Utrecht University

Bachelor's Programme in Artificial Intelligence

Machine Learning as a classification tool for Histopathological analysis

by

[D. den Heijer d.heijer@students.uu.nl]

Abstract Decision support systems and automated classification are a vast area of research in the intersection of medicine and artificial intelligence. Significant accomplishments are delivered in the last decade, thanks to the revolution of computer vision. Automated classification of histopathological slides can potentially provide well-needed assistance to overworked pathologists, improve classification accuracy, and reduce inter and intra-observer error. Research proves that convolutional neural networks outperform pathologists on several tasks. However, these promising results are not yet integrated into the pathology department because of some intricate problems. The integration of deep learning algorithms can potentially enhance diagnostics and efficiency, which in turn makes healthcare more accessible and affordable.

Bachlor's Thesis (7.5 credits ECTS)
July 2019
Supervisor: Prof. Dr. Ir. M.J.C. Eijkemans
Examiners: Prof. Dr. Ir. M.J.C. Eijkemans
Dr. B.G. Rin
Word Count: 6030

Contents

1

Introduction

1.1 Research Problem

Decision support systems and automated classification are a vast area of research in the intersection of medicine and artificial intelligence. Significant accomplishments are delivered in the last decade, thanks to the revolution of computer vision. Sophisticated algorithms already outperform human experts on tasks as image classification Krizhevsky et al. (2012) and facial recognition Taigman et al. (2014).

Recent techniques in whole slide image processing[1] and cost reduction in digital storage has made the digital evaluation of stained tissue feasible. Combined with the improvement of computer power, this created an entirely new field on the cutting-edge of medicine and artificial intelligence. This area produced impressive results in the past few years, such as the automated detection of melanoma Esteva et al. (2017), classification of retinopathy Gulshan et al. (2016) and neural networks outperforming radiologists Rajpurkar et al. (2018).

Models created for diagnostic purposes, need to be more explainable and reliable than those in many other fields. It is less critical why Google recommends a website based on a query, but when an algorithm diagnoses an individual with severe disease, the patient would like an explanation of why it decided to do so. That process is called the explainability of a network and is an important area of research Montavon et al. (2018).
Most of these models are trained using supervised learning. In the case of histopathological analysis, this means that the input is composed of stained tissue slides, and the corresponding output labels are benign or malign [2]. The model is trained with these examples and gets tested on new data afterward.

Classifying microscopic images is a labor-intensive task and has a relatively sizeable inter-and intra-observer error Jackson et al. (2017), Su et al. (2016). Recent research proves that automated classification of histopathological tissues has an error rate similar or even better than a highly trained pathologist Bejnordi et al. (2017).

[1] The process of scanning and digitally representing tissue slides. For more information, see chapter 3: Histopatholgy

[2] This is just for the sake of an example. More sophisticated models can attach a probability to the chance of tissue being malign.
A central definition in pathology is the difference between benign and malign. Benign is good, and malign is bad. Cancer is malign; a tumor can be either malign or benign.

So why is the pathologist still looking through his microscope? The simple explanation is that not all these promising results are used in day to day hospital care. There is a significant distinction between theoretical results and the implementation of such algorithms. Several factors attribute to this, such as the black box problem, the cost of pathological automatization and implementation issues.

'A successful solution would hold great promise to reduce the workload of pathologists while at the same time, reduce the subjectivity in diagnosis Madabhushi & Lee (2016).' The ability to mine "sub-visual" image features from digital pathology slide images, features that may not be visually discernible by a pathologist, offers the opportunity for better quantitative modeling of disease appearance and hence possibly improved prediction of disease aggressiveness and patient outcome Madabhushi & Lee (2016).

There is a pressing need for computer-assisted diagnoses to relieve the workload on pathologists by sieving out benign areas so that pathologists can focus on the more challenging to diagnose suspicious cases. For example, approximately 80% of the 1 million prostate biopsies performed in the United States every year are benign; this suggests that prostate pathologists are spending 80% of their time sieving through benign tissue Gurcan et al. (2009).

1.2 Aim and Scope

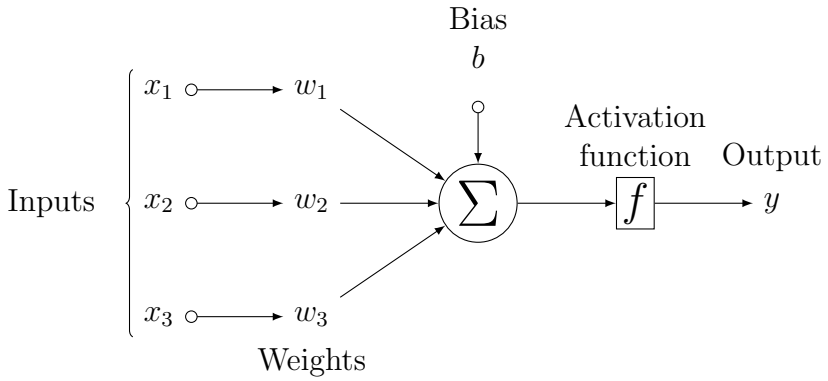
This thesis aims to explore the status quo of histopathological classification and discuss prospects. Does the classification of histopathological images work, and why is it not used in daily hospital care? Subsections within this question that are addressed are the following: Methods of classifying; State of the art classification results; Theory of classification; Current pitfalls and how to address these in the future.

