Classification of High Risk Prostate Cancer using Deep Learning

Elin Olofsson
ine14eol@student.lu.se

June 4, 2019

Master's thesis work carried out at
the Department of Mathematics, Lund University.

Supervisors: Anders Heyden
Ida Arvidsson

Examiner: Niels Christian Overgaard

Abstract

Prostate cancer is one of the most common types of cancer for men, making proper diagnostic essential. Using machine learning as a tool to help in digital pathology has become increasingly popular and helps to limit the high intra observer variability between pathologists. Due to the many cases of prostate cancer and the large differences between tumours, treatments have to be individualized for each patient. The Active Surveillance was introduced for patients with low risk prostate cancer were treatment in the form of surgery or radiation was deemed too invasive for the cancers current state. Instead the progression is supervised and when, if ever, a certain threshold is surpassed further treatment is discussed.

In this thesis it is investigated if a Convolutional Neutral Network (CNN) can be trained to find high risk patients before pathologists can see cancer progression and if benign tissue holds vital information about future development. A CNN was trained on two different datasets, the first containing all of the available data and the second only including the biopsies from the latest examination in a patient's timeline.

The results indicate that the problem is hard and the biggest struggle has been to limit the data without introducing new biases. The variability within each class was seemingly large in relation to the possible underlying patterns containing clues about the cancer making the accuracy low. Generalization was overall bad but it was found that when combing the results to make a patient grading, instead of grading individual biopsies, accuracies increased. Peak performance was found when only training on the last biopsies and was for the patient grading 67%. Although no outstanding results were found further research has to be done in the area of predictive prostate cancer classification in order to draw any final conclusions.

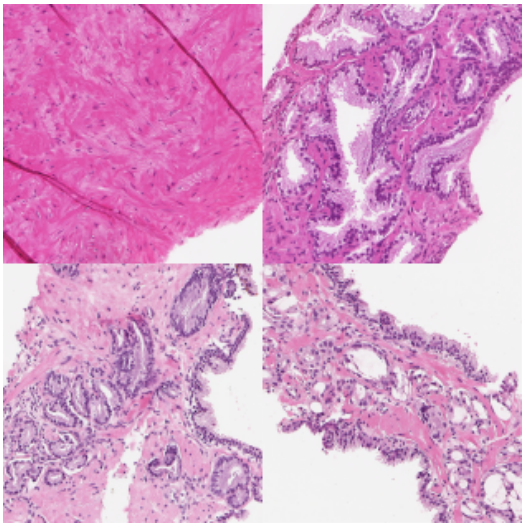
Keywords: Prostate Cancer, CNN, Deep Learning, PRIAS, Active Surveillance, Digital Pathology

Klassificering av högrisk-prostatacancer med hjälp av maskininlärning

Prostatacancer är den näst vanligast förekommande typen av cancer för män i världen. Att ha tillgång till ordentliga och tillräckliga diagnostiseringsverktyg är därför en avgörande faktor för hur väl vården kan hantera och behandla patienterna. Idag används Gleason gradering for att bedöma hur allvarlig cancern är. Metoden introducerades på 60-talet och går ut på att granska hur välstrukturerade cellerna är och därmed definiera stadiet av tumören.

För att undvika överklassificering av cancern och därmed introduktion av invasiva och onödiga behandlingar finns det en studie som övervakar svårigheten av tumören. Patienter som bedöms ha en lågrisk cancer som växer långsamt och inte riskerar att sprida sig eller bidra med svåra symptom får vara med i en så kallad Active Surveillance (AS) där de får komma på upprepade besök och lämna biopsiprover. Om cancern sedan bedöms ha utvecklats för mycket och medför en risk för spridning får patienterna vidare behandling.

Skånes Universitets Sjukhus har under tio år samlat in ett dataset över patienter som medverkar i AS. Målet med arbetet var att ta reda på om det går att använda maskininlärning för att prediktivt hitta patienterna som löper större risk att utveckla värre cancer och därmed även undersöka om den friska vävnaden innehåller information kring hur cancern stagnerar.



Figur 1: Exempelbild på hur de olika vävnadstyperna i datasetet kan se ut. Med börjar uppe till vänster innehåller bilderna; bindväv, körtlar, cancer med Gleason grad 3 samt cancer med Gleason grad 4.

Sedan idén att modelera den mänskliga hjärnan introducerades har metoderna och användningsområdena mångfaldigats. I takt med att datorerna förbättrats har maskininlärningen utvecklats och även blivit populär inom medicinsk analys. Speciellt applicerbart för bildanalys är ett så kallat Convolutional Neural Network (CNN) som tar hänsyn till den spatiala strukturen i bilder. Det enda som behövs för träning av nätverket är ett dataset med bilder samt deras förbestämda klass. Träningen utförs sedan genom optimering av de variabler som nätverket byggs upp utav för att den slutgiltiga klassificeringen ska ligga så nära den ursprungliga som möjligt.

Ett CNN har i detta arbete tränats på urklipp med storlek 500×500 RGB-pixlar från nålbiopsier tagna av patienter med prostatacancer. Urklippen delades först upp i två olika klasser, de som fortfarande är aktiva i studien och de som blivit exkluderade på grund av att cancern försämrats. Två olika dataset har använts för att träna och testa nätverket, först urklipp från alla tillgängliga biopsier och sedan enbart från de biopsier som togs vid varje patients senaste provtillfälle.

Modellerna som presenteras i rapporten hade stora problem med att generalisera klassificeringen och hitta relevanta drag i bilderna som skulle kunna beskriva framtida utvekling av cancervävnaden. Några problem som belyses i rapporten är svårigheten i att begränsa datasetet så att det innehåller relevant information utan att bli för partiskt samt hur tidsaspekten på bilderna ska tas hänsyn till.

Acknowledgements

Firstly a special thank you to my supervisor Ida Arvidsson for her valuable insight in the project and for never ending feedback, discussions and suggestions regarding model selection and data processing.

Moreover, I would also like to thank Anders Heyden for support and feedback as well as Agnieszka Krzyzanowska at SUS for invaluable help with the concepts of cancer classification in addition to detailed descriptions of the dataset.

Elin Olofsson
Lund, May 2019

Abbreviations

ANN Artificial Neural Network

AS Active Surveillance

CNN Convolutional Neural Network

H&E Hematoxylin-Eosin

LU Lund University

MLP Multilayer Perceptron

PRIAS Prostate Cancer Research International Active Surveillance

PSA Prostate-Specific Antigen

ReLU Rectified Linear Unit

RMSprop Root Mean Square Propagation

RP Radical Prostatectomy

SGD Stochastic Gradient Descent

SUS Skåne Univerity Hospital

Contents

1

Introduction

Prostate cancer was one of the most common types of cancer found amongst men worldwide in 2018, as well as one of the types with the highest mortality rate [3]. Given these statistics it becomes evident that fully functioning diagnostic tools are crucial for pathologists when determining the appropriate treatment for each individual patient. However, it has been concluded by several studies that the variability between pathologists' classifications is large, leading to possible under- or over-grading of the malignant tissue and hence assignment to the wrong treatment [26, 21, 6].

Another problem to tackle is the sheer amount of data that the pathologists have to go through on a daily basis in order to find the suspicious sections of the biopsies. For a standard diagnostic procedure a total amount of 10 needle biopsies are taken, only a couple of these contain cancer cells and the malignant tissue cover only parts of the samples making the pathologists spend a lot of time looking at benign tissue. Figure 1.1 shows an example of a whole biopsy as well as an enlargement of it to illustrate the amount of data in one needle biopsy of prostate cancer. Litjens et al. [25] introduced deep learning as a tool to lighten the workload of pathologist by excluding images that did not contain cancer cells with the help of machine learning, giving them more time to focus on correctly diagnosing the malignant tissue. They concluded that up to 40% of the biopsies containing only benign tissue could be automatically identified.

Neural networks, because of their ability to study large amounts of data in a seemingly short time, are helpful tools when it comes to classifying objects, specifically images [22]. Machine learning is a collective name wherein Artificial Neural Networks (ANNs) are a subgroup. The main idea is that the network trains itself based on given data and optimizes it's parameters in comparison to predefined outputs. Incorporating machine learning in medical analysis has thus become popular and can hopefully, if used correctly, help regulate data that has to be manually inspected and reduce inconsistency in diagnostic.

1.1 Background

The Prostate Cancer Research International Active Surveillance (PRIAS) study was introduced in 2006 with the aim of reducing over-treatment of prostate cancer by supervising patients with tumours that grow slowly yielding none to mild symptoms. Patients that are classified with this type of low risk cancer can take part in an Active Surveillance (AS) where they are scheduled for regular biopsies to monitor the state of their cancer. The goal with the AS is to find anomalies in the tumours as quickly as possible without introducing unnecessary and invasive treatments [2].

A research collaboration between Skåne Univerity Hospital (SUS) and Lund University (LU) is in progress at writing time where the scope is to investigate how digital image analysis can be used for optimization of Gleason grading, the leading diagnostic tool used for prostate cancer. For SUS the project is denoted Digital Pathology for Optimized Gleason Score (DOGS-2) and is the latest of two projects sponsored by Vinnova. At LU the project is a part of a collective base, Analytic Imaging Diagnostics Arena (AIDA), and is called Decision Support for Classification of Microscopy Images in Digital Pathology Using Deep Learning Applied to Gleason Grading. As a side study to this collaborative project, SUS has collected a PRIAS cohort since 2007 resulting in a dataset containing needle biopsies from 180 patients participating in an AS. Because of the long scope of the study ground truth over how the patients' cancer have progressed and whether they needed treatment or not can be retrieved with approximate certainty. Since the labels of the data correspond directly to the cancer progression, detailed annotations from pathologists are not needed thus reducing the screening time for each patient.

A Convolutional Neural Network (CNN) is a type of Artificial Neural Network (ANN) that is specialized in classifying images. Given a set of images and their respective labels the network is trained to mimic the label given any of the input images. During the training, like humans, the network finds features in the images that help them create knowledge over how a certain type of image should look. The training is well defined using mathematical formulas, which are described in Section 2.4, and have the sole purpose of trying to minimize the miss-classification. Because of the mathematical optimization the features might be different from what we humans believe are good cues. Incorporating deep learning as a diagnostic tool might therefore help find features that are not yet used in cancer classification but that contain significant information in the development of the tumours.

As previously described the AS was introduced to help diagnose high risk cancer in an early stage to prevent eventual spreading and progression as soon as possible. Since it is yet debated amongst pathologists whether cancer cells of Gleason grade 3 can have metastases or not it is important to find the high risk patients as soon as possible to reduce cancer spread without introducing unnecessary treatments.

This thesis aims to identify high risk prostate cancer in an earlier stage than only AS can by predicting future development of the cancer using deep learning in the form of a CNN, i.e. investigating if the seemingly benign tissue holds any vital information about the cancer.

1.2 Previous Work

Previous work in the field of digitized medical image analysis of prostate cancer has been focused on improving existing diagnostic tools by introducing automated Gleason grading in different forms. In 2007 Doyle et al. [6] published a paper stating that with the use of a support vector machine classification of the four classes; benign stroma, benign epithelium, Gleason grade 3 and Gleason grade 4, could be determined with up to 92.7% accuracy. The extracted features for the model was based on textural patterns and graphs over nuclear structures.

