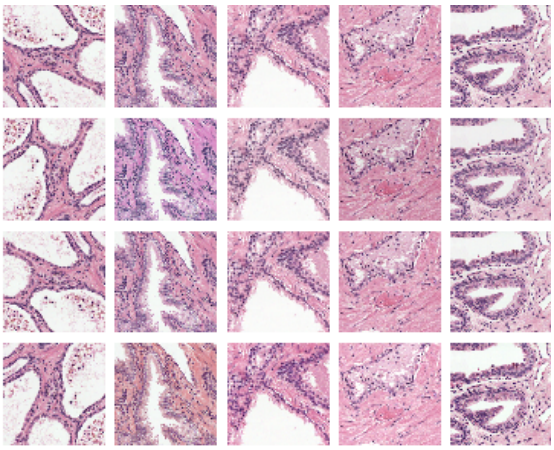

Figure 8: Examples of data augmentation applied to H&E patches. Columns from left to right: flipping, hue, saturation, brightness, contrast.

3.3.4 Data augmentation

A simple data augmentation procedure was adopted to introduce more morphological patterns and stain variance. The augmentations applied during training consist of flipping horizontally/vertically as well as hue, saturation, brightness and contrast adjustments. The range of these adjustments was initially determined by visual inspection, and is shown in Fig. 8. The ranges of augmentation were further finetuned by training models with different settings, and picking the setup that resulted in the lowest validation MSE. The colour augmentation is relatively mild, as more extreme augmentation ranges seemed to negatively affect performance. Additionally, the colour augmentations were only applied on the input patches; the reconstruction target patches were left as is.

3.4 Hardware & software

DIAG maintains a small deep learning cluster[1] with approximately 80 NVidia GPUs (GTX 1080, RTX 2080 Ti and a few TITAN cards) that was available to schedule and run experiments on using Docker and Slurm [43]. All experiments were implemented in Python using TensorFlow [44], making full use of GPU acceleration. The available memory in these cards places restrictions on the maximum size of the networks used, as well as parameters like batch size. While no problems were encountered with network size, batch size was set depending on the type of card running the experiment.

[1] https://rse.diagnijmegen.nl/software/sol (visited on: 08-29-2019)

4 Experiments

As mentioned before, three main approaches are compared. Each one is detailed more in depth in its respective subsection, but the general setup of the pipeline can be found in Fig. 9. In every approach, the model is trained to predict all three classes (stroma, benign epithelium and tumour). However, only the F1 score for tumour versus no tumour binary classification is reported. The reason that models are trained to predict three classes is that it forces the models to learn the difference between benign epithelium and tumour tissue in particular, which is advantageous when performing the binary classification task. These three class models can additionally be used to further investigate their quality through application in a segmentation task.

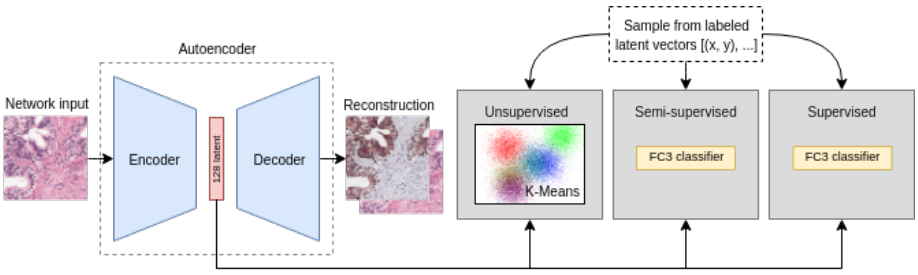

Figure 9: The flow of data for each of the three methods. Note that the data flow for the supervised method is identical to that of the semi-supervised method, but does not utilize unsupervised pre-training.

A varying number of labeled patches is used to assign labels (unsupervised) or train the classifier portion of the approach ((semi-) supervised). Since manual prostate cancer grading is done on H&E data, all classifiers take H&E patches as input, except for the supervised network trained on IHC data (included for completeness). The numbers of labeled patches were chosen through some preliminary experiments, and can be found in the tables in Section 5. In those tables, the results obtained using autoencoders that were further finetuned using the learned perceptual measure are included as well.

4.1 Unsupervised

In this approach, no labels or balanced data are used at all during training and fitting. An autoencoder model M is trained on D_r to reconstruct either H&E or IHC, and the latent vectors $M(D_r)$ are used to fit a k-means classifier. Labeled data is used to assign classes to clusters through majority voting, by predicting labels for a certain number of patches from D_b. If a certain cluster does not get any patches during label assignment, it is assigned the stroma class. This is just a technicality; it almost never happens and does not significantly skew results. To validate the model, clusters are predicted for the test set and each patch is assigned the label corresponding with that cluster. The number of clusters used in the experiments is 50, as increasing this number even more leads to diminishing returns in classification performance. This number was found experimentally.

4.2 Semi-supervised

Similar to the unsupervised approach, an autoencoder model M is trained on D_r to reconstruct either H&E or IHC. However, instead of fitting a k-means classifier, a single layer neural network classifier is trained on subsets of labeled patches from D_b by chaining it to the (frozen) encoder part of M. Only the weights for the classification layer are updated, but it is trained including data augmentation, contrary to the unsupervised model. This approach offers some additional expressivity over the unsupervised method, and is trained to extract information from the autoencoder embeddings that is relevant to the specific class. To validate the model, the test patches get assigned a label by feeding them through the encoder and classification layer. In this approach, we use some labeled data during training, but the main part of the model is still trained completely without labels.

4.3 Supervised

The supervised approach uses the same network architecture as the semi-supervised model, but the whole network is trained in an end-to-end fashion (e.g. no pre-training of the encoder). This effectively corresponds to training a simple CNN classifier. The supervised approach is used to illustrate the full expressive capacity of the chosen architecture when all labeled data is used, and to show the advantage of semi- and unsupervised methods when the amount of available labeled data is limited. This network is only trained on a certain number of labeled patches from D_b.