2

Neural Networks

2.1 Neural networks

A neural network (NN), also called a multilayer perceptron, is loosely inspired by the human brain. A typical NN consists of an input layer, one or more hidden layers, and an output layer. Each of these layers holds n nodes, and all the nodes inside these layers are connected with weights. The weights of the network can be thought of as a line connecting a node in layer l to a node in layer $l-1$ — the higher the weight, the more value it represents in the next node. All nodes have a certain threshold that determines whether the node should be active or not. This threshold is called the bias. A single neuron receives its input from the output of its n predecessors, multiplied with n corresponding weights and an added bias b. This value goes through an activation function f before it transmits its output to the neurons in the next layer.



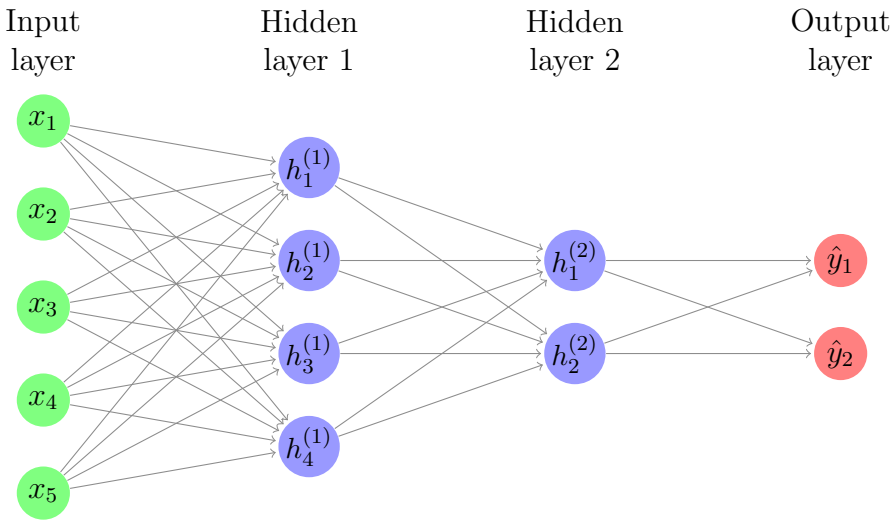
In the figure above, a single neuron is figured. It receives input from three input nodes $\vec{\mathbf{x}} = [x_1, x_2, x_3]$, these input values are multiplied with their corresponding weights $\mathbf{w}^T = [w_1, w_2, w_3]$ which results in the input of the neuron. After that, the bias is added, and the product sum is passed through the activation function before generating its output. An often used activation function is the sigmoid:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

The reason for its popularity is that it is easily differentiable and squashes the output of each node between 0 and 1. Activation functions allow the model to introduce

non-linearity into it. This means that without using an activation function, the model can only represent linear functions.

Below a simple neural network with one input layer, two hidden layers, and one output layer is shown. When for example, the task of an NN is to classify the presence or absence of a tumor, the output layer consists of two neurons representing the two output values: 1. present or 2. absent.



To further formalize neural networks, a few definitions are needed:

w_{ij}^k: weight for node j in layer l_k for incoming node i.
b_i^k: bias for node i in layer l_k.
a_i^k: product sum plus bias for node i in layer l_k.
o_i^k: output for node i in layer l_k.
r_k: number of nodes in layer l_k.

The matrix equation for calculating the output values of the nodes in a layer is denoted as:

$$\sigma = \left(\begin{bmatrix} o_0^{l-1} \\ o_1^{l-1} \\ \vdots \\ o_n^{l-1} \end{bmatrix} \begin{bmatrix} w_{0,0} & w_{0,1} & \dots & w_{0,n} \\ w_{1,0} & w_{1,1} & \dots & w_{1,n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{k,0} & w_{k,1} & \dots & w_{k,n} \end{bmatrix} + \begin{bmatrix} b_0^l \\ b_1^l \\ \vdots \\ b_k^l \end{bmatrix} \right)$$

The first vector represents all output of the neurons in the preceding layer. The matrix consists of all the weights connecting the neurons in the previous layer to the current one, and the last vector contains the biases corresponding to each neuron. The σ represents the activation function that is applied to all the product sums, generating the output value for each neuron in the current layer. A short notation of the formula for this calculation reduces to the following math:

$$\mathbf{o}^l = \sigma(\mathbf{W}_l \mathbf{o}^{(l-1)} + \mathbf{b}_l) \tag{2.2}$$