Källen et al. [19] and Gummeson et al. [14, 15] both used CNNs for automated Gleason grading with increasingly accurate results, the scope focused on introducing interactive tools for pathologists for lower inter observer variability. In [25] Litjens et al. introduced deep learning as a tool to pre-process whole biopsy samples in order to automatically exclude images containing all benign tissue.

Furthermore, Lee et al. [24] and Cordon-Cardo et al. [4] have studied how to incorporate digital pathology in recurrence prediction after Radical Prostatectomy (RP). They both concluded that benign tissue surrounding malignant areas held crucial information about the progression of the cancer.

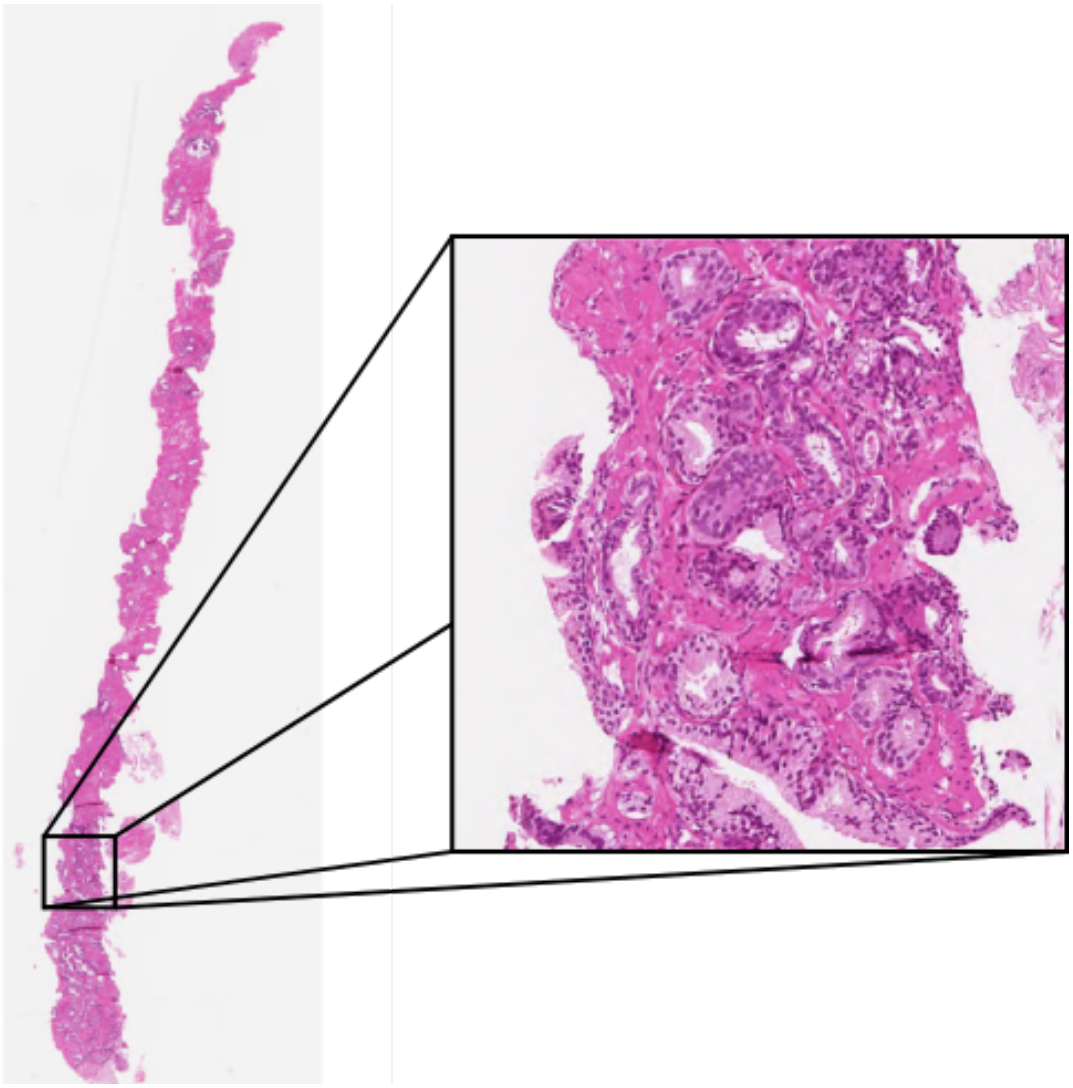


Figure 1.1: Comparison between a full needle biopsy sample from the prostate and an enlarged section wherein the cellular structures can be properly seen. The enlarged section contains cancer cells of Gleason grade 3.

2

Deep Learning

Deep learning is a collective name for all types of machine learning that is considered deep i.e. that has multiple layers that are hidden from the user and contributes to the so called "black-box" behaviour. This chapter aims to describe the relevant theory behind the machine learning methods used in this thesis. Initially the more general models included in ANNs will be described before more specific theory about how CNNs are built is introduced.

2.1 Artificial Neural Networks

The idea behind machine learning is that, using some dataset, an artificial structure can be taught to replicate outputs correlated with the individual inputs in the dataset. An ANN is built up by perceptrons that contain weights determining the behaviour of the network and that during training are optimized to label the data.

2.1.1 Simple Perceptron

As can be heard from the name, ANNs are constructed as to imitate the neural structure in the human brain [13, p. 165]. Subsequently, the simplest part of a neural network, called a perceptron, show close resemblance to actual neurons, both of which consists of an input layer, some activation and an output layer.

Because of the complexity of the biological neuron the structure of the perceptron has been simplified down to a straight forward mathematical formula, which can be seen in (2.1). Hence, the output, y, is a combination of weighted inputs, x_i, with w_i being the corresponding weight, evaluated with some activation function, ψ. The activation function is most commonly a non-decreasing function that can be bound either from below or in

both directions. Since different neurons might have individual constant values, a bias is introduced. To incorporate this in the summation in (2.1) we can denote $x_0 = 1$.

$$y = \psi(\sum_{i=0}^{N} w_i \cdot x_i) \tag{2.1}$$

Figure 2.1 shows a graphic representation of the formula in (2.1), where the flow goes from left to right.

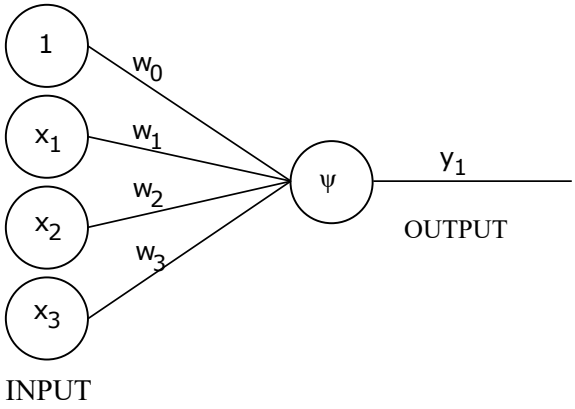


Figure 2.1: A schematic representation of a simple perceptron including bias and activation function.

2.1.2 Multilayer Perceptron

The Multilayer Perceptron (MLP) is a combination of multiple layers of simple perceptrons, a schematic can be seen in Figure 2.2. Increased complexity in the architecture combined with using non-linear activation functions is a way to model more involved functions [12, p. 2628].

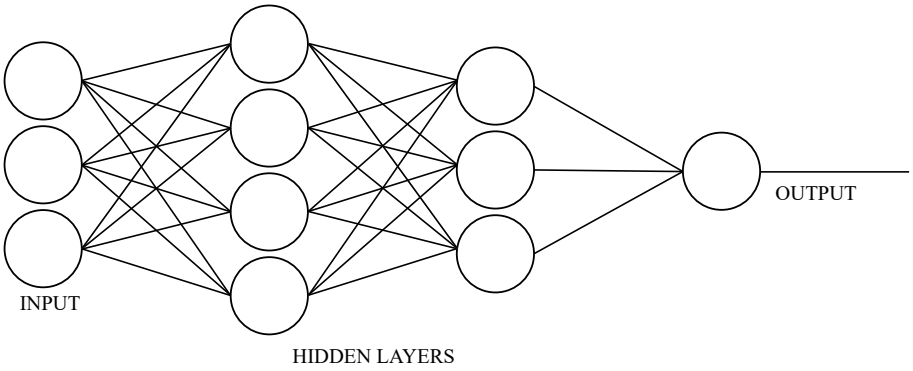


Figure 2.2: A fully connected MLP with two hidden layers.

The activation function determines the behaviour of the neuron. For description of non-linear correlations between input and output, the activation function has to be non-linear

as well. Since CNNs became popular, it has been concluded that Rectified Linear Unit (ReLU), $\psi(x) = \max(0, x)$ is an efficient choice for hidden layers [22, p. 439] [5, p. 8610].

2.2 Convolutional Neural Networks

A CNN is a type of ANN that considers spatial dependencies in the different layers, making it perfect for image analysis. In 1989 LeCun et al. introduced a method for digital classification of the MNIST dataset, a collection of handwritten digits from 0 to 9 [23], increasing the exposure of CNNs. Combined with the development in performance and memory size for computers it became possible to train neural networks with deeper architectures and higher complexity, giving birth to the new name deep learning. To better understand how a CNN processes images, theory about the three primary layers of such a network will be described; the convolutional layers, pooling layers and the final dense layer.

2.2.1 Convolutional Layer

The convolutional layers of a CNN are used for extracting features from the input. Images are essentially arrays with high spatial dependencies, meaning that if a pixel in an image belongs to a tree it is likely that neighbouring pixels also belong to a tree as opposed to a dog or a car. To exploit this characteristic filters, or kernels as they are also called, are used in the convolutional layer and are essentially sparsely connected MLPs with shared weights.

Multiple kernels of a smaller size than the input image, usually 3×3 or 5×5 pixels, are moved across the input, i.e. convolved, to create a set of new images to be processed by the next layer. For a deep system these convolutional layers are combined in order to create larger structures of features. As described by LeCun, Bengio & Hinton [22] the first layer often detects edges in different directions, the second layer then combines edges in order to construct parts of objects. Further convolutional layers are used to assemble more complex structures in hopes of finally separating images of different classes.

Mathematically the convolutional layers are 2-dimensional discrete convolutions, as can be seen in (2.2). Here g and f represents the input and filter respectively and the indexation represents the pixel in row i and column j [14, 22, 13].

$$y_{ij} = \sum_m \sum_n f_{i-m,j-n} g_{m,n} \tag{2.2}$$

From this we can describe how the filtering works, namely, for every pixel in the input image a kernel is multiplied with pixels in a neighbourhood of the original. The filter size determines the neighbourhood which should be considered. As can be seen from the summations in the equation, the filter is not directly multiplied with the image but first flipped in both directions. However, for our intuition about filtering this holds no

significance, especially since an image of a dog is still an image of a dog even if it is flipped upside down and left to right. Figure 2.3 shows a schematic over how the kernel is convoluted with the original image.