4.4 Training specifics

All networks are trained using the Adam optimiser shipped with Tensorflow. Different optimisers were tested (SGD, RMSprop and Adam), but Adam showed the quickest convergence. When optimising MSE, a learning rate of 10^{-3} resulted in the best reconstructions. During the finetuning training, two separate Adam optimisers are used. The autoencoder optimizer utilised a learning rate of 10^{-4}, and the discriminator optimizer utilised a learning rate of 10^{-5} in order to not surpass the autoencoder. Other parameters were left as default.
In order to track training progress, Tensorboard was used to plot loss values and store validation reconstructions after each epoch. This allowed for quick judgement in regard to overfitting, mode collapse or other issues. Other values of particular interest were the discriminator classification accuracy and classifier prediction accuracy, as these seemed to correlate with the model's stability.

4.5 Additional investigation

Besides comparing the aforementioned methods on their classification performance, it is also interesting to investigate possible explanations for their differences. The models are applied to larger tissue areas as a sliding window classifier to obtain a visual indication of their ability to correctly identify all three classes. One benign area and one tumour area were selected, in order to get an idea of the performance in different contexts. Additionally, t-SNE plots are provided to visualise the autoencoder latent space. In these plots, a clear

clustering of tissue classes should at least indicate the usefulness of the learned features.

5 Results

5.1 Lower bound performance

In order to get an idea of what F1 score should be easily achievable on the given test sets, it is useful to look at the ratio of the test set labels. In the binary classification problem, there is a 50/50 label distribution: stroma and benign epithelium are labeled as 0 (not cancer, amounting to 50% of the test set), while tumour is labeled as 1. Of these classes, benign epithelium and tumour look very much alike while stroma is very different. Let there be a classifier that can, by looking at the center-pixel, flawlessly classify stroma patches, but cannot distinguish benign epithelium from tumour. This classifier would reach an F1 score of 0.8 with the used data distribution.

$$\begin{aligned}
\text{real labels} &= \{0, 0, 1, 1\} \\
\text{predictions} &= \{0, 1, 1, 1\} \\
\text{F1 score} &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \\
&= 2 \cdot \frac{\frac{2}{3} \cdot 1}{\frac{2}{3} + 1} \\
&= 0.8
\end{aligned}$$

In the more likely scenario that the classifier was unable to distinguish between benign epithelium and tumour and classified these patches randomly as either cancer or not cancer, it would however reach an average F1 score of ~ 0.52. This is the lower bound value that is realistic, since each model is trained to predict three classes. While stroma is easily identifiable from the other two classes, the distinction between benign epithelium and tumour is difficult for uropathologists as well.

5.2 Supervised

The obtained F1 scores for the supervised method are listed in Table 1. As expected, its performance on both test sets increases with the available number of labeled patches. When trained on H&E patches, this approach obtains a maximum F1 score of 0.88 on the prostatectomy test set and 0.74 on the biopsy test set. When training solely on IHC data, a maximum F1 score of 0.91 is reached. This suggests that the IHC data is indeed more informative than the H&E data. However, more IHC data is needed for a robust average; the standard deviations at 2000 and 10.000 labeled patches are very large.
On the biopsy test set, large standard deviations are reported along the whole range of labeled patches. This indicates that features learned from the prostatectomy data do not transfer very well to the external biopsy test set with this approach (i.e. overfitting).

5.3 Unsupervised

The F1 scores obtained by the completely unsupervised approach are listed in Table 2. Without adversarial finetuning, this method reaches an F1 score of 0.71

	Biopsy	Prostatectomy	
NLP	**H&E**	**H&E**	**IHC**
100	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
500	0.40 ± 0.16	0.58 ± 0.12	$\mathbf{0.67 \pm 0.09}$
1000	0.58 ± 0.10	0.76 ± 0.01	$\mathbf{0.77 \pm 0.02}$
2000	0.54 ± 0.11	$\mathbf{0.73 \pm 0.03}$	0.56 ± 0.27
10.000	0.50 ± 0.15	$\mathbf{0.74 \pm 0.04}$	0.71 ± 0.29
100.000	0.74 ± 0.09	0.88 ± 0.02	$\mathbf{0.91 \pm 0.02}$

Table 1: F1 score of the supervised network on the available stainings of both datasets (NLP = number of labelled patches).

	Biopsy		Prostatectomy	
NLP	**H&E →H&E**	**H&E →IHC**	**H&E →H&E**	**H&E →IHC**
	Regular		**Regular**	
100	0.53 ± 0.02	$\mathbf{0.69 \pm 0.03}$	0.71 ± 0.02	0.71 ± 0.02
500	0.63 ± 0.02	$\mathbf{0.68 \pm 0.03}$	0.72 ± 0.02	$\mathbf{0.74 \pm 0.01}$
1000	0.62 ± 0.02	$\mathbf{0.67 \pm 0.05}$	0.72 ± 0.01	$\mathbf{0.74 \pm 0.01}$
2000	0.63 ± 0.02	$\mathbf{0.67 \pm 0.06}$	0.74 ± 0.02	$\mathbf{0.75 \pm 0.00}$
10.000	0.64 ± 0.02	$\mathbf{0.68 \pm 0.03}$	0.70 ± 0.01	$\mathbf{0.74 \pm 0.01}$
100.000	0.49 ± 0.03	$\mathbf{0.64 \pm 0.02}$	0.73 ± 0.01	$\mathbf{0.75 \pm 0.01}$
	Finetuned		**Finetuned**	
100	0.56 ± 0.06	$\mathbf{0.57 \pm 0.09}$	0.63 ± 0.05	$\mathbf{0.64 \pm 0.07}$
500	$\mathbf{0.63 \pm 0.03}$	0.60 ± 0.05	0.69 ± 0.01	$\mathbf{0.71 \pm 0.01}$
1000	$\mathbf{0.65 \pm 0.03}$	0.60 ± 0.04	0.70 ± 0.02	$\mathbf{0.73 \pm 0.01}$
2000	$\mathbf{0.62 \pm 0.03}$	0.61 ± 0.04	0.71 ± 0.02	$\mathbf{0.72 \pm 0.01}$
10.000	$\mathbf{0.64 \pm 0.03}$	0.63 ± 0.05	0.71 ± 0.02	$\mathbf{0.72 \pm 0.01}$
100.000				

Table 2: F1 score of the unsupervised clustering approach on both datasets with 50 clusters (NLP = number of labelled patches).

on the prostatectomy test set for both autoencoder reconstruction targets with only 100 labeled patches. As the number of labeled patches increases, there is a divergence in the performance when using different reconstruction targets. For example at 1000 labeled patches, the IHC reconstruction target results in an F1 score of 0.74 versus 0.72 with the H&E reconstruction target. As the number of labeled patches increases further, the maximum F1 score does not improve, indicating an upper bound plateau for this approach.