2.2 Training a network

After training for sufficient epochs[1], we want the generated output label to resemble the expected output label, i.e., the network classifies the input correct. To train for sufficient epochs does not sound very accurate, and it indeed is not. The number of epochs is one of the hyperparameters of a NN. Hyperparameters make up an essential part of training a network. More on this is addressed later in this chapter. To achieve the goal of classifying the input correct, we want the network to learn. What we mean by learning is that the network adjusts all its parameters in such a manner that it produces the desired output for a given input. To train the network in accomplishing this resemblance is an optimization problem that uses a cost function to measure how far away the network is from accurately classifying the input with its corresponding label. To train the network, we need a few components: a feedforward neural network θ, a cost function, an optimization tool, hyperparameters and a dataset consisting of N input-output pairs denoted

$$X = \{(\vec{\mathbf{x}}_1, \vec{\mathbf{y}}_1), \ldots, (\vec{\mathbf{x}}_N, \vec{\mathbf{y}}_N\}) \tag{2.3}$$

where $\vec{\mathbf{x}}_i$ is the input and $\vec{\mathbf{y}}_i$ is the desired output for input $\vec{\mathbf{x}}_i$.

2.2.1 Forward propagation

As briefly addressed above, one of the ingredients required to train the network is a feedforward neural network. The method by which we propagate over this network is called forward propagation. As the name suggests, the input data is fed in the forward direction through the network. Each hidden layer accepts the input data, processes it as per the activation function, and passes its output to the succeeding layer. Forward propagation generates the desired output values we can use to calculate the cost function. This process is repeated every epoch to calculate the current error after the weights have been updated.

2.2.2 Gradient Descent

Gradient descent is a technique for minimizing the above-mentioned cost function. A typically used analogy to describe Gradient Descent is a ball rolling down an irregular surface. If the ball is placed on a hill, it will roll down, coming to rest at the bottom of a valley Abu-Mostafa et al. (2012). Note that if there exists only one minimum, the ball will always roll into this global minimum. However, in most cases, the function will be made up of many variables and contains multiple local optima. Therefore, finding this global optimum is not always feasible.
To minimize the error, gradient descent needs a cost function. This cost function can best be imagined as a 'surface' in a high-dimensional space.
Minimizing the cost function implies minimizing the error rate and therefore maximizing the accuracy. An often used cost function applied to gradient descent is the

[1] An epoch is the process in which all the input data has been propagated once through the network

Mean Squared Error (MSE). MSE is defined as:

$$C(\hat{y}, y) = \frac{1}{2} \sum_{i} (\hat{y}^{(i)} - y^{(i)})^2 \tag{2.4}$$

Note that if the expected output of the network differs a lot from the generated output, the cost will be high, and if it close to the expected output, the cost is small.
Gradient Descent works by calculating a gradient, taking a step in the direction of this gradient multiplied by the learning rate and repeats this process until it ends up at a minimum. The algorithm for efficiently calculating this gradient is called backpropagation.

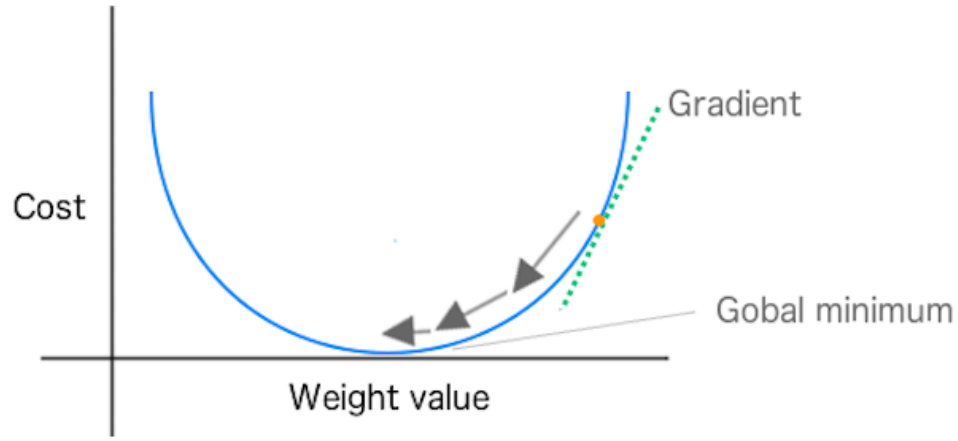


Figure 2.1: A simplified visual representation of gradient descent Patterson & Gibson (2017)

2.2.3 Backpropagation

Backpropagation efficiently calculates the gradient for the gradient descent algorithm. It is a method used in artificial neural networks to calculate the gradient that is needed to adjust the weights and biases used in the network. The backpropagation algorithm decides how much to update each weight and bias of the network after comparing the predicted output to the desired output given an example. Accordingly, it is needed to compute how the error changes with respect to each weight. For a more mathematical description of backpropagation, we refer to appendix B.