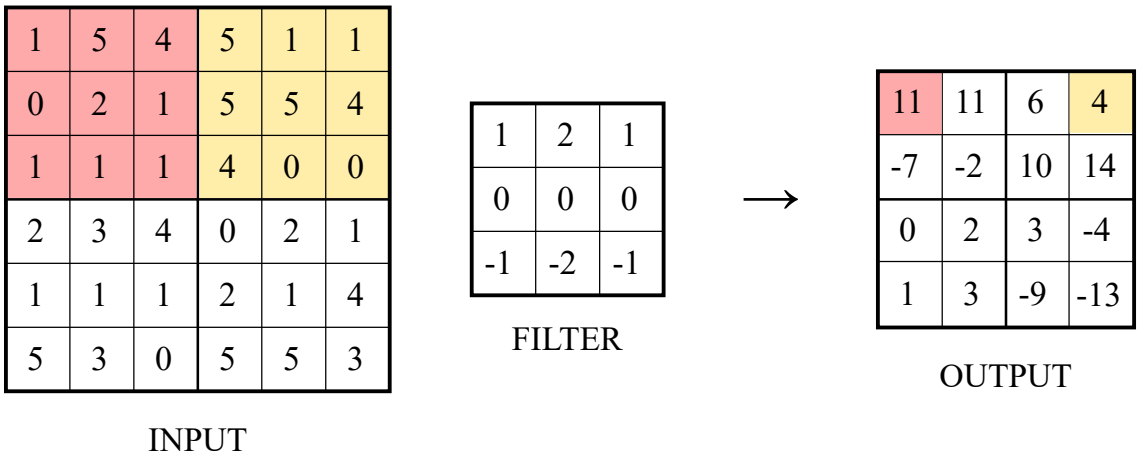


Figure 2.3: Example of a convolutional layer with no padding. The represented filter is flipped in the schematic.

As previously described, a CNN is essentially a sparsely connected MLP where the spatial dependency of the input images determines the connectivity of the network. This feature reduces the parameters that need to be trained and reduces the risk of over-training.

The idea of using a CNN is that instead of manually constructing kernels, the network itself can better determine the optimal values for the filters by training on a given dataset. Specifying the location of a feature too precise will have no effect on the final classification which is why filters of much lower dimensions than the input image are used [23, p. 542]. To further generalize the exact position of a component, a pooling layer is used to complement the convolutional layer described in this section.

2.2.2 Pooling Layer

Some generalization of object position is an important detail for image classification using CNNs since a small tilt or zoom on the image should not change the results. However, keeping the relative distances between features are of utmost importance since they later make up more complex structures deeper into the network [27, p. 94]. To utilize the images spatial dependence without emphasizing exact locations a max-pooling layer is used.

Application of this layer entails an image reduction, defined by the size and spacing of the pooling [22, p. 439]. The output is simply the maximum value encountered in the neighbourhood where the pooling filter was applied.

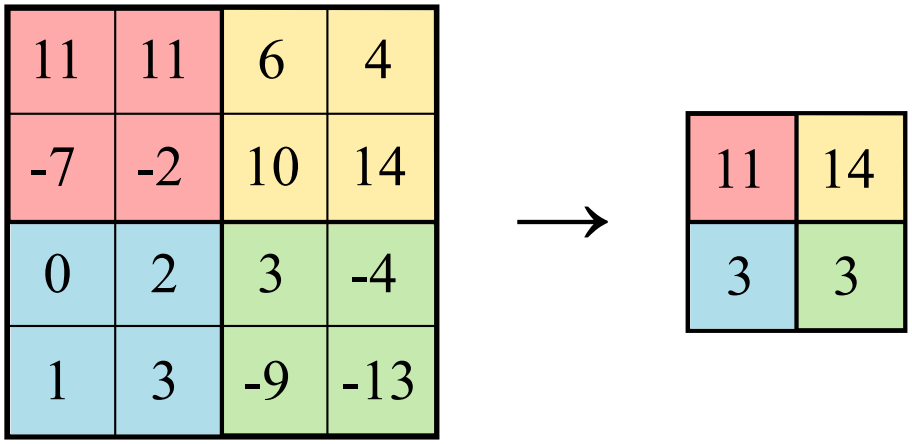


Figure 2.4: Example of a 2×2 max-pooling layer.

2.2.3 Dense Layer

The final step of the CNN consists of a dense layer, i.e. a MLP that is fully connected. Because the dense layer is added to the end of the network to classify the input, the number of hidden layers and neurons will be decided by the complexity of the images. However, constructing a more intricate network will make it more prone to over-fitting since the CNN then adapts too well to the training set, making it bad at generalizing. Over-fitting is a serious problem with deep learning, but can be avoided with the use of e.g. dropout [5], this technique will be described in more detail later.

2.3 Classification

By adding the previously described layers together in different formations we can construct a CNN. With each layer the intention is to reduce the dimensions necessary to describe the input, from e.g. 500×500 RGB-pixels down to a single variable used for classification of the images. Because of the complexity of the network and the non-linear activation functions the input is transformed into a space wherein the classes can be linearly separable. For a binary classification problem the sigmoid function, $y(x) = \frac{1}{1+e^{-x}}$ [13, p. 65], is used as output function and represents a probability of the input to belong to one of the two classes, $P(\text{class } 1)$. The probability for the other class can then be calculated as $P(\text{class } 2) = 1 - y(x)$. It is also worth to mention that it is possible to classify more than two classes without much alteration. The biggest change is in the choice of output- and loss-function, which will not be more thoroughly described.

2.4 Back-Propagation

The idea of modelling the learning process of the human brain has been around for many decades but it was not until the introduction of back-propagation that ANNs became especially useful [27]. Back-propagation is an algorithm that is used during training to optimize the network parameters to best describe the input data. For a CNN, in each training cycle a number of images are sent through the network with randomly initiated weights returning some output, $y(x_i)$, i indexes the input images and labels. After N images has been evaluated by the network the results are compared to the ground truth labels, y_i, via a loss function. The loss is then back-propagated through the network to optimize the weights and find better representations of the images [12, 22].

2.4.1 Loss Function

The binary cross entropy function, $L(\mathbf{x})$, is a loss function that can be used to determine the loss of a network, where $\mathbf{x}$ is the list of all images used for training, see equation 2.3 [20].

$$L(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(y(x_i)\right) + (1 - y_i)\log\left(1 - y(x_i)\right) \tag{2.3}$$

Since the goal is to minimize the loss function it becomes evident that the classification $y(x_i)$ should be as close to 1 as possible if the true label is $y_i = 1$ and close to 0 if the label is $y_i = 0$. N describes the number of images used in the current training batch and is used for normalization of the loss [22].

2.4.2 Optimizer

The purpose of the optimizer is to change the weights of the network in a direction that minimizes the loss function. The optimizer determines how the loss function should be used for minimization of the error and therefore somewhat defines the properties of the network. One of the simplest models for optimization is called Stochastic Gradient Descent (SGD) and had been used diligently since the introduction of back-propagation [22, 13].

Stochastic Gradient Descent

Optimizing over all of the available training samples simultaneously has proven inefficient and the stochastic approach of the gradient descent is therefore used. In this case stochastic means that the training data is split into smaller batches and the optimization is done once for each batch. A full epoch has run when every image in the training folder has been evaluated once and the batch split is randomized for every epoch [13, ch. 8].

The base of the SGD algorithm is simple, it uses the gradient of the loss function with respect to the internal weights, θ_j, to determine the direction in which the loss function has the steepest descent. The weights are then updated and moved some small step in that direction. The size of the step depends both on the learning rate, ϵ, and the gradient, $\Delta\theta_{j,k}$. The algorithm is described in Table 2.1.

Table 2.1: Stochastic Gradient Descent

SGD Algorithm	
Initial parameters	$\theta_{j,0}$
Iteration index	$k = 0$
For each batch in the epoch:	
Sample mini-batch	$\mathbf{x}_k = \{x_1, x_2 \cdots, x_N\}$
Compute gradient estimate	$\Delta\theta_{j,k} = -\frac{1}{N}\sum^{N}\frac{\delta E(\mathbf{x}_k)}{\delta\theta_{j,k}}$
Update iteration	$k = k + 1$
Update weights	$\theta_{j,k} \leftarrow \theta_{j,k-1} - \epsilon \cdot \Delta\theta_{j,k}$

Here $\theta_{j,k}$ represents the weights for layer j at iteration k and ϵ is the learning rate. The error function for each layer is represented by $E(\mathbf{x})$ where the output of the previous layers has been put into the activation function of the current layer. The gradient for the binary cross entropy loss function can be seen in equation 2.4. For each layer in the network the gradient will be increasingly complex and the chain rule for derivation has to be utilized.

$$\nabla L(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i}{y(x_i)} + \frac{y_i - 1}{1 - y(x_i)}\right)y'(x_i) \tag{2.4}$$

A non-adaptive learning rate is used for simplicity but this is not feasible for applied optimization since the method then would have trouble with convergence due to the vanishing gradient close to the minima [13].

Other Optimizers

Other optimizers include Root Mean Square Propagation (RMSprop) and Nadam which both are methods with adaptive learning rates. RMSprop keeps a running average of the squared gradients for each weight to make the learning rate adaptive. Nadam is Adam with Nesterov momentum and works similar to RMSprop but additionally keeps track of past gradients [7]. The momentum term keeps the algorithm going more or less in the same direction as previously, this to prevent it from oscillating in a valley perpendicular to the optimal descent direction. They both work well for image classification but Nadam tends to find less extreme optimums because of its momentum term.

2.5 Regularization

As previously stated, over-fitting is a problem when the network or model becomes too complex in relation to the available data. The relationship between the input and output becomes reliant on measuring noise in the training set making the generalization to unseen images very poor. Regularization is a process where the relationship between input and model is randomly changed to prevent over-fitting. Following are three regularization methods used in this thesis.

2.5.1 Dropout

Dropout is a regularization method that removes a node in a hidden layer with the probability p, example given in Figure 2.5. For each new training epoch random nodes are dropped with the given probability and the thinned network is optimized [28]. The method can also be seen as a type of ensemble technique where the output of multiple thinned networks is used to approximate the input.

In a CNN dropout is only used for the MLP since the convolutional layer has substantial weight sharing, reducing the over-fitting of the filters. For a CNN, Dahl et al. [5] describes that dropout is best used in combination with the ReLU activation function for the hidden nodes.

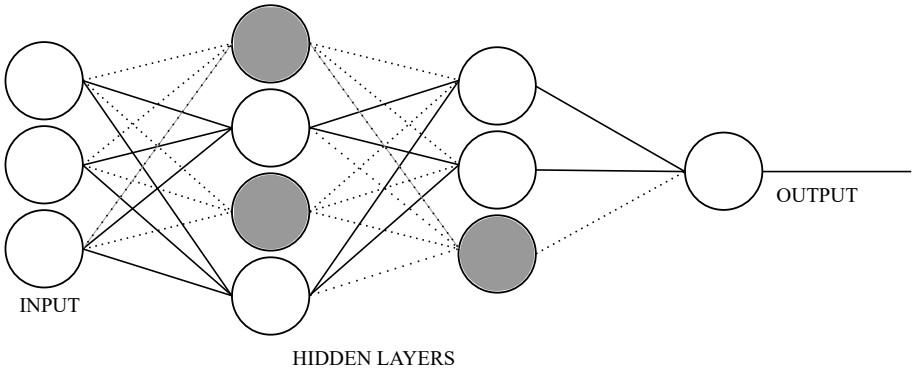


Figure 2.5: An example of a thinned network after applying dropout with $p = 0.5$ to a fully connected MLP.

2.5.2 Image Augmentation

Another way to prevent a network from becoming over-trained is to introduce augmentation on the training images so that each epoch a new version of the same image is seen. This includes rotations, mirroring and shifts in height and width. Figure 2.6 shows an example from the dataset.