On the biopsy test set, the same trend is visible albeit the absolute F1 scores are lower. This is to be expected due to the fact that this data comes from a different source. At 100 labeled patches, there is a big difference in F1 score between the reconstruction targets, H&E resulting in 0.53 and IHC resulting in 0.69. This gap closes as more labeled patches are available, but the IHC reconstruction target results in overall higher F1 scores. The same upper bound plateau is visible here as well, with scores not rising above 0.69 at all.

With adversarial finetuning, the difference in performance between reconstruc-

	Biopsy		Prostatectomy	
NLP	**H&E →H&E**	**H&E →IHC**	**H&E →H&E**	**H&E →IHC**
	Regular		**Regular**	
100	0.48 ± 0.06	$\mathbf{0.60 \pm 0.04}$	0.56 ± 0.07	$\mathbf{0.69 \pm 0.04}$
500	0.52 ± 0.03	$\mathbf{0.64 \pm 0.06}$	0.70 ± 0.03	$\mathbf{0.75 \pm 0.01}$
1000	0.49 ± 0.02	$\mathbf{0.61 \pm 0.05}$	0.73 ± 0.01	$\mathbf{0.77 \pm 0.01}$
2000	0.49 ± 0.03	$\mathbf{0.59 \pm 0.04}$	0.74 ± 0.01	$\mathbf{0.77 \pm 0.01}$
10.000	0.50 ± 0.02	$\mathbf{0.58 \pm 0.03}$	0.76 ± 0.01	$\mathbf{0.78 \pm 0.01}$
100.000	0.48 ± 0.01	$\mathbf{0.58 \pm 0.03}$	$\mathbf{0.75 \pm 0.02}$	0.74 ± 0.02
	Finetuned		**Finetuned**	
100	$\mathbf{0.64 \pm 0.04}$	0.48 ± 0.25	$\mathbf{0.64 \pm 0.04}$	0.53 ± 0.27
500	0.62 ± 0.04	$\mathbf{0.65 \pm 0.05}$	0.69 ± 0.03	$\mathbf{0.71 \pm 0.03}$
1000	$\mathbf{0.65 \pm 0.03}$	0.63 ± 0.07	0.70 ± 0.04	$\mathbf{0.74 \pm 0.02}$
2000	0.50 ± 0.25	$\mathbf{0.63 \pm 0.06}$	0.58 ± 0.29	$\mathbf{0.73 \pm 0.05}$
10.000	$\mathbf{0.60 \pm 0.04}$	0.40 ± 0.33	$\mathbf{0.74 \pm 0.02}$	0.45 ± 0.37
100.000	$\mathbf{0.60 \pm 0.03}$	0.57 ± 0.08	$\mathbf{0.75 \pm 0.02}$	$\mathbf{0.75 \pm 0.02}$

Table 3: F1 score of the semi-supervised networks on both datasets (NLP = number of labelled patches).

tion targets seems to mostly disappear. On the prostatectomy test set, the scores are comparable to those achieved without finetuning, if slightly lower. On the biopsy test set, the H&E reconstruction target actually results in better F1 scores. For example, at 1000 labeled patches the classifier reaches an F1 score of 0.65 with the H&E reconstruction target as opposed to 0.60 with IHC. The upper bound plateau can again be observed with both test sets.

5.4 Semi-supervised

The F1 scores obtained by the semi-supervised approach are listed in Table 3. While an upper bound plateau can also be observed with this method, the added expressivity results in better F1 scores when compared to those of the unsupervised method. Without adversarial finetuning, this approach reaches a maximum F1 score of 0.75 on the prostatectomy test set using 500 labeled patches. As the number of labeled patches increases, the F1 score climbs to 0.78 with 10.000 labeled patches, indicating that this method benefits from more labeled data. The difference in F1 scores using different reconstruction targets is more emphasised here as compared to the unsupervised approach. For example, at 1000 labeled patches the IHC reconstruction target results in an F1 score of 0.77 as opposed to 0.73 with the H&E reconstruction target.
On the biopsy test set, the F1 scores obtained using the IHC reconstruction target are also higher. However, the absolute scores here are lower than those achieved by the unsupervised method. The scores decrease when many labeled patches are used. This showcases the fact that this approach ends up focusing on prostatectomy data specifics to a greater extent.
When adversarial finetuning is introduced, there is a decrease in F1 scores across the board. On the prostatectomy test set, F1 scores are generally greater when using the IHC reconstruction target up to 2000 labeled patches. Of particular

interest here are the large standard deviations found at 100, 2000 and 10.000 patches. These are the result of runs that produced a NaN F1 score, which was reported as 0. NaN scores are the result of a zero division that occurs when the precision and recall are both 0, which happened in two instances for each of the aforementioned results. Since the applied models were identical for the biopsy test set, the large standard deviations occur at the same points.
Interestingly, the models with adversarial finetuning perform better on the biopsy test set than those without when the H&E reconstruction target is used. This trend was not visible in the unsupervised method results. The difference between reconstruction targets is not readily observable here, as the maximum F1 score pivots between H&E and IHC reconstruction targets for various numbers of labeled patches.

5.5 Comparison of methods

In order to get a better idea of the difference in performance between the tested methods, line plots of the F1 scores are provided as well. All these plots are created from the same numbers listed in Tables 1 to 3. For the prostatectomy test set, Figs. 10 and 11 both show the advantage of the semi- and unsupervised methods over the supervised approach when few labeled data is available.