2.2.4 Hyperparameters

Hyperparameters, as presented above, are a critical element in training the network. Hyperparameters are the variables that determine the networks structure and the variables that determine how the network is trained. These parameters are set before the training process begins. This is contrary to a parameter, which is changed during the training of the network Goodfellow et al. (2016).

Network Structure

1. Number of hidden layers and nodes in each hidden layer.
   A common way to determine how many hidden layers to add is to keep adding them until the test error does not improve anymore. A to shallow network can cause underfitting.

2. Activation function
   Until now, the only activation function discussed is sigmoid. However, many activation functions can be used to optimize the performance of the network.

Training Parameters

1. **Learning rate**
   The learning rate α is already briefly addressed in the section about Gradient Descent. In plain English, this parameter determines how fast the network learns. If it learns to fast, it might overshoot the optimum and might not converge. If it learns too slow, it will converge but slows down the learning process. A popular way to deal with this trade-off is to use a learning rate proportionate to the slope of the gradient vector. 'The size of the learning rate is limited mostly by factors like how curved the cost function is. You can think of gradient descent as making a linear approximation to the cost function, then moving downhill along that approximate cost. If the cost function is highly non-linear (highly curved) then the approximation will not be very good for very far, so only small step sizes are safe' Goodfellow et al. (2016). Those step sizes are determined by the batch-size.

2. **Batch size**
   The batch size is the number of input-output pairs that will be propagated backward through the network, before the weights and biases are adjusted.
   'Figure 2-2 shows the paths taken by three Gradient Descent algorithms in parameter space during training. They all end up near the minimum, but Batch GD's path actually stops at the minimum, while both Stochastic GD and Mini-batch GD continue to walk around. However, don't forget that Batch GD takes a lot of time to take each step, and Stochastic GD and Mini-batch GD would also reach the minimum if you used a good learning schedule' Goodfellow et al. (2016).

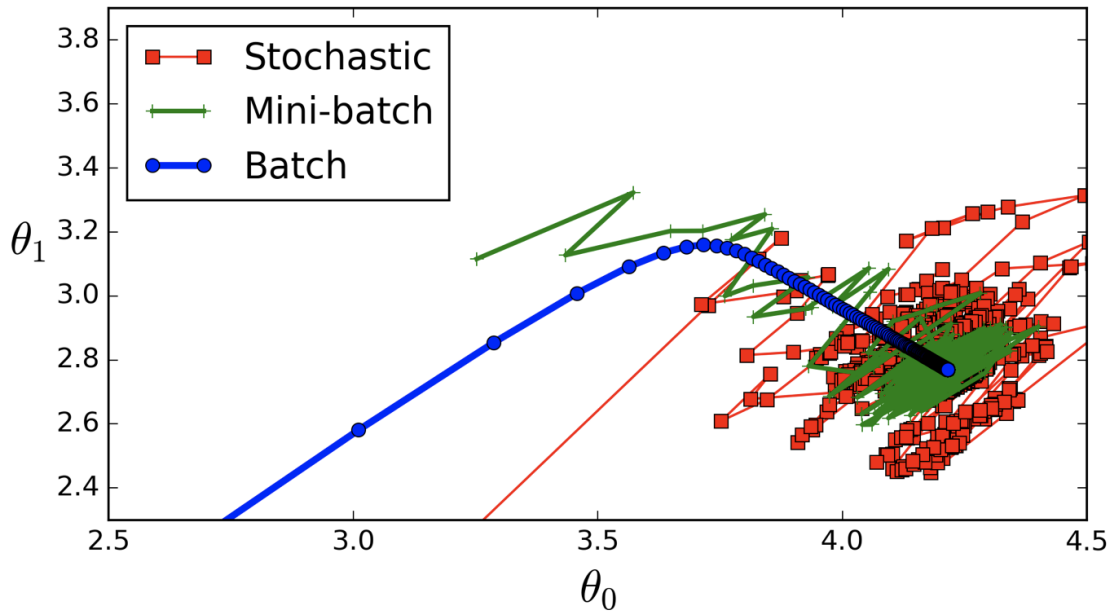


Figure 2.2: Comparison of Stochastic GD, Mini-Batch and BarchGéron (2017)

3. **Number of epochs**
   An epoch is a complete learning cycle in which the network processes all training examples. A rule of thumb is to increase the number of epochs until the validation accuracy starts decreasing.

2.2.5 Regularisation

Regularisation is particularly important in neural networks with a huge amount of parameters. Convolutional neural networks, the networks that are exceptionally good at recognizing images, consist of an explosive amount of parameters because of the enormous input space. Regularisation techniques are used to avoid overfitting in these networks.