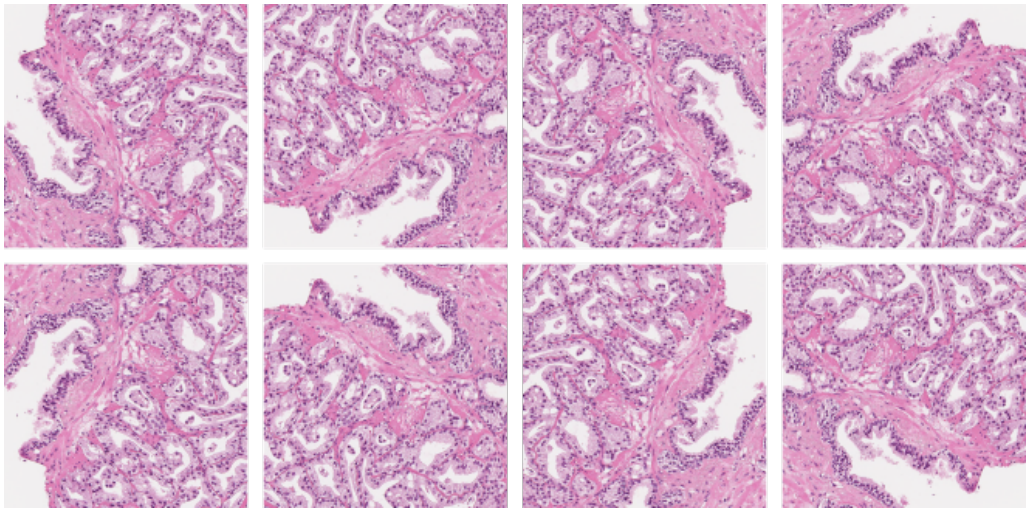


Figure 2.6: Example of how one image from the dataset might be augmented during training. The first row corresponds to the 4 rotations with 90 degree angle and the following row contains mirrored representations.

2.5.3 Early Stopping

A third way of preventing over-fitting and keeping the model generalized is the incorporation of early stopping. To utilize this tool a validation set has to be included during training in addition to the training dataset. After each training epoch the validation dataset is sent through the network and the loss and accuracy is measured. If the loss for the training data keeps descending each epoch while the validation loss starts to increase it is a sign that the model has become too closely fitted with the specifics in the training set and training should therefore be stopped even if the desired number of epochs has not been executed. Afterwards one can restore the weights of the model where the network was not yet over-trained in order to obtain the best configuration.

3
Cancer Diagnostic

Prostate cancer was the second most common form of cancer for men in 2018 [3]. Combined with the increase in patient specific treatments and images containing non-malignant tissue highly increases the work load for the pathologists [25] making it desirable to automate the process. This section will describe current diagnostic tools and connect traditional and digital pathology.

3.1 Gleason Grading

Classification of prostate cancer has been done more or less the same since the 1960s when the Gleason grading first was introduced. The grading is based on architectural patterns of the malignant tissue samples, where a higher grade indicates less structure in the cellular patterns and a maximum of 5 can be given. Since its introduction it has been concluded that Gleason grades 1 and 2 cannot properly be determined through needle biopsy and is therefore considered as benign tissue [8, 10]. Figure 3.1 shows a schematic over how the cellular structures are divided for the different grades.

The final Gleason score for the entire biopsy sample is given as a sum of the two primary patterns which are: the most common type of cancer cells and the highest occurring Gleason grade different from the primary grade. If the biopsy only contains one pattern or the second composes less than 3% of the sample the primary score is doubled [18]. Since this method leads to a Gleason score between 6 and 10 for all cases it was in 2014 proposed that the grading should be divided into 5 groups that better differentiated the scoring [9, 11], Table 3.1 shows the updated grading groups ranging from 1 to 5.

Table 3.1: The updated grading groups based on Gleason grading

Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**
total sum of 6 or less	3+4	4+3	total sum of 8	total sum of 9-10

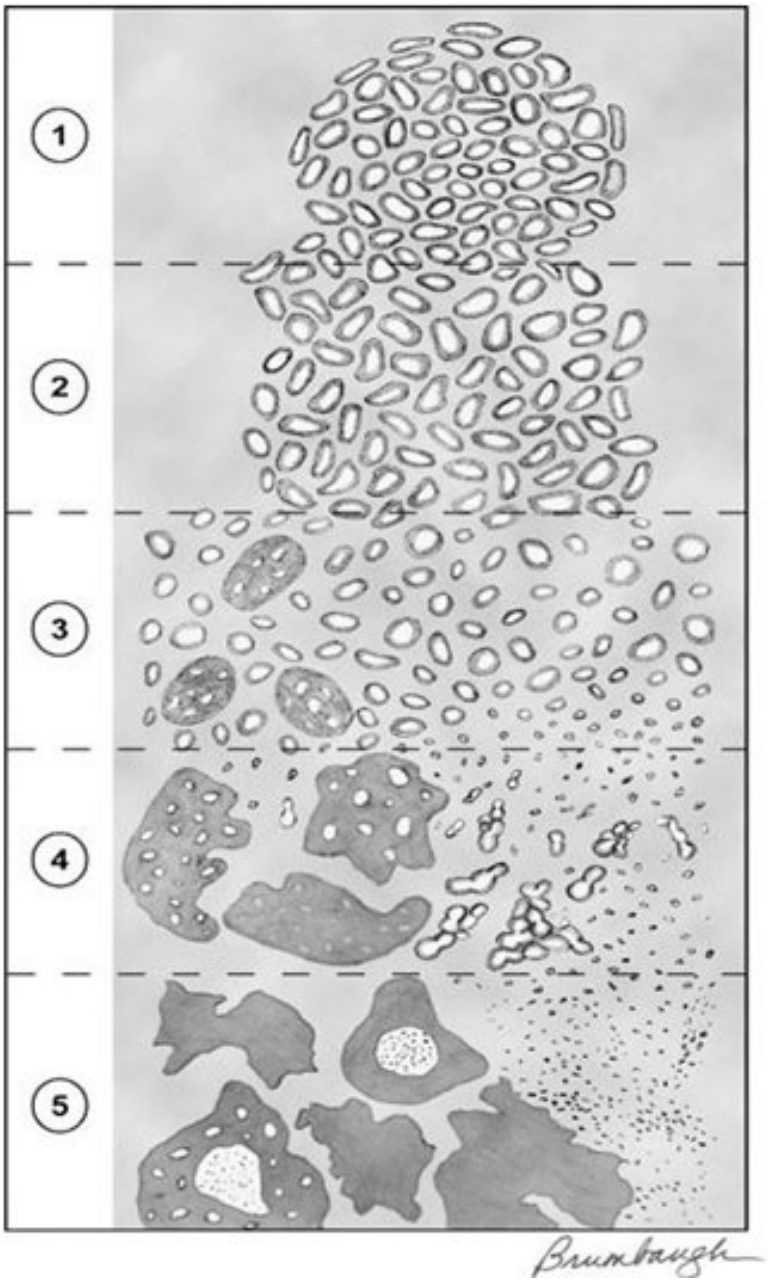


Figure 3.1: Schematics for the updated Gleason grading system done in 2005, demonstrates the cellular structures for the five grading stages where the darker parts have higher nuclei density. Image modified from [1].

3.2 Hematoxylin-Eosin Staining

The most common way to visualize the cellular components of the biopsies is by Hematoxylin-Eosin (H&E) staining before the microscopic analysis [16]. Hematoxylin stains the nuclei blue whereas eosin stains the cytoplasm and connective tissue pink making the classifica-

tion into Gleason scores possible. However, even if H&E is a proven method, variations in saturation, tissue sample and staining duration change the appearance of the samples complicating automation of the classification process [15]. Figure 3.2 shows examples of how the staining might differ both in the gland structure and stroma.

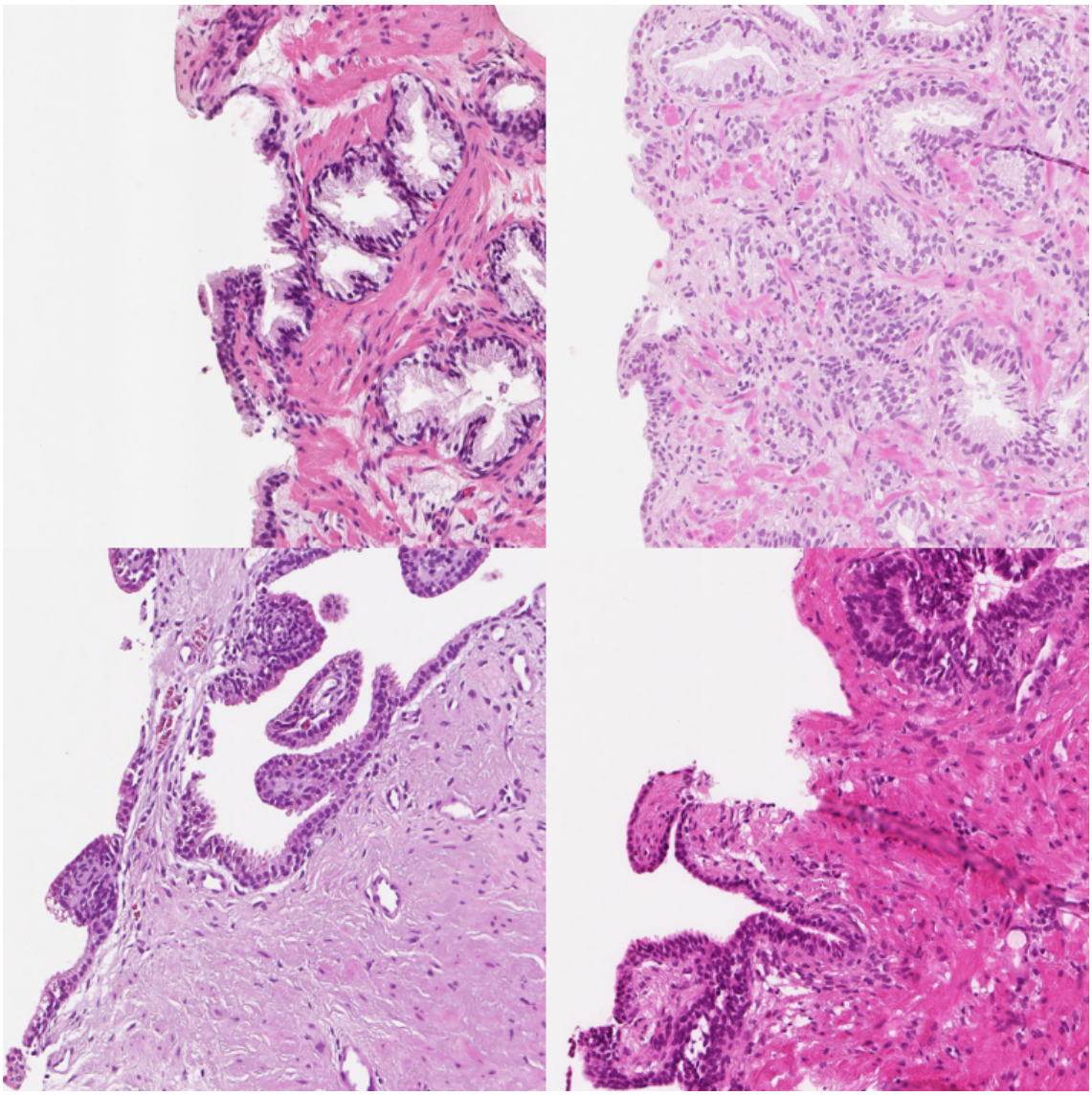


Figure 3.2: Example of how the H&E staining differs in the dataset.