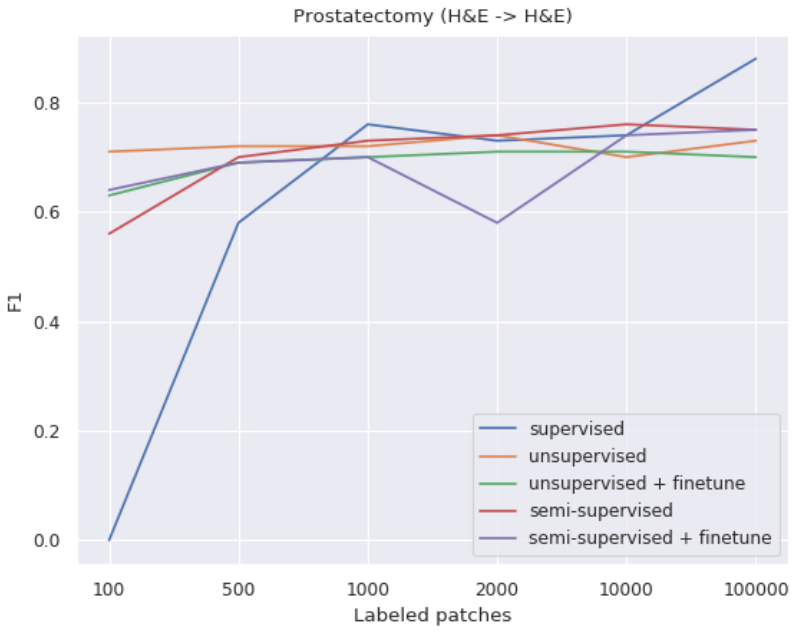

Figure 10: F1 scores of all methods on the prostatectomy test set. Autoencoders were trained to reconstruct H&E → H&E.

On the biopsy test set, the same trends are visible in Figs. 12 and 13.
Additionally, each approach is applied to a larger prostatectomy tissue area as a sliding window classifier. One benign region and one tumour region were selected in order to evaluate the classifier accuracy to give a more complete idea

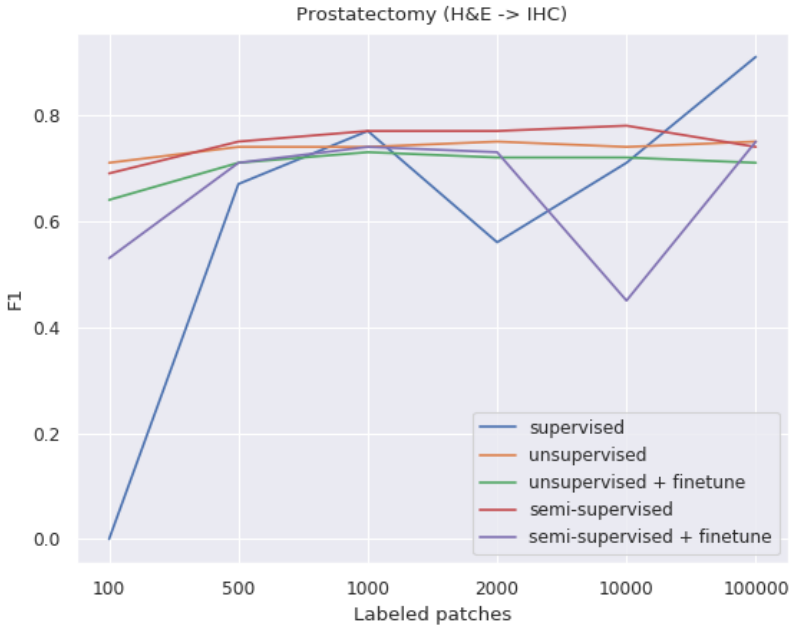

Figure 11: F1 scores of all methods on the prostatectomy test set. Autoencoders were trained to reconstruct H&E → IHC.

Figure 12: F1 scores of all methods on the biopsy test set. Autoencoders were trained to reconstruct H&E → H&E.

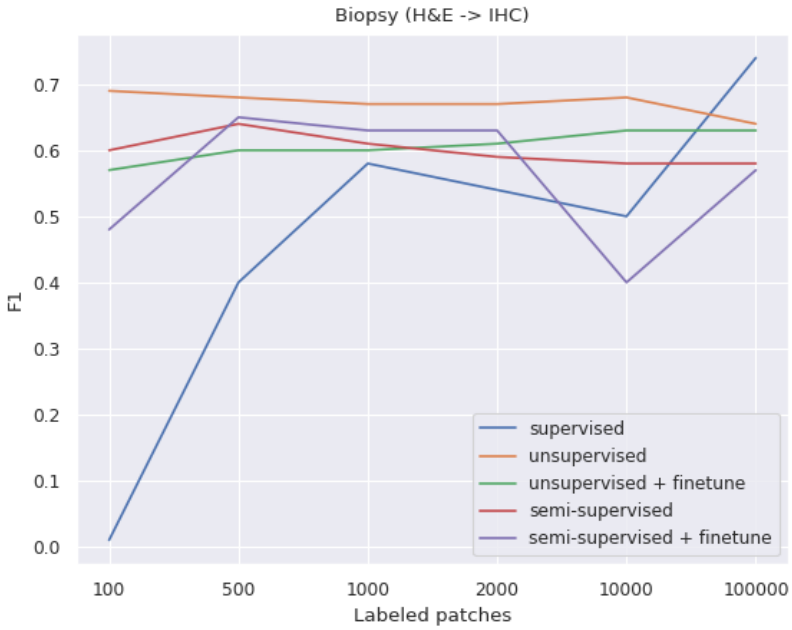

Figure 13: F1 scores of all methods on the biopsy test set. Autoencoders were trained to reconstruct H&E → IHC.

of their quality. Each classifier was selected at 1000 labeled patches, since that is the number where the F1 scores approximately intersect. The benign tissue classification maps can be found in Fig. 14, and the tumour tissue classification map in Fig. 15.
In the benign classification maps, the unsupervised approach is fairly coarse and incorrect. However, the semi-supervised approach is actually more accurate than the supervised approach (note that there should not be any tumour labels at all). This indicates that the unsupervised learning part of the semi-supervised method provides additional useful information.
In the tumour classification maps, the difference between the unsupervised method and the other two methods is less obvious. However, the tissue structure is much more easily visible in the semi-supervised and supervised maps.
These results, excluding the adversarial finetuning portion, were accepted as an extended abstract for the MIDL 2019 conference [45].