Dropout

Dropout is one of the most common techniques to avoid overfitting in deep NN's. As the name suggests, it drops out neurons in a particular layer according to a probability. By using this technique, the network does not rely too much on a specific set of features, which could induce overfitting.

2.3 Deep Learning

Deep learning is a subset of neural networks. A deep neural network is a neural network with a lot of hidden layers. A commonly used definition is that a neural network is 'deep' when it has more than two hidden layers. So why is a deep neural network better than a non-deep neural network? That is a topic of discussion in research and is still undecided. However, empirical results show that a deep NN classifies better than a shallow one and therefore it is commonly used.

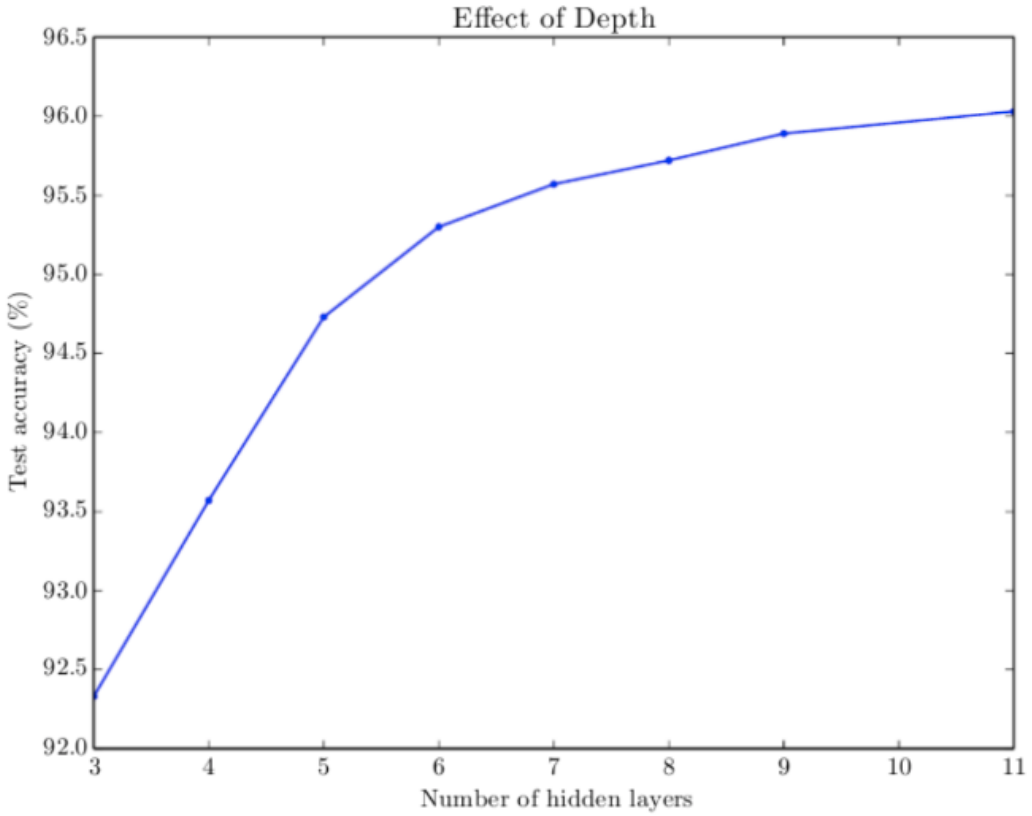


Figure 2.3: Empirical results showing that deeper networks generalize better when used to transcribe multi-digit numbers from photographs of addresses. The test set accuracy consistently increases with increasing depth Goodfellow et al. (2016).

2.4 Computer Vision

For more specific tasks, specialized ANN's are created, such as Convolutional Neural Networks to roughly emulate the human visual system. People are continually looking at the world surrounding them and are subconsciously making predictions about what they see. This seeing apparatus is trained during their life and is the reason a trained pathologist can make accurate predictions whether a particular piece of tissue is malign or not. This task has recently shifted to computed processing of visual information, hence computer vision. Computer vision is the branch that has to do with images and videos. It seeks to automate tasks the human optical system can do. An artificial neural network specialized in this task is the convolutional neural network.

2.4.1 Convolutional Neural Networks

A convolutional neural network (CNN) is a specialized version of a neural network. It is called convolution because it convolves over an image. A CNN is especially useful in classifying images, the reason why it is so popular and successful in classifying histopathological slides. A few reasons why CNN's work so good are
(i) the availability of much more extensive training sets, with millions of labeled examples;
(ii) robust GPU implementations, making the training of huge models practical and

(iii) better model regularization strategies, such as dropout.
In this section, we explore how a CNN sees and understands the images we feed it. The function of a CNN consists out of two important tasks: extracting features through the convolutional and pooling layers, and classification through its fully-connected layers. The more convolutional layers used, the more intricate the detected features will be. The first layers identify lower level features like edges and orientation. The deeper we look into the network, the more high level features patterns we encounter. To illustrate what a low-level feature is, two pictures of a dog are presented below.