3.3 Cancer Recurrence

Another part of cancer diagnostic is determining the risk for cancer recurrence after eventual treatment. Radical Prostatectomy (RP) is an operation to remove the entire prostate and closely surrounding tissue. As many as 40% of patients with a successfully performed RP will experience some sort of recurrence; rise in Prostate-Specific Antigen (PSA) level,

local, or distant recurrence and cancer death. Lee et al. [24] preformed a preliminary study in 2017 where they used machine learning to try to predict the recurrence rate based on biopsy samples, PSA levels, Gleason grade etc. They concluded that the best predictive model did not only consider tissue samples with malignant cells but incorporated benign tissue surrounding the cancer to better estimate the recurrence rates.

Similarly, Cordon-Cardo et al. [4] performed a study where they used random forest classification in order to determine recurrence rates. In their research they included graphs over nuclei positions for best prediction and reached a concordance index of 0.82 in the validation set. The concordance index is a measure used in survival analysis and logistic regression, also known as area under the ROC curve, where a score of 1 indicates that the model fits the test data perfectly and 0.5 indicates randomized outputs [17, p. 161-164].

3.4 Digital Pathology

The inter observer variability between the Gleason grading of different pathologist has proven to be high and is thus a problem when it comes to assignment to the correct treatment [21, 26]. Variability in classification combined with huge amounts of data, in the forms of images, makes the process especially desirable to automate. Researchers have investigated how CNNs can be incorporated in the grading to reduce miss-classifications and lighten the workload for pathologists. Litjens et al. [25] concluded that as many as 40% of the benign biopsy samples could be excluded from inspection with the use of a neural network.

Similarly, Gummeson et al. [15] and Källen et al. [19] presented results that supported the usage of machine learning as a tool for prostate cancer classification. Their respective networks reach a peak performance at 92.3% and 89% in correlation with the assigning pathologist. These accuracies can be compared to results presented by Lattouf et al. [21] where the accuracy between pathologists when comparing Gleason scores from needle biopsies with scores of the entire prostate after RP stayed at 48.2%. However, note that the presented numbers should be considered carefully since the results are based on limited datasets and might not generalize to new data.

4
Dataset

The dataset is a PRIAS study [2] aimed to better generalize low and high risk cancer by following men with early-stage cancer under AS and is supplied by SUS. For a patient to be eligible for AS there can be cancerous tissue in no more than 2 biopsies with a maximum classification of Gleason group 2, see Table 3.1. If the cancer has grown in following biopsies the patient is excluded from the cohort and gets further treatment in the form of radiation therapy or surgery (RP). The study has been approved by the Local Ethics committee at Lund University no. 708/2008. This includes the use of the dataset for the present study.

The cohort on which the network was trained contains H&E stained needle biopsies from 79 patients with about 10 biopsy samples per test and between 1 and 7 recurrent tests during the years 2007-2018. For this project the images have been saved at 10x enlargement with a resolution of 500×500 RGB-pixels. Up to five images has been retrieved from every biopsy and the total number of available biopsies was 2300. Each patient belonged to one of two classes depending on if they were still active in the AS or not and the data was labeled accordingly. 39 patients were *active* in the AS and 40 were *excluded*, thus classified as high risk patients. Figure 4.1 shows examples of the most common cellular patterns that can be found in the dataset.

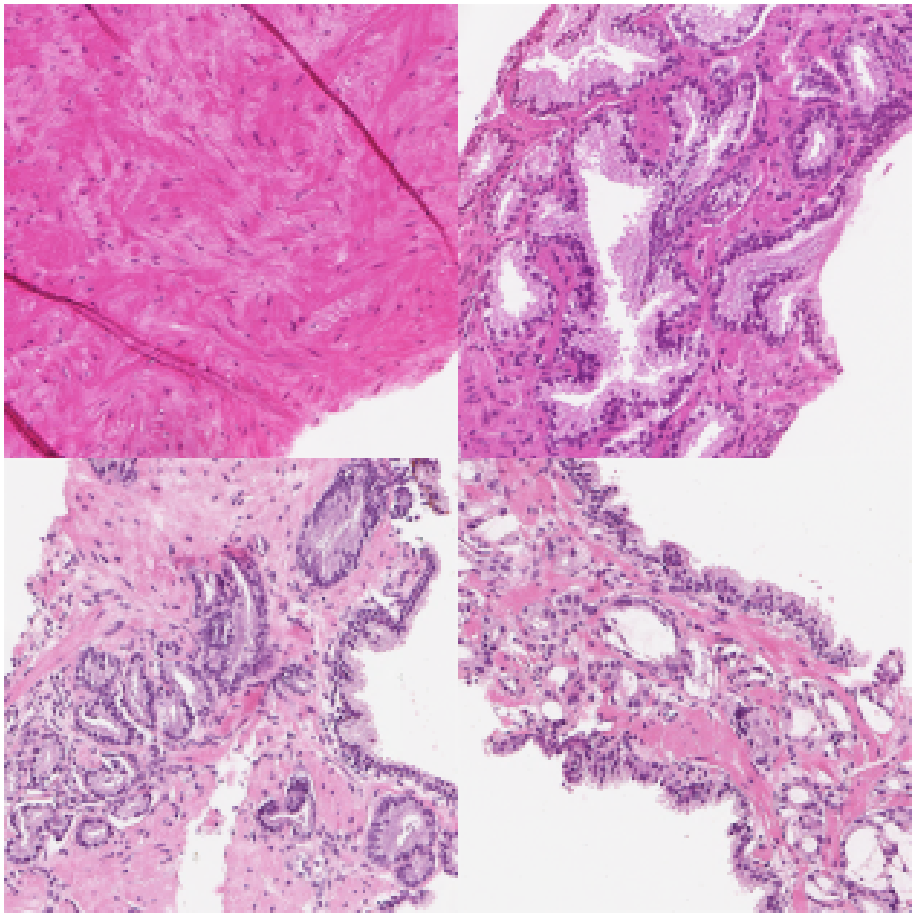


Figure 4.1: Images from the dataset with different types of cellular patterns included. Top left; benign stroma, top right; benign glands, bottom left; Gleason grade 3, bottom right; Gleason grade 4.

5
Method

Because of the large amount of biopsies available the first step was to decide how to approach the data. This step was important because we did not want to introduce any new biases by showing the network what to focus on but at the same time the data had to contain relevant information for the problem at hand. For all of the following datasets the chosen biopsies were sent through a script where a maximum of five patches of size 500×500 RGB-pixels were extracted. The extraction was solely chosen based on thresholding to exclude images with too large white areas as well as images where the staining had become too faded in relation to other biopsies.

We ended up testing two different datasets. Firstly, the network was trained on all the available data where no images were discriminated based on where in the patients timeline they were taken. The second try was based on data only from each patient's latest available biopsies. These images are namely the basis for the pathologists when choosing to give some patients further treatment and the classification is thus known to be feasible.

5.1 Network Training

Before training the network each dataset was split into three subsets: *train*, *validation* and *test*. The purpose of splitting the dataset was to make sure that the model did not over-train and to test the generalization of the network. During training the *train* dataset was used when optimizing the loss function and the final loss and accuracy of an epoch is a weighted sum of each of the batch optimizations. The *validation* set was sent through the network at the end of each epoch to validate that the optimization has not biased too much on the *train*-images and early stopping was used.

After patch extraction on the full biopsies the only pre-processing that was done was image augmentation and colour normalization. The patches in the *train* and *validation* dataset were augmented. For the first dataset the patches were randomly mirrored horizontally and vertically before each epoch and for the second dataset each patch was additionally

rotated four times each with 90 degree angles. Colour scaling was done from (0, 255), which is the colour range for an RGB-image, down to (0, 1), i.e. division with 255.

5.1.1 Network Architecture

The CNN used for classification of the images belonging to one of the two following classes, *active* (denoted as class 0) and *excluded* (denoted as class 1), was constructed with Keras, an open source library implemented in Python [20]. When designing the network the number of convolutional and max-pooling layers was altered as well as the number of kernels in each layer and the optimizer.

5.2 Post-processing

After the network was trained, prediction and evaluation was performed on the test images. Further post-processing was done to better estimate the progression for each patient. Since each biopsy had multiple patches the full biopsy score was determined as an average of the predicted labels for the patches belonging to that biopsy. Finally a patient grading was done where the grade was an average based on the biopsies.

5.2.1 Accuracy Measures

In order to understand the importance of the results a few additional accuracy measures are introduced. For binary classification one can use the following terms to describe the two classes:

- **True Positive (TP)** correctly classified 1
- **False Positive (FP)** classified as 1 but belongs to class 0
- **True Negative (TN)** correctly classified 0
- **False Negative (FN)** classified as 0 but belongs to class 1

Accuracy

The accuracy is determined as the sum of correctly classified cases divided with the total number of tested images.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Confusion Matrix

An important feature to illustrate during testing is the miss-classification for each of the classes. The rows of the confusion matrix represent the two actual classes and the columns specify the label as supplied by the network. Optimally, the matrix would therefore have all non-zero elements on the diagonal.

	Predicted class	
True class	TP	FN
	FP	TN

Precision and Recall

The precision and recall measures the relevancy of the predictions. Precision can be described as: fraction of the selected instances that are relevant, and recall as: fraction of relevant instances that are selected. For measures for class 1 the positive and negative terms change positions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

6

Results

The architecture for all of the presented networks is the same and can be seen in Figure 6.2. There are 6 sections of convolutional and max-pooling layers followed by a sparsely connected MLP with dropout and finally the net is flattened and fully connected with the output node. The kernels for the convolutional layers has size 3×3 except for layer C8 where the kernel is of size 1×1. Max-pooling was done over a region of 2×2 pixels. ReLU was used as an activation function for all layers except for the output which has a sigmoidal function for classification and the loss function used was binary cross entropy. The training batches were of size 32 patches. There were a total of 150 353 trainable parameters in the net. Early stopping was used for all models with a patience of 15 epochs where after the best network was restored.

6.1 Networks

Two separate experiments were performed with different datasets, the first containing patches from all available images, hereon referred to as *all images*, and the second only containing patches from the latest available biopsies of that patient, referred to as *last biopsy*.

6.1.1 All images

The best performing network during training of all images is denoted *Net 1* and Table 6.1 shows the accuracy measures for this model. The model was first optimized with RMSprop and finally fine-tuned with Nadam. 816 patches were used in the train set and 260 in the validation set. Testing was done on two datasets, the first containing images from the same biopsies that were included in the training set, consisting of 271 patches, and the second only containing images from unseen biopsies, a total of 1400 patches. Figure 6.1 displays the loss and accuracy measures for the train and validation dataset during training.

Confusion matrices for the two tests performed on all images can be seen in Table 6.2 and 6.3.

Table 6.1: Measures for network 1 trained on all images. Test 1 contains images from the same biopsies as in the training dataset and test 2 only unseen biopsies.