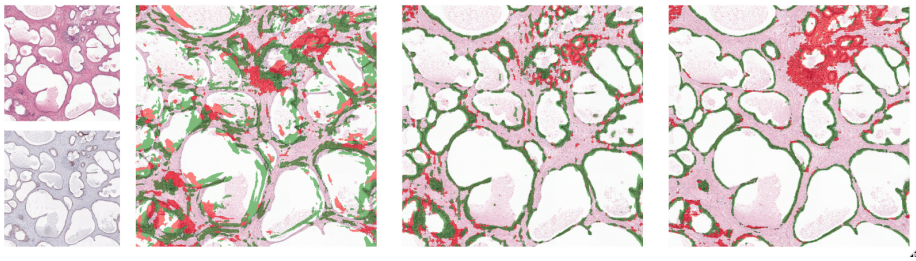

Figure 14: Classification maps of models trained with 1000 labelled patches applied to a benign tissue region (transparent = stroma, green = benign epithelium, red = tumour). From left to right: H&E/IHC original, supervised, semi-supervised, unsupervised.

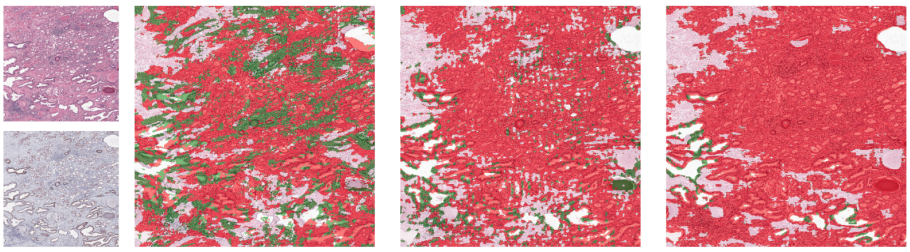

Figure 15: Classification maps of models trained with 1000 labelled patches applied to a tumour tissue region (transparent = stroma, green = benign epithelium, red = tumour). From left to right: H&E/IHC original, supervised, semi-supervised, unsupervised.

6 Discussion

6.1 Data

While the amount of data available for this project is very large, the data is still fairly heterogenous. Artifacts are introduced during staining or scanning. The baseline masks (background masks, epithelium segmentation masks) that were used in this thesis were generated by machine learning models developed in house. Additional annotations of benign and tumour regions were provided by a uropathologist in the test slides only; the training and validation slides were annotated by myself and checked by my daily supervisor (Wouter Bulten). This means that additional mistakes were very likely introduced, and this probably influences the final performance of the models, especially since so few labeled data is used to relate the features learned without supervision to class labels. In particular at low numbers of labeled patches, like 100 or 500, mistakes could have a big impact. Ideally, the labeled patches would be perfect, but perfect labeling is one of the big challenges of this particular problem.

6.2 Model performance

In order to further inspect the internals of the autoencoders used for the semi- and unsupervised methods, Fig. 16 shows 2D visualisations of the embeddings created after training to reconstruct H&E or IHC data using regular MSE optimisation.

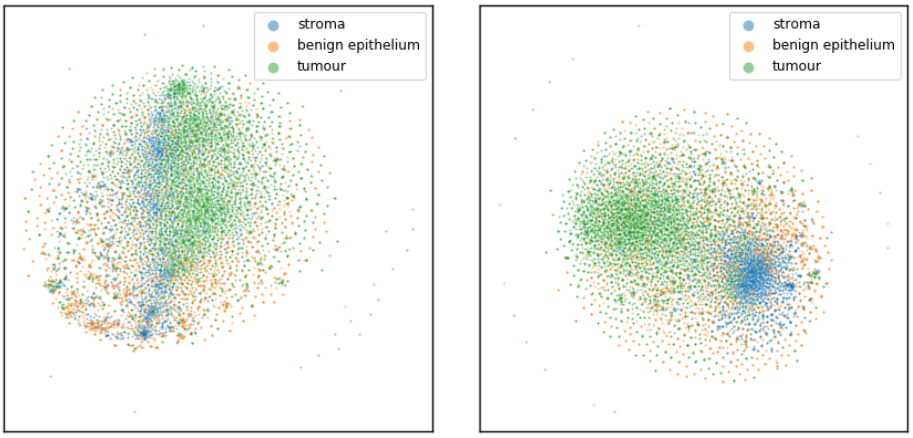

Figure 16: t-SNE visualization of H&E → H&E (left) and H&E → IHC (right) autoencoder embeddings.

For the H&E reconstruction target, there is some clustering visible, but the clusters are mostly localised even within a single class. Ideally, all the samples of one class would be assigned to the same cluster; this would make it easier for a classifier to assign samples to a certain class. Such global clustering is much more apparent in the plot created from autoencoders that were trained to reconstruct IHC. This suggests that the features learned by training to reconstruct IHC are more indicative of the class of a respective sample.

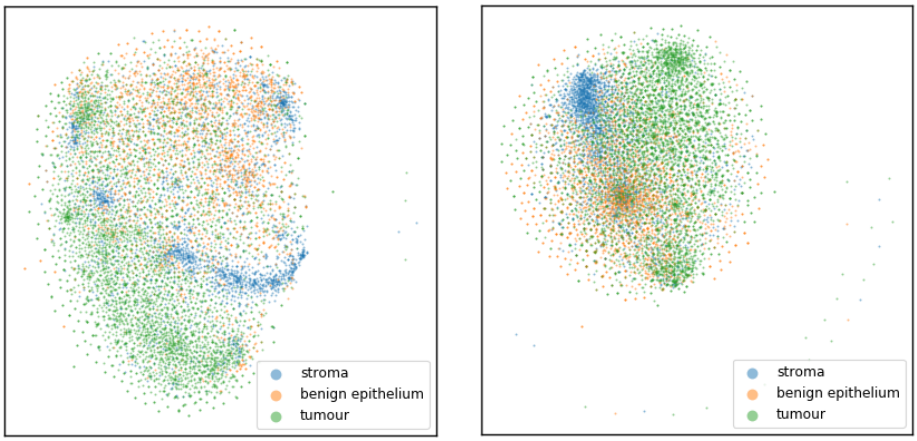

Figure 17: t-SNE visualization of H&E → H&E (left) and H&E → IHC (right) finetuned autoencoder embeddings.