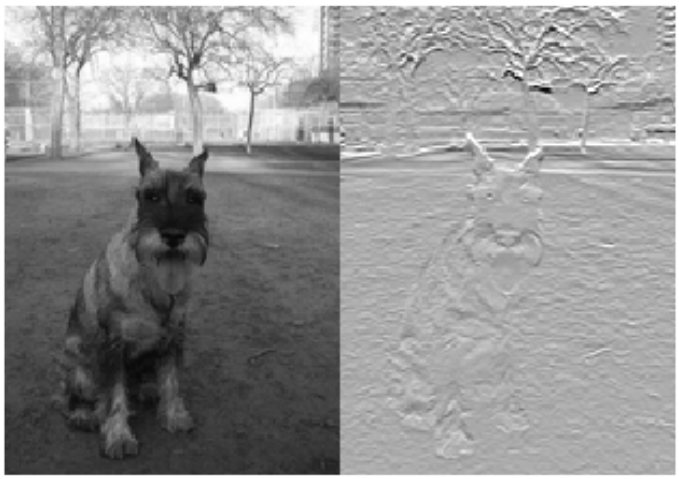


Figure 2.4: Filtered image of a dog using horizontal edge detection.

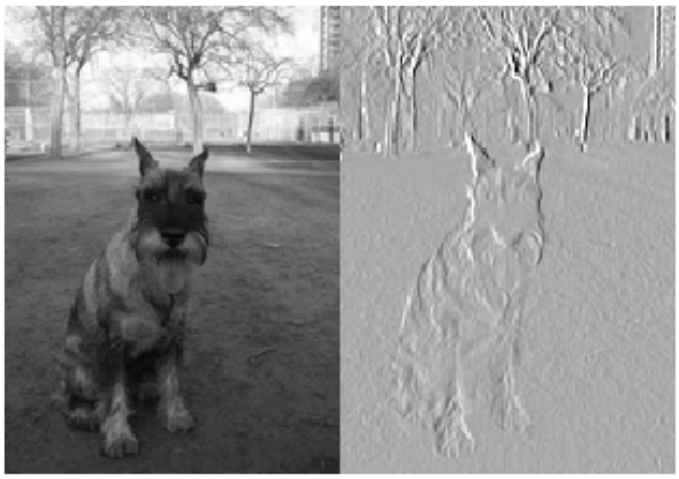


Figure 2.5: Filtered image of a dog using vertical edge detection.

This is essentially what a filter does in a CNN. In the first image of the dog, the filter detects horizontal edges and in the second one, it detects vertical edges.

At the most basic level, a CNN is a special sort of neural network that contains at least one convolutional layer. A common CNN receives an input image, runs it through a number of convolutional layers, a nonlinear activation function, one

or more pooling layers (downsampling), and a feedforward network and outputs a classification label.
A computer does not recognize images the same way as humans do. Their input consists entirely out of numbers. More specifically, a 2-dimensional array for black and white photos. For colored images, the data consists of a 3-dimensional array — the red-blue-green pixel values make up the third dimension in this array.

The first layers in a CNN are looking for simplistic features such as horizontal or vertical edges that can be seen in the pictures of the dog above. The deeper layers are looking for more complex patterns.
Pooling is used to downsize the image; it reduces the number of hyperparameters and therefore reduces the computation required for training the network.

2.4.2 Convolutional layer

The convolution layer uses filters that perform convolution operations as it scans the input image with respect to its dimensions. The hyperparameters of the convolutional layer consist of the filter size and stride. The resulting output O is called the feature map, this map has all the features calculated from the Hadamard product in the input layers and filters. In the figure below, this process is portrayed.

Mathematically, the convolution operation is quite simple. Suppose we have N x N matrices called A and B.
Let M_{ij} denote the entry in the i^{th} row and j^t column of matrix M. A $\bullet$ B is called the Hadamard product and multiplies each index in matrix A with the corresponding index in matrix B.

Stride

The stride S is defined as the number of squares in the matrix to move after each convolution.

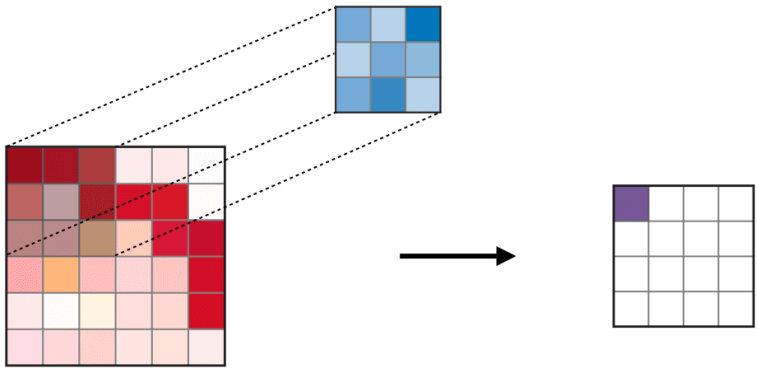


Figure 2.6: The working of a convolutional layer Afshine Amidi (2018)