Measure	Test 1	Test 2
Patch accuracy	0.79	0.47
Loss	0.43	1.46
Precision 1	0.77	0.45
Recall 1	0.68	0.19
Biopsy accuracy	-	0.47
Patient accuracy	-	0.65

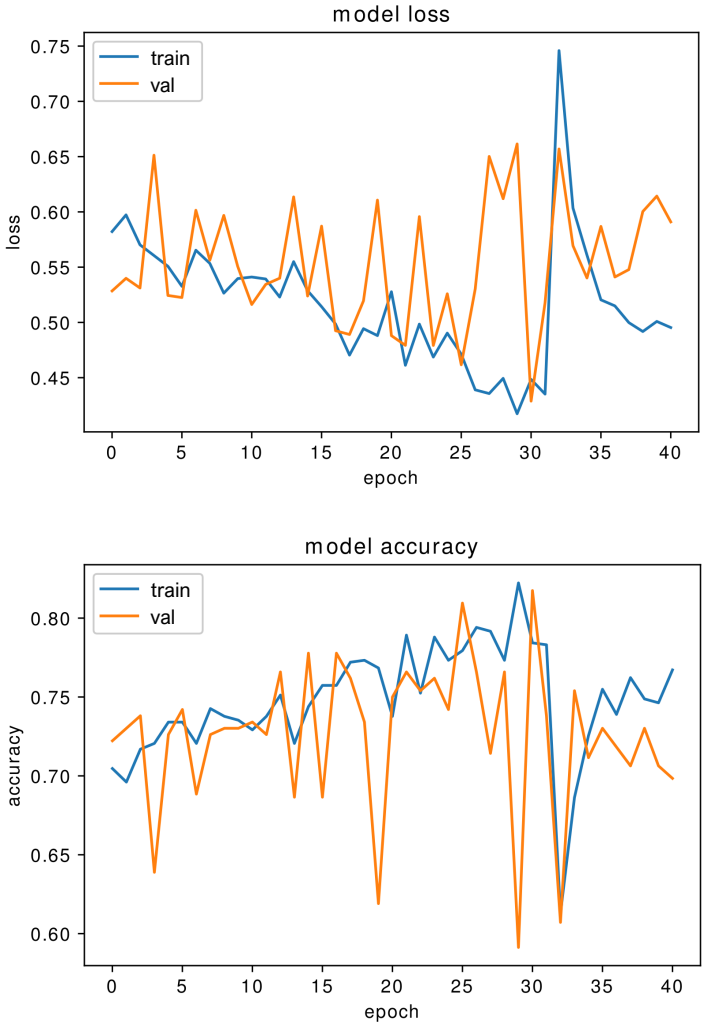


Figure 6.1: Loss and accuracy plots for model 1.

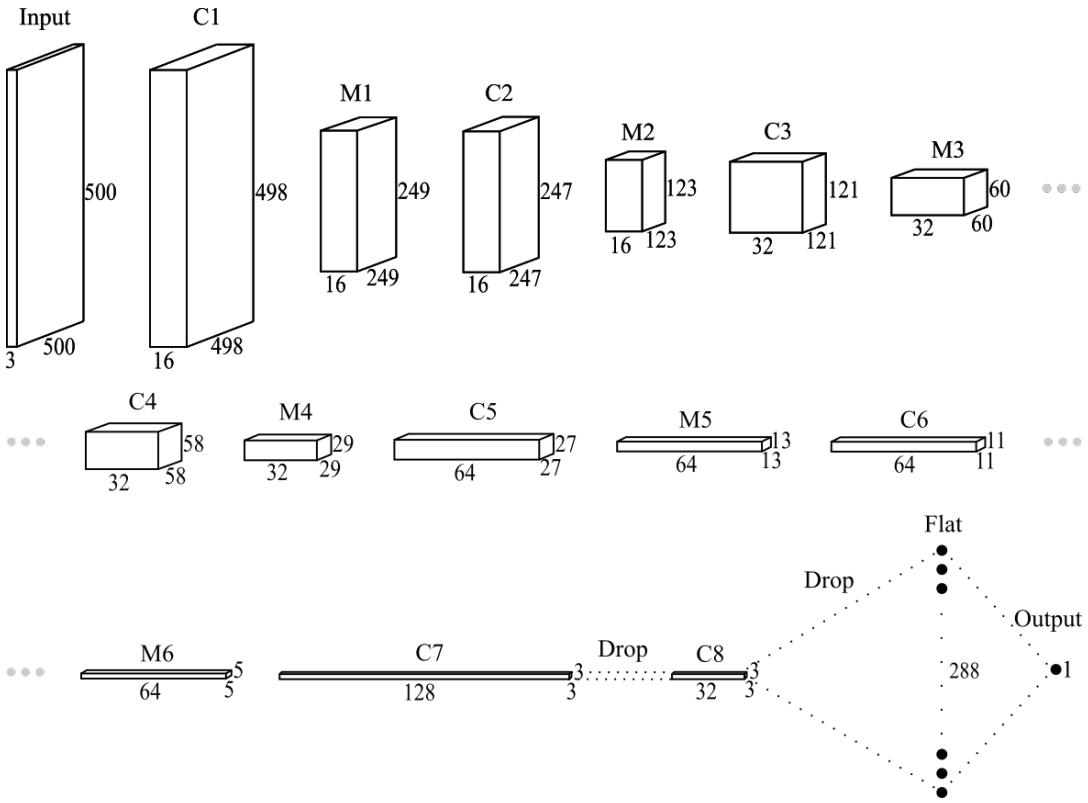


Figure 6.2: Network architecture. The notes corresponds to: C - convolutional layer, M - max-pooling layer, Drop - dropout with drop-probability $p = 0.5$ and the depth correspond to the number of kernels in each layer.

Table 6.2: Confusion matrices for net 1 on all images.

Test 1	Predicted class	
True class	75	35
	22	139

Test 2	Predicted class	
True class	133	580
	164	523

Table 6.3: Confusion matrix for test 2 showing the patient grading for all images.

Test 2	Predicted class	
True class	2	4
	3	11

6.1.2 Last biopsy

For this experiment the number of patches available for the different datasets were: train - 1119, validation - 560 and test - 383. Two models are presented to show the results for the second experiment, *Net A* and *Net B*, and their respective accuracy measures can be seen in Table 6.4. RMSprop was used as an optimizer for both nets. Figure 6.3 and 6.4 show loss and accuracy for the train and validation sets during training. The test set for all presented results only contain images from unseen biopsies and patients to remove all bias from individual gland structure of patients. Confusion matrices for patch prediction and patient grading respectively can be seen in Table 6.5 and 6.6.

Table 6.4: Measures for the networks trained only on last biopsy images.

Measure	Net A	Net B
Patch accuracy	0.577	0.551
Loss	0.772	0.682
Precision 1	0.89	0.42
Recall 1	0.53	0.52
Biopsy accuracy	0.595	0.560
Patient accuracy	0.667	0.667

Table 6.5: Confusion matrices for the two models trained only on last biopsies.

Net A	Predicted class	
True class	159	19
	143	62

Net B	Predicted class	
True class	75	103
	69	136

Table 6.6: Confusion matrices for the final patient grading with model A and B.

Net A	Predicted class	
True class	4	0
	3	2

Net B	Predicted class	
True class	1	3
	0	5

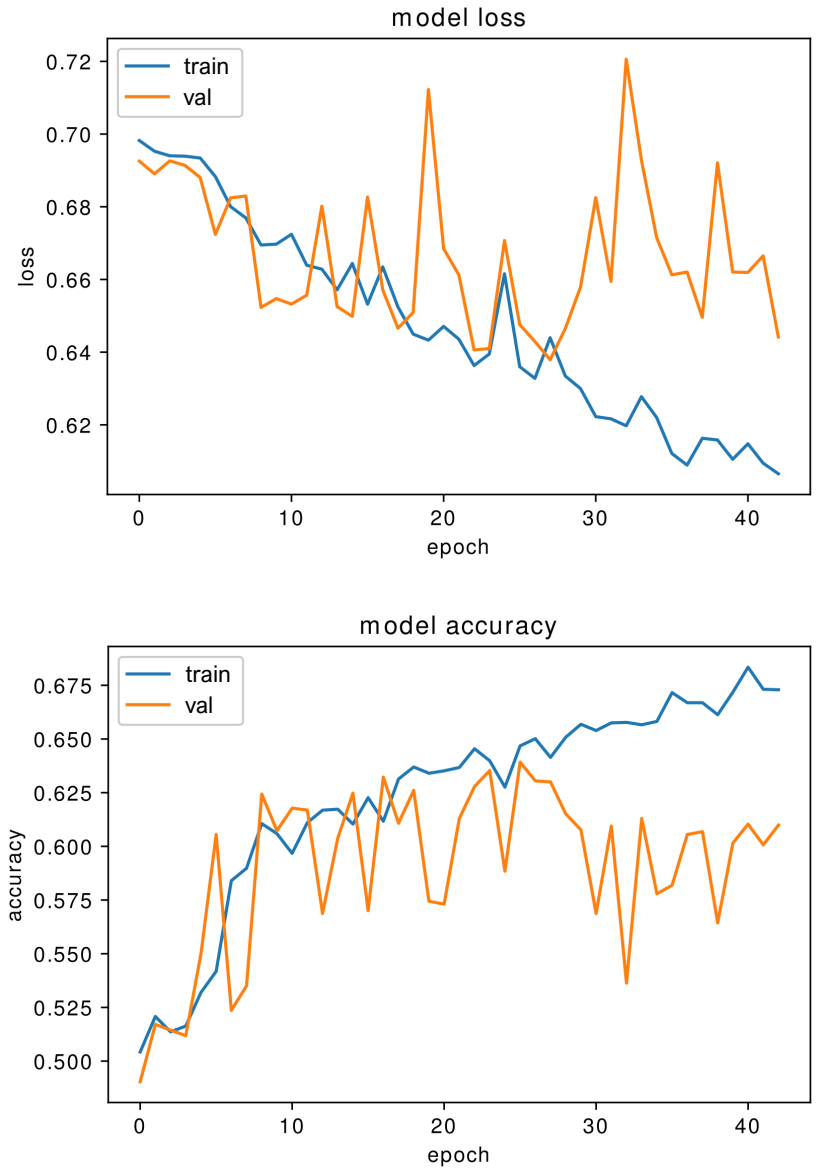


Figure 6.3: Loss and accuracy plots for model A.