In the plots in Fig. 17 that were created with adversarially finetuned autoencoders, the global clustering in particular occurs to a lesser degree than with the regular autoencoders. This might be considered a partial explanation of why the adversarial finetuning step seems to decrease the model's performance. However, it is important to note that because of the flexibility of the t-SNE algorithm, it can in principle be used to display *structure* (or lack of structure) for almost any dataset[46]. While the difference in the t-SNE plots confirms the difference in the tables in Section 5, it is not a final conclusion about the quality of the latent space of the autoencoders.

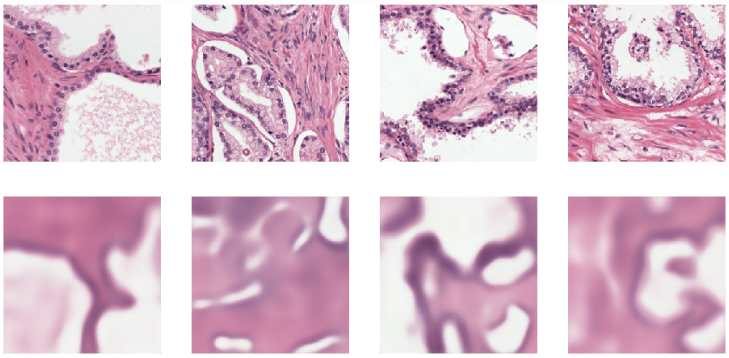

Figure 18: Reconstructions of regular autoencoder trained to reconstruct H&E (top = original, bottom = reconstruction).

Arguably it is more useful to look at the reconstruction quality of the autoencoders. As mentioned before, intuitively the autoencoder latent space quality correlates with the reconstruction quality; the better the reconstruction, the more information the latent space holds. However, there are some important points to consider here. One of the benefits of autoencoders is that data can be compressed to latent space with a much smaller size. This is inherently a lossy

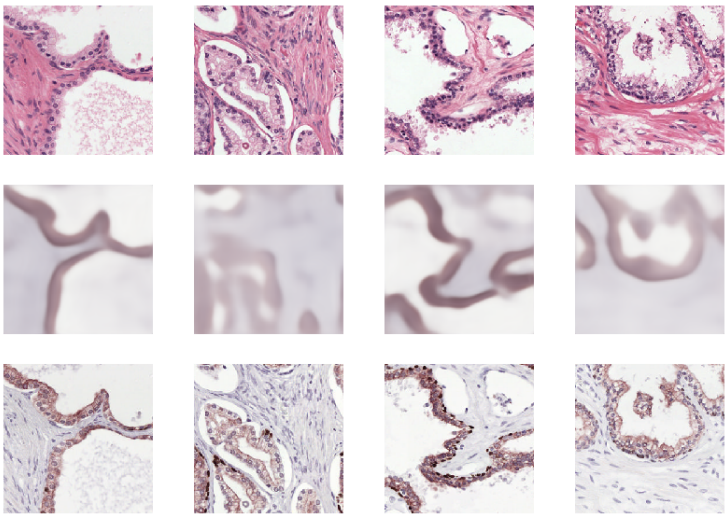

Figure 19: Reconstructions of regular autoencoder trained to reconstruct IHC (top = original, middle = reconstruction, bottom = target).

operation, which forces the model to encode the data as *high level features*. The larger the latent space is, the more information can be stored there and thus the better the reconstructions will be; it however decreases the need for high level feature encoding, since the latent space starts to act almost as an index into all the previously seen data. This means that aside from running into the curse of dimensionality, the features might not be as suitable for a classifier to use because they do not contain the necessary higher level information.
Some example reconstructions of the autoencoders trained to reconstruct H&E or IHC are shown in Figs. 18 and 19 respectively. The reconstructions are fairly blurry and they do not clearly show the (lack of a) basal cell layer. However, the general structure of the epithelium and stroma is easy to see, and the colours approximately line up for both reconstruction targets.

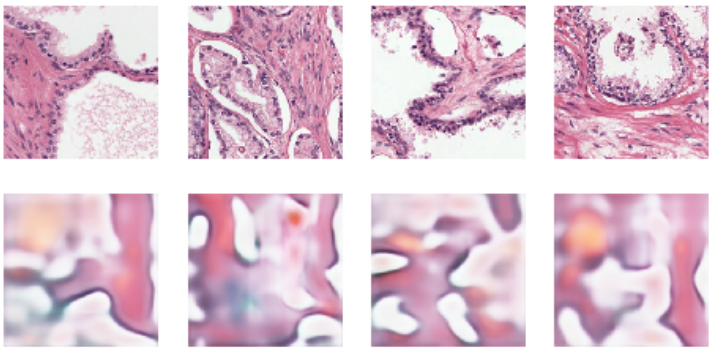

Figure 20: Reconstructions of finetuned autoencoder trained to reconstruct H&E (top = original, bottom = reconstruction).

For the autoencoders that were adversarially finetuned, the H&E and IHC recon-

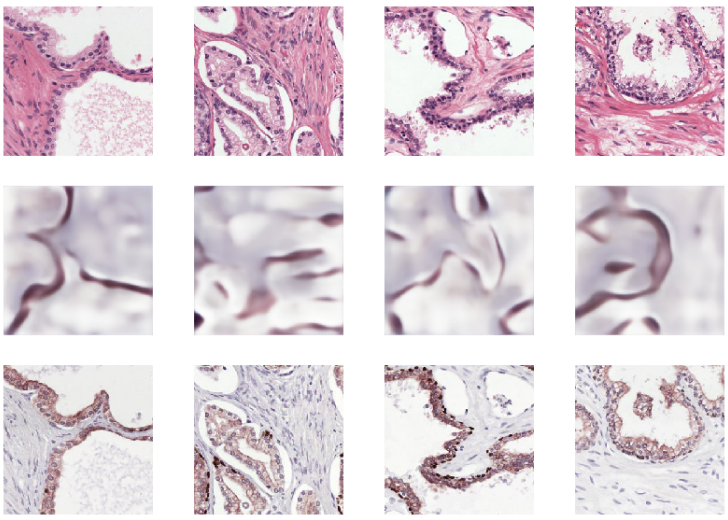

Figure 21: Reconstructions of finetuned autoencoder trained to reconstruct IHC (top = original, middle = reconstruction, bottom = target).

structions are shown in Figs. 20 and 21 respectively. In the H&E reconstructions, weird colour artifacts are visible and the reconstructions do not look like the intended target colour-wise. The main structure of the tissue can be followed for both reconstruction targets, but the quality of the reconstructions is less realistic than that of those obtained without adversarial finetuning. This indicates more definitively that the latent space representations are of lesser quality in the finetuned models. It also explains why the classification performance is worse overall for these models.