2.4.3 Pooling layer

A pooling layer is used for the downsampling of features, typically applied after a convolution layer. It reduces the dimensionality of each feature map to reduce the number of parameters and computations in the network, therefore controlling overfitting. There are two types of operations that a pooling layer can do; max and

average pooling, where the maximum and average value of the features are taken respectively. A pooling layer summarises the features of the convolutional layer, this leads to better generalization. The reason for its functionality is that a convolutional layer represents a very specific feature of a particular image. The pooling layer is used to represent this as a more general feature that is less prone to rotations and locations in the picture.

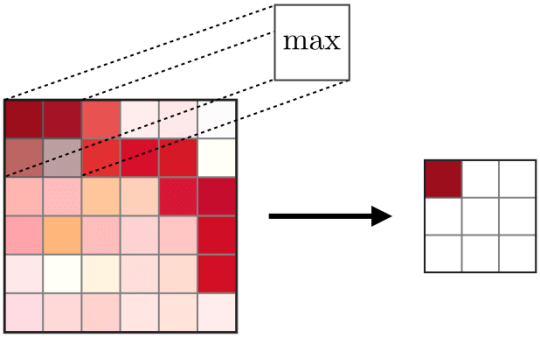


Figure 2.7: Max pooling Afshine Amidi (2018)

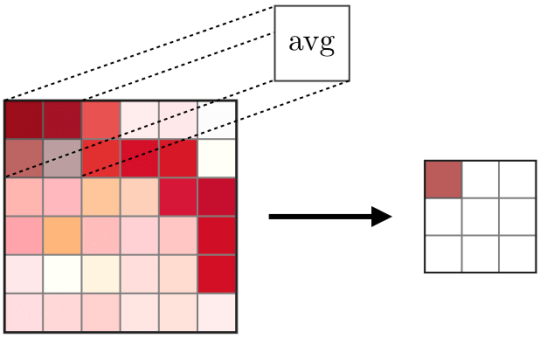


Figure 2.8: Average pooling Afshine Amidi (2018)

2.4.4 Forward feeding layers

The convolutional and pooling layers represent the high-level features of the input. The forward feeding layers use these features for classifying the input image into a class based on its training examples. Forward feeding layers also intro the non-linearity as discussed above. An FC layer is a plain vanilla NN attached to the convolutional and pooling layers.

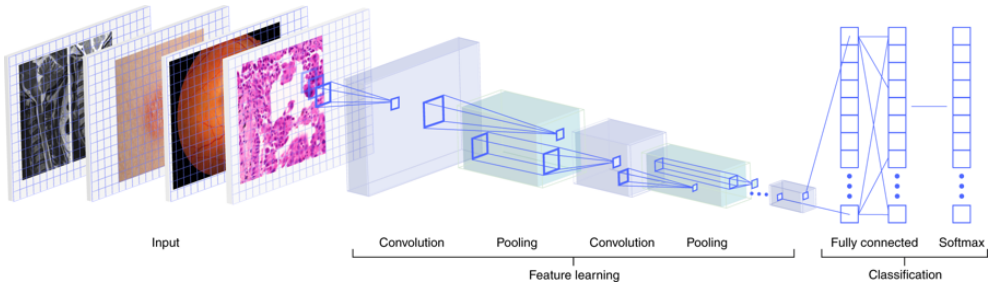


Figure 2.9: Overview of a convolutional neural network consisting of the various elements described above Esteva et al. (2019).

3

Histopathology

To get a better grasp on the intuition behind the automated classification of histopathological slides, a brief introduction is needed in histology (from the Greek histos, 'tissue' and logica, 'study') and histopathology (from the Greek pathos, 'suffering'). Histology is the study of the microscopic structures of tissues, whereas histopathology is the study of diseased cells. Histopathology plays a vital role in today's practice of medicine and is used as a gold standard for many diseases, including nearly all types of cancer. To fulfill such a diagnostic assessment, an individual first has to follow a six-year residency after a medical degree to become a pathologist. Most pathologists even specialize in a specific tissue afterward. The ability to classify tissue accurately requires at least twelve years of practice. We can, therefore, conclude that the classification of pathological tissue is a complicated task. The start of the pathological process begins with the collection of the tissues. This may be done using surgery, biopsy, or autopsy. After that, the tissue is either fixed using chemical fixation or frozen. To make the tissue visible, various pigments are used to stain the tissue. The most frequent used pigments are hematoxylin and eosin given the tissue its typical pink, purple look. Hematoxylin is used to stain nuclei blue, while eosin stains cytoplasm and the extracellular connective tissue pink Griffin & Treanor (2017).