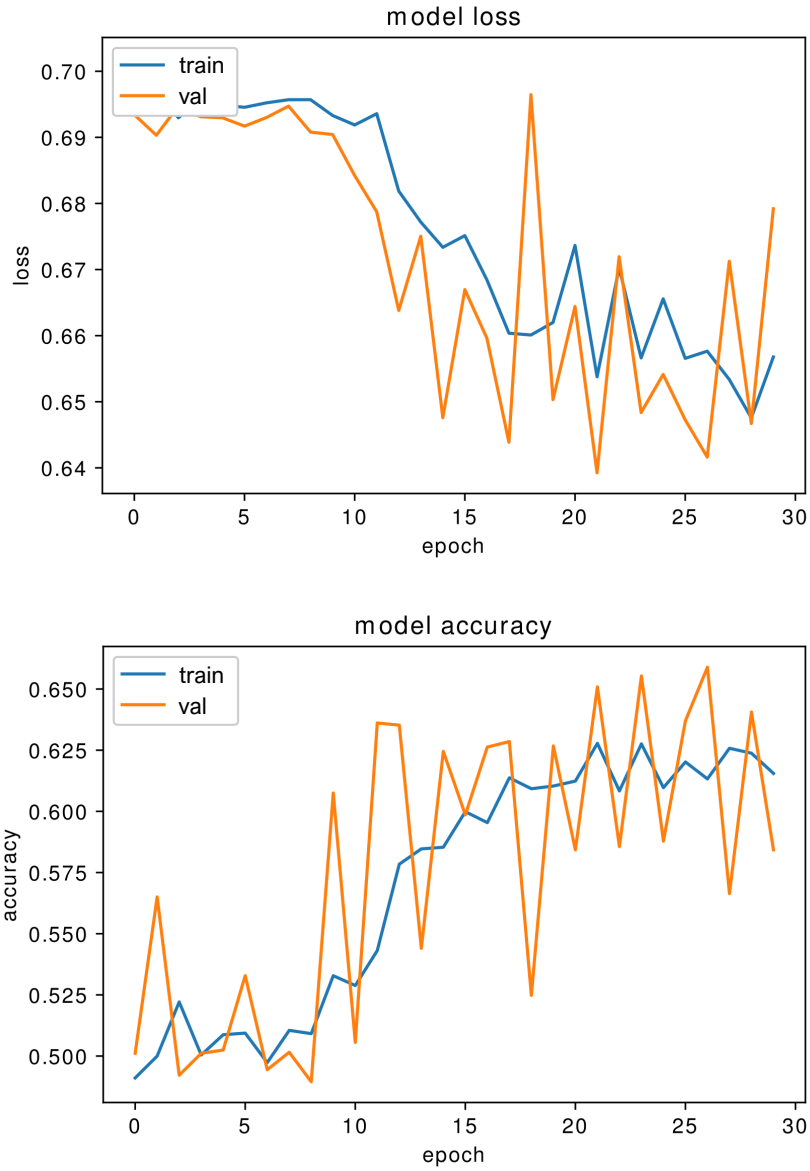


Figure 6.4: Loss and accuracy plots for model B.

6.2 Layers

To illustrate what the kernels in networks A and B focus on, one of the images from the test set was sent through layer C1 and M1 in the two nets. The original image can be seen in Figure 6.5. It was chosen since it contains areas of stroma, glands, both semi-structured and with low structure, as well as background. Furthermore, it represents the most common staining colours in the dataset. The label for the image is 0.

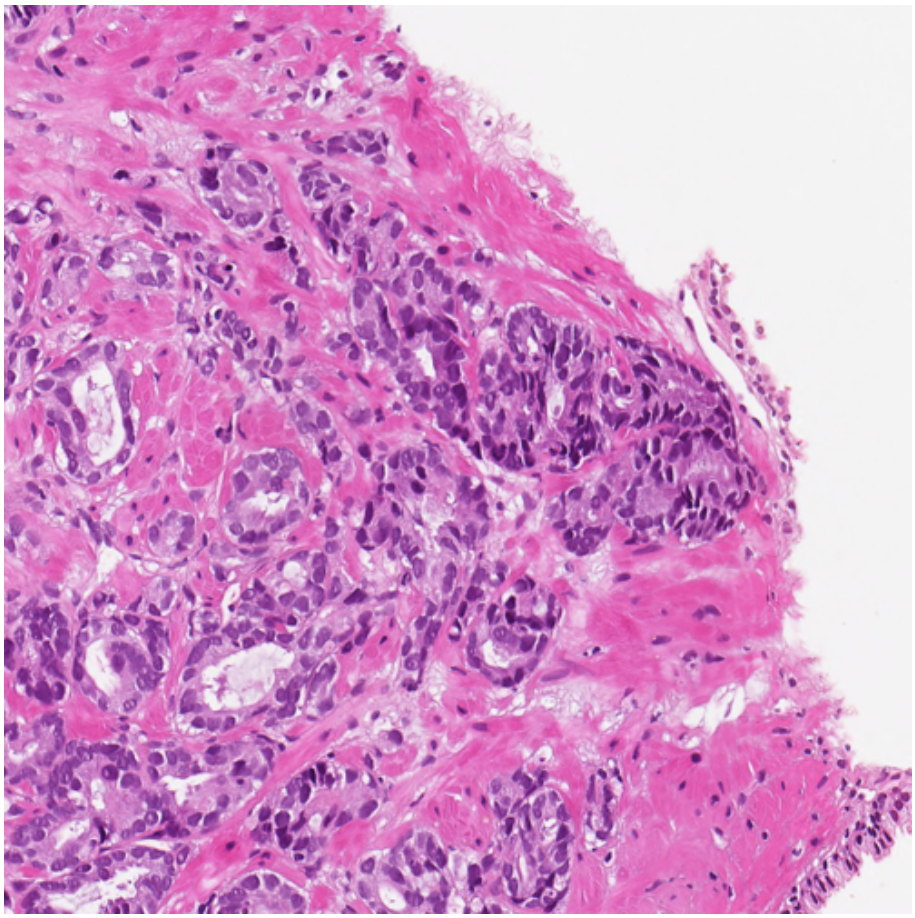


Figure 6.5: The original image sent through C1 and M1 of net A and L.

Figure 6.6 and 6.7 shows one of the images in the test set after the first convolutional layers of network A and B has been applied to it.

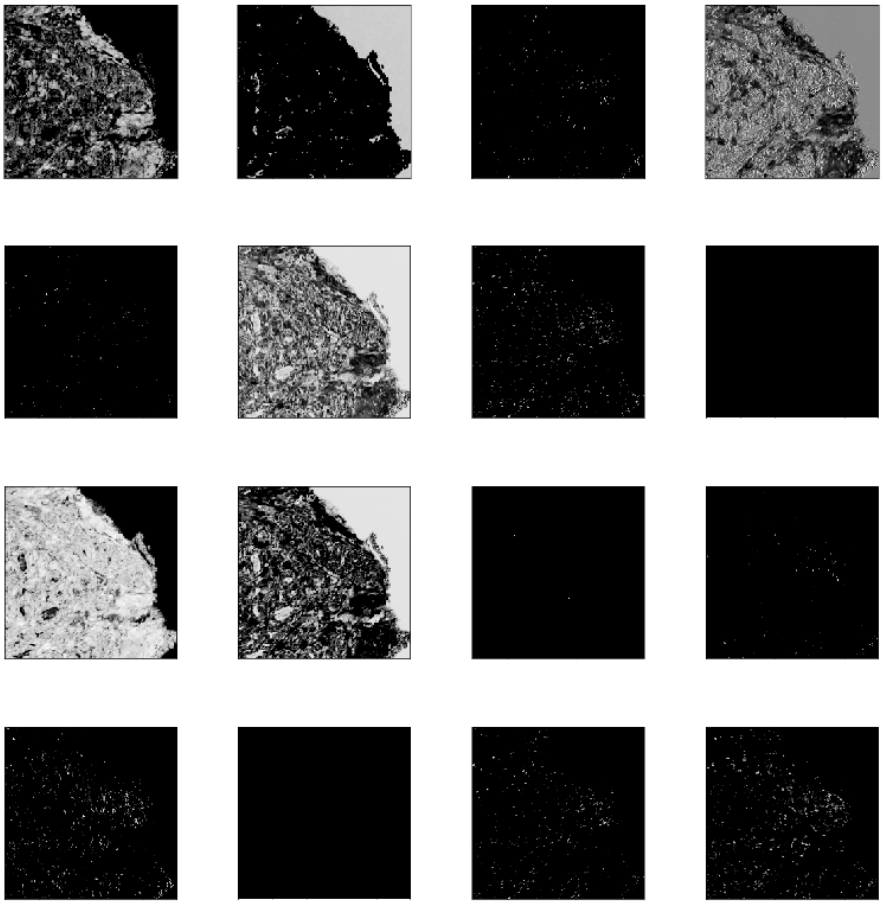


Figure 6.6: The first layer convolution from net A applied to one of the test images. The grey scale has been modified for better illustration of the features.

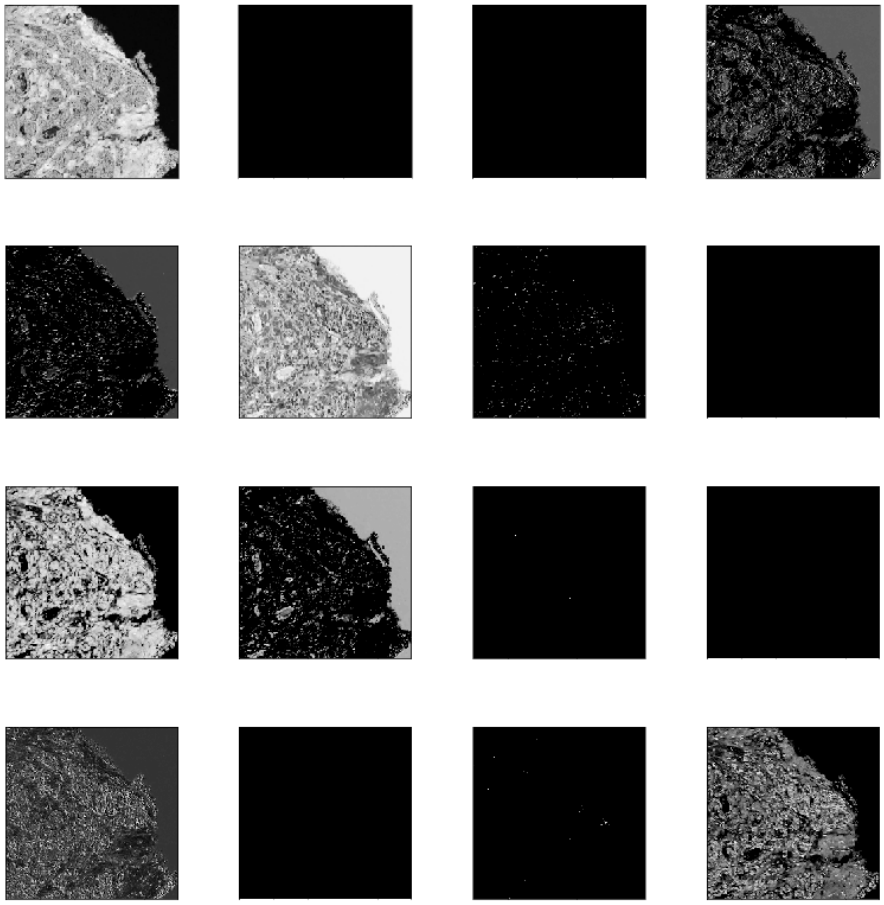


Figure 6.7: The first layer convolution from net B applied to one of the test images. The grey scale has been modified for better illustration of the features.

7
Discussion

The general scope of the thesis was to predicatively identify high risk cancer and investigate whether the benign tissue holds any vital information of the progression of the cancer or not. As the results indicate the classification was not trivial, although the nets seemed to learn something the models were not good at generalizing. Several tests were performed but none with results even remotely close to be of clinical use and only a few better than guessing, making the inter observer variability between pathologist small in relation. So why is the problem so hard? Is it only due to the difficulty in the dataset or more closely connected with the model choice and pre-processing of the data?