6.3 Adversarial finetuning

Getting adversarial finetuning to work without mode collapse or other artifacts proved very difficult. The original authors of the VAE-GAN paper also said that "the devil is in the details" which turned out to be very true. Mode collapse is a common phenomenon in general adversarial networks, where the generator network (the autoencoder in this research) only generates one specific image. The key to get to a result without mode collapse was to tweak learning rates such that the discriminator stayed around 70% classification accuracy, and making sure that the perceptual loss term decreased steadily. Other tricks that proved useful[2] were label flipping, using two optimisers (one for the autoencoder and one for the discriminator), and adding some dropout to the discriminator. The kernel size of the discriminator also had a big impact; too small sizes (i.e. $3x3$, $4x4$) resulted in mode collapse.
The implementation of the adversarial training in this research also has two glaring differences from the VAE-GAN paper. First, the latent space distribution restriction was removed, since the autoencoder used in this thesis did not utilise it either and it seemed fair to keep it identical in that way. Second, the

[2] https://github.com/soumith/ganhacks (visited on: 08-29-2019)

paper used $64x64$ images and a 2048 latent space size. This means that the model has to encode much less data into a significantly larger latent space. For comparison, the network in this thesis compresses images with a factor of 1536 whereas the VAE-GAN from the paper only compresses with a factor of 6. While it seemed logical at first that the latent space held more information than the reconstruction showed, perhaps being held back by the training regimen, this may have caused some issues after all.

6.4 Future work

While this thesis provides data that points out some advantages of unsupervised learning, many open questions need further investigation. For example, the upper bound performance plateau of the semi- and unsupervised techniques would be interesting to look into; is the problem in the quality of the latent space, the network architecture itself, or the simple classifiers used? Other clustering algorithms or more expressive classifiers might result in a better performance without changing anything else. A more extensive architecture and parameter exploration might result in better performance as well.
With regard to adversarial training, there is much to be explored as well. Different data sizes, different training regiments and different architectural features might offer a great improvement over what this thesis has established.
It would also be interesting to extend the deep learning approaches explored in this thesis to be used for actual Gleason grading. This is a more difficult task than simple cancer versus no cancer classification, but it would provide more clinical usefulness since the Gleason grade given by uropathologists is a more detailed prognosis that indicates survival rates and further treatment. This is already being worked on at DIAG, and it is definitely an interesting research subject.
Finally, the way cancer grading is performed by uropathologists might serve as an inspiration for future deep learning systems. For example, one might utilise tissue data at multiple resolutions to capture a greater balance of large scale context as well as low scale detail. This might even allow moving on from patch based classification and grading complete WSIs, where the lower scale data provides localisation information while the higher scale provides a global image of the extensiveness of any present cancer.

7 Conclusion

In conclusion, this thesis provides results that can confidently answer three out of the four posed research questions.

RQ 1 The unsupervised methods tested in this research definitely provide a clear advantage over a completely supervised approach in a situation with few labeled data. Completely unsupervised deep learning still extracts useful features that can be used to classify, in this case, prostate tissue.

RQ 2 The inclusion of an additional IHC tissue staining does provide a good increase in ultimate classification performance over simply using H&E only. Higher F1 scores on the cancer classification task were observed, and the investigation of the latent space seemed to corroborate the suggestion that the features learned using both IHC and H&E are more informative.

RQ 3 Adversarial finetuning does not seem to provide an advantage in this research. An increased performance on the external dataset is visible in certain experiments, but this does not give a definitive conclusion on the usefulness of this technique in this specific context. Several problems and possible explanations were listed in Section 6. Unfortunately it is not possible to conclude for certain that adversarial training does or does not provide an additional advantage as opposed to conventional training.

RQ 4 In general, the same trends observed on the prostatectomy test set between various methods were also visible on the biopsy test set. While the absolute classification scores were lower, this does bode well for the transferability of the features learned using unsupervised techniques. In most cases, the performance on the biopsy test set was well above lower bound.
In this thesis, a clear benefit of unsupervised techniques was shown for the particular use case. With additional improvements on the tested methods, a system could be developed to aid uropathologist in the grading of prostate tissue samples.

References

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012", *CA: a cancer journal for clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[3] A. Stangelberger, M. Waldert, and B. Djavan, "Prostate cancer in elderly men", *Reviews in urology*, vol. 10, no. 2, p. 111, 2008.

[4] P. A. Humphrey, "Gleason grading and prognostic factors in carcinoma of the prostate", *Modern pathology*, vol. 17, no. 3, p. 292, 2004.

[5] J. I. Epstein, W. C. Allsbrook Jr, M. B. Amin, L. L. Egevad, and ISUP Grading Committee, "The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma ", *The American journal of surgical pathology*, vol. 29, no. 9, pp. 1228–1242, 2005.

[6] D. F. Gleason, "Histologic grading of prostate cancer: A perspective", *Human pathology*, vol. 23, no. 3, pp. 273–279, 1992.

[7] D. G. Bostwick and M. K. Brawer, "Prostatic intra-epithelial neoplasia and early invasion in prostate cancer", *Cancer*, vol. 59, no. 4, pp. 788–794, 1987.

[8] R. Totten, M. Heinemann, P. Hudson, E. Sproul, and A. Stout, "Microscopic differential diagnosis of latent carcinoma of prostate.", *AMA archives of pathology*, vol. 55, no. 2, p. 131, 1953.