3.1 Tumor Grade Classification

As discussed before, the job of a pathologist is not to solely assign the label benign or malign to a tissue slide. The treatment and 5-year survival depend primarily on the grade of the tumor. The tumor grade is the description of a tumor based on how abnormal the cells look under a microscope. The abnormality is an indicator of how quickly a cancer is likely to grow and spread. The factors used to determine tumor grade differs between types of cancer NCI (2013). An important distinction to make is that the grade of a tumor is not the same as the stage of cancer. Cancer stage refers to the size and whether or not the primary tumor has spread. Cancer stage is based on factors as location, tumor size, lymph node involvement, and the number of tumors.
Depending on the type of cancer, different grading systems are used. 'In general, tumors are graded as 1, 2, 3 or 4, depending on the amount of abnormality. In Grade 1 tumors, the tumor cells and the organization of the tumor tissue appear close to normal. These tumors tend to grow and spread slowly. In contrast, the cells

and tissue of Grade 3 and Grade 4 tumors do not look like normal cells and tissue. Grade 3 and Grade 4 tumors tend to grow rapidly and spread faster than tumors with a lower grade.' NCI (2013)

If a grading system for a tumor type is not specified, the following system is generally used Edge et al. (2010):

GX: Grade cannot be assessed (undetermined grade)

G1: Well-differentiated (low grade)

G2: Moderately differentiated (intermediate grade)

G3: Poorly differentiated (high grade)

G4: Undifferentiated (high grade)

In order to develop a particular treatment and determine a patient's prognosis, tumor grade plays a vital role. Generally, a lower grade indicates a better prognosis. Higher-Grade cancer may grow and spread more quickly. It is therefore important to not only label a tumor benign or malign but also detect the grade NCI (2013).

3.2 Whole-Slide Images

The rapid improvement in whole-slide imaging (WSI) technologies has transformed pathology in the last few years. Following the FDA approval of the Philips IntelliSite imaging system in 2017, whole-slide images may now be used for primary clinical diagnosis. It is an essential step towards the complete digitalization of pathology. Pathologists have been using WSI since the 1980s for remote pathology diagnosing using digital image transmission. WSI is high-resolution digitization and storage of entire glass pathology slides as digital slides. 'These images can be stored, viewed locally, or transmitted over a network for remote viewing Griffin & Treanor (2017).' Remote viewing is a great advantage to consult an expert pathologist for ambiguous cases.

The digitization of pathology would allow fewer specialized pathologists to serve more patients while increasing diagnostic accuracy and precision. Another important reason for the integration of WSI in pathology is that its necessary to digitize tissue slides before a computer can read those. It is, therefore, an essential part of the automated classification process.

4

Automated classification

One of the questions this thesis tries to answer is whether the status quo of automated histopathological classification is reliable and accurate enough to be used in daily hospital care. To address this question, seven articles were collected. Three of these look at the comparison between deep neural networks and doctors in general medicine and are discussed in section Machine learning in medicine. The other four articles are mainly designed to compare an algorithm to a pathologist. These papers are discussed in section Machine learning in histopathology. All articles concluded that the best tool for classifying images is a convolutional neural network. Hence, the comparisons in this section are between a pathologist and a convolutional neural network. The variety in the papers is in the method and statistics used to measure this comparison. Intuitively, the best way to compare a pathologist to an algorithm is by using a third variable, the so-called ground truth of gold standard. Consequently, the pathologist and the network can be individually examined to this ground truth. However, such a ground truth is not always existing. To compare the cases where such a standard is lacking, various statistical tools were used. For a more detailed description of the statistics used by the discussed articles, we refer to Appendix A.

4.1 Machine learning in medicine

Histopathology is not the only area of research where deep learning and, more precise image analysis, is used to improve diagnostics and increase efficiency. One of the first significant achievements of deep learning in medicine was in the use of an algorithm detecting retinopathy, a signal of the presence of diabetes Gulshan et al. (2016). The test used to diagnose retinopathy is called fluorescein angiography; the process where a fluorescent dye is injected in the bloodstream. The color highlights the blood vessels in the eye so they can be photographed. The results of the algorithm are promising:

	AUC algorithm
Testset 1	0.991 (95% CI, 0.988-0.933)
Testset 2	0.990 (95% CI, 0.986-0.995

Another promising algorithm is an example is of the automated detection of skin lesions Esteva et al. (2017). Skin cancer and in particular melanoma is cancer that

is hard to spot and does not give many infirmities early on. With the use of deep learning applications, people can take a picture of a birthmark and let the diagnostic algorithm run while at home. Such algorithms are cost-efficient and more critical, can improve the detection of early manifestations of melanoma.

	Overall accuracy Dermatologist	Overall accuracy Algorithm
Three-class disease partition	65.78%	72.10%
Nine-class disease partition	54.15%	55.41%

Another prominent area of research is automated diagnostics in radiology. A study