The first problem was how to handle all the data. Initially, all of the images were used, randomly distributed over a full needle biopsy, with no further weights introduced on where in the patients timeline they were taken. Because of the progression over time for each patient the latter images contain more up to date information meaning that they should be able to become more accurately classified and thus have larger importance during training. However, this statement is only derived from current diagnostic tools used by pathologists. Since the scope was to construct a predictive model we wanted to introduce as small a bias as possible based on current knowledge. The results from the all images dataset, tested on biopsies from the same biopsies as used for training, reached a maximum accuracy of 0.79. When more test samples were introduced and the test folder no longer contained patches from the same biopsies the accuracy dropped notably and stabilized around 0.5, meaning that it performed no better than guessing. An interesting fact can be derived from this, the staining for each biopsy introduces difficulties when it comes to automation of the process and that the differences in the cellular patterns of each patient contain variations that suppress possible cancer progression cues.

Furthermore, when studying the relevancy of the results in Table 6.1 we can see that less than 20% of the relevant instances of class 1 were correctly classified. This indicates that certain non-relevant features for the class selection was more prevalent in the training set for class 1 making the network focus on unwanted class separators. Another reason might be that the net was only able to find interesting sections in one fifth of the images due to the many images containing benign tissue far from any cancerous cells, indicating that the progression could not be seen here. To investigate which of the two possible facts being

the underlying reason for the low recall, biopsy and patient grading was introduced.

In Table 6.1 we can see that when combining all of the patches from one biopsy and averaging the results the accuracy does not change notably. For the patient grading the accuracy increases slightly but if we analyse Table 6.3 it is clear that the higher accuracy corresponds to a higher precision and recall for class 0 and not for the class 1 which we are interested in. Thus, we can conclude that network 1 cannot find any relevant features for the separation of the two classes if not introduced to the particular patient beforehand. Consequently, the dataset had to be constrained in some sense before further testing.

The second experiment was instead performed with patches only extracted from biopsies from the last available sample series, mainly because these were used by pathologists when deciding if the patients should be included in the AS or excluded and given further treatment. In other words, current diagnostic tools can spot large enough differences in the cellular patterns to make informed decisions about the cancer progression in these images. Using a dataset that was known to be separable into two classes could help us find important features for this separation. However, as the results indicate the model could not find any strong correlations even for patient gradings, see Table 6.4.

Since the pathologists use simple count of cancer occurrence in the biopsies as well as Gleason scores it would be reasonable to believe that it would not be until the final patient grading that the true score could be determined, particularly for patients with low Gleason scores that were excluded due to too many biopsies containing malignant tissue. The low precision and recall measures for class 1 in Table 6.4 indicate that non-relevant features are found in the dataset and out weight potential progression factors even if we tried limiting the dataset. Additionally, since patients with higher Gleason scores than 3+4 were all excluded, i.e. all part of the same dataset, it is evident that the network does not find and focus on gland structure of the cancerous tissue. Again, restraining the dataset would be needed in order to find relevant features.

Additionally, if we analyse the accuracy and loss plots for all of the three models the convergence of the curves for the train and especially validation set is highly irregular. This might suggest over-fitting to the training set and also that there is too little data to create a model that properly generalizes the classes.

For network 1 we can see that after about 30 epochs the loss function suddenly experiences a high increase in value. Due to the fact that the model uses smaller batches for optimization it is possible that the updated weights are moved in an unlucky direction where the overall loss increases. The choice of optimizer further determined how well the model recovers after such an incidence. Here Nadam is used and the momentum term might decrease the convergence rate so that early stopping comes into play before the loss has relocated to its lowest value.

Studying Figure 6.6 and 6.7 we can see what parts of the image the different layers focus on. Both for network A and B we can see that half or less of the kernels in the first layer focus on large reoccurring patterns such as stroma, background and gradients. Instead most kernels target smaller, less prevalent arrangements making a majority of the convoluted images almost entirely black with a few white dots. When comparing the original image with the convoluted representations we can see that most of the small patterns are edges

in certain angles and between strict colour-layers. The strong fixation of particular colour combinations might be one contributing factor why the network has trouble generalizing, especially if that exact colour combination was not used during training as it was for the first test with the all images dataset.

If we compare general patterns in Figure 6.6 and 6.7 we see that network B endorse larger general patterns whereas net A has a stronger focus on edges in different formations. Moreover, we can see that there are several convoluted images in both nets that are entirely black. One possible explanation for this is that the network is too large in relation to the dataset and that it over-trains on non-relevant features in the train set. Additionally, the respective kernels might be correlated with other colour combinations than the ones present in the original image in Figure 6.5. Reinforcing this statement is the fact that multiple convoluted images in Figure 6.6 and 6.7 are sparse and only have very few points with high intensity. This implies that, although rare in this particular image, there are relevant colour combinations in the train dataset that correspond to these filters making it credible to believe that there exists multiple such patterns, thus emphasising the importance of coherent staining when using digital pathology.

The models presented in Chapter 3.3 that were able to predicatively classify cancer recurrence after RP only incorporated benign tissue closely surrounding the malignant tissue. It is possible that progression only can be seen in benign tissue that lie close enough to the cancer cells both spatially and temporally. When not restricting the patches enough the in class variability was perhaps too large for the model to be able to find underlying predictive features for progression.

Following, it might not be appropriate to use randomly extracted patches for classification but instead focus on interesting areas such as cancerous regions or glandular structures. To properly determine which sections contain such interesting information it would make sense to let pathologists study and try to classify these biopsies thus also making the results comparable with human diagnostics.

Even if the final score of the patients for the second dataset indicate low correlations the train and test folders should contain more samples in order to produce more substantial results. As stated in section 2.5 CNNs tend to over-fit if they are given too much room to bend to the training data but they also rely on the dataset being large and diverse enough in order to generalize properly. Time was a limiting factor when writing this thesis constraining the amount of available data to around half the AS cohort from SUS. Incorporating all 180 patients would hopefully make the model fit better hence giving more reliable results.

Conclusively, the presented models were not able to predicatively determine which patients that would in time get high risk cancer progression. Even as Gleason grading has its restraints and the inter observer variability between pathologists is high this thesis was not able to find any alternative cellular structures that could be used in clinical medical analysis to help with diagnostics.

7.1 Future work

Since it has not, to the author's knowledge, been tested to make a predictive model for the AS cohort there is a lot of research that is still to be done. Firstly, restricting the dataset to only contain malignant and closely surrounding benign tissue might help the network focus on relevant patterns to correctly split the classes instead of finding larger patches that are present in both but favourable for one due to more instances in the training data.

Furthermore, since the introduced model did not find Gleason grading of importance when predicting the outcome it might be fitting to ensemble multiple networks, biased on different things such as Gleason grade, nuclei density and relevancy based on the current patients timeline, to create a true patient grading.

Finally, it should be investigated how the images could be used in a time series for better predictions. It is reasonable to believe that how the tumours development over time differs for high and low risk cancer and that the same tumours could be seen in multiple longitudinal biopsies. However, for this to be possible more test subject need to be included in the dataset to restrict over-fitting of the available patients. Additionally, detailed annotations over cancer cells and their positions would be needed.

Bibliography

[1] American Urological Association. Prostatic adenocarcinoma: Gleason grading (modified grading by isup). https://www.auanet.org/education/auauniversity/education-products-and-resources/pathology-for-urologists/prostate/adenocarcinoma/prostatic-adenocarcinoma-gleason-grading-(modified-grading-by-isup). Accessed: 2019-03-18.

[2] Leonard P Bokhorst, Riccardo Valdagni, Antti Rannikko, Yoshiyuki Kakehi, Tom Pickles, Chris H Bangma, Monique J Roobol, PRIAS study group, et al. A decade of active surveillance in the prias study: an update and evaluation of the criteria used to recommend a switch to active treatment. *European urology*, 70(6):954–960, 2016.

[3] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[4] Carlos Cordon-Cardo, Angeliki Kotsianti, David A Verbel, Mikhail Teverovskiy, Paola Capodieci, Stefan Hamann, Yusuf Jeffers, Mark Clayton, Faysal Elkhettabi, Faisal M Khan, et al. Improved prediction of prostate cancer recurrence through systems pathology. *The Journal of clinical investigation*, 117(7):1876–1883, 2007.

[5] George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8609–8613. IEEE, 2013.

[6] Scott Doyle, Mark Hwang, Kinsuk Shah, Anant Madabhushi, Michael Feldman, and John Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1284–1287. IEEE, 2007.

[7] Timothy Dozat. Incorporating nesterov momentum into adam. https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ, 2016. Accessed: 2019-04-20.

[8] Jonathan I Epstein. An update of the gleason grading system. *The Journal of urology*, 183(2):433–440, 2010.

[9] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.

[10] Jonathan I Epstein, Michael J Zelefsky, Daniel D Sjoberg, Joel B Nelson, Lars Egevad, Cristina Magi-Galluzzi, Andrew J Vickers, Anil V Parwani, Victor E Reuter, Samson W Fine, et al. A contemporary prostate cancer grading system: a validated alternative to the gleason score. *European urology*, 69(3):428–435, 2016.

[11] Jonathan Ira Epstein. A new contemporary prostate cancer grading system: message to the italian pathologists. *Pathologica*, 107(3-4):205–207, 2015.

[12] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. Accessed: 2019-03-30.

[14] Anna Gummeson. Prostate cancer classification using convolutional neural networks. *Master's Theses in Mathematical Sciences*, 2016.

[15] Anna Gummeson, Ida Arvidsson, Mattias Ohlsson, Niels Christian Overgaard, Agnieszka Krzyzanowska, Anders Heyden, Anders Bjartell, and Kalle Aström. Automatic gleason grading of h and e stained microscopic prostate images using deep convolutional neural networks. In *Medical Imaging 2017: Digital Pathology*, volume 10140, page 101400S. International Society for Optics and Photonics, 2017.

[16] Metin N Gurcan, Laura Boucheron, Ali Can, Anant Madabhushi, Nasir Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147, 2009.

[17] DW Hosmer and S Lemeshow. *Applied Logistic Regression (2nd Edition)*. John Wiley & Sons, New York, NY, 2000.

[18] Peter A Humphrey. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern pathology*, 17(3):292, 2004.

[19] Hanna Källén, Jesper Molin, Anders Heyden, Claes Lundström, and Kalle Åström. Towards grading gleason score using generically trained deep convolutional neural networks. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1163–1167. IEEE, 2016.

[20] Keras. Keras: The python deep learning library. https://keras.io/, 2019. Accessed: 2019-05-05.

[21] JB Lattouf and F Saad. Gleason score on biopsy: is it reliable for predicting the final grade on pathology? *BJU International*, 90:694–699, 2002.

[22] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[23] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[24] George Lee, Robert W Veltri, Guangjing Zhu, Sahirzeeshan Ali, Jonathan I Epstein, and Anant Madabhushi. Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: preliminary findings. *European urology focus*, 3(4-5):457–466, 2017.

[25] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6:26286, 2016.

[26] Josefin Persson, Ulrica Wilderäng, Thomas Jiborn, Peter N Wiklund, Jan-Erik Damber, Jonas Hugosson, Gunnar Steineck, Eva Haglind, and Anders Bjartell. Interobserver variability in the pathological assessment of radical prostatectomy specimens: findings of the laparoscopic prostatectomy robot open (lappro) study. *Scandinavian journal of urology*, 48(2):160–167, 2014.

[27] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[28] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.