[9] W. C. Allsbrook Jr, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein, "Interobserver reproducibility of gleason grading of prostatic carcinoma: General pathologist", *Human pathology*, vol. 32, no. 1, pp. 81–88, 2001.

[10] G. D. Carlson, C. B. Calvanese, H. Kahane, and J. I. Epstein, "Accuracy of biopsy gleason scores from a large uropathology laboratory: Use of a diagnostic protocol to minimize observer variability", *Urology*, vol. 51, no. 4, pp. 525–529, 1998.

[11] R. J. Babaian and W. A. Grunow, "Reliability of gleason grading system in comparing prostate biopsies with total prostatectomy specimens", *Urology*, vol. 25, no. 6, pp. 564–567, 1985.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition", *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[15] A. Gertych, N. Ing, Z. Ma, T. J. Fuchs, S. Salman, S. Mohanty, S. Bhele, A. Velásquez-Vacca, M. B. Amin, and B. S. Knudsen, "Machine learning approaches to analyze histological images of tissues from radical prostatectomies", *Computerized Medical Imaging and Graphics*, vol. 46, pp. 197–208, 2015.

[16] W. Bulten and G. Litjens. (2018). Unsupervised prostate cancer detection on H&E using convolutional adversarial autoencoders, [Online]. Available: https://arxiv.org/abs/1804.07098 (visited on 08/28/2019).

[17] W. Bulten, P. Bándi, J. Hoven, R. van de Loo, J. Lotz, N. Weiss, J. van der Laak, B. van Ginneken, C. Hulsbergen-van de Kaa, and G. Litjens, "Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard", *Scientific reports*, vol. 9, no. 1, p. 864, 2019.

[18] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis", *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[19] E. Heim, T. Roß, A. Seitel, K. März, B. Stieltjes, M. Eisenmann, J. Lebert, J. Metzger, G. Sommer, A. W. Sauter, F. R. Schwartz, A. Termer, F. Wagner, H. G. Kenngott, and L. Maier-Hein, "Large-scale medical image annotation with crowd-powered algorithms", *Journal of Medical Imaging*, vol. 5, no. 3, p. 034 002, 2018.

[20] M. Rajchl, M. C. Lee, F. Schrans, A. Davidson, J. Passerat-Palmbach, G. Tarroni, A. Alansary, O. Oktay, B. Kainz, and D. Rueckert, "Learning under distributed weak supervision", *arXiv preprint arXiv:1606.01100*, 2016.

[21] J.-M. Bokhorst, H. Pinckaers, P. van Zwam, I. Nagtegaal, J. van der Laak, and F. Ciompi, "Learning from sparsely annotated data for semantic segmentation in histopathology images", 2018.

[22] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation", in *International conference on medical image computing and computer-assisted intervention*, Springer, 2016, pp. 424–432.

[23] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do CIFAR-10 Classifiers Generalize to CIFAR-10?", *arXiv preprint arXiv:1806.00451*, 2018.

[24] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, C. Feldmann, A. F. Frangi, P. M. Full, B. Van Ginneken, A. Hanbury, K. Honauer, M. Kazubek, B. A. Landman, K. März, O. Maier, K. Maier-Hein, B. H. Menze, H. Müller, P. F. Neher, W. Niessen, N. Rajpoot, G. C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. Van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, and

A. Kopp-Schneider, "Why rankings of biomedical image analysis competitions should be interpreted with care", *Nature communications*, vol. 9, no. 1, p. 5217, 2018.

[25] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in CT images", in *2015 12th Conference on Computer and Robot Vision*, IEEE, 2015, pp. 133–138.

[26] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran, and J. H. Saltz, "Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images", *Pattern recognition*, vol. 86, pp. 188–200, 2019.

[27] J. Arevalo, A. Cruz-Roa, V. Arias, E. Romero, and F. A. González, "An unsupervised feature learning framework for basal cell carcinoma image analysis", *Artificial intelligence in medicine*, vol. 64, no. 2, pp. 131–145, 2015.

[28] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis", in *International conference on machine learning*, 2016, pp. 478–487.

[29] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm, "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring", *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1322–1331, 2016.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", *Stat*, vol. 1050, p. 10, 2014.

[31] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders", *arXiv preprint arXiv:1711.01558*, 2017.

[32] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders", in *International Proceedings on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1511.05644.

[33] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution", in *European Conference on Computer Vision*, Springer, 2016, pp. 694–711.

[34] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric", pp. 1558–1566, 2016.

[35] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network", *arXiv preprint arXiv:1505.00853*, 2015.

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", in *International Conference on Machine Learning*, 2015, pp. 448–456.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] R. Bellman, "Dynamic programming", *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[39] D. Tellez, J. van der Laak, and F. Ciompi, "Gigapixel whole-slide image classification using unsupervised image compression and contrastive training", 2018.

[40] A. Odena, V. Dumoulin, and C. Olah. (2016). Deconvolution and checkerboard artifacts, [Online]. Available: http://distill.pub/2016/deconv-checkerboard (visited on 08/29/2019).

[41] W. Bulten, P. Bándi, J. Hoven, R. v. d. Loo, J. Lotz, N. Weiss, J. v. d. Laak, B. v. Ginneken, C. Hulsbergen-van de Kaa, and G. Litjens. (November 2018). PESO: Prostate Epithelium Segmentation on H&E-stained prostatectomy whole slide images, [Online]. Available: https://doi.org/10.5281/zenodo.1485967 (visited on 08/29/2019).

[42] G. Litjens, *Automated Slide Analysis Platform (ASAP)*, 2017.

[43] A. B. Yoo, M. A. Jette, and M. Grondona, "Slurm: Simple linux utility for resource management", in *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, 2003, pp. 44–60.

[44] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org, [Online]. Available: https://www.tensorflow.org/ (visited on 08/29/2019).

[45] K. Dercksen, W. Bulten, and G. Litjens, "Dealing with label scarcity in computational pathology: A use case in prostate cancer classification", *Medical Imaging with Deep Learning 2019 (MIDL)*, 2019.

[46] M. Wattenberg, F. Viégas, and I. Johnson. (2016). How to Use t-SNE Effectively, [Online]. Available: http://distill.pub/2016/misread-tsne (visited on 08/29/2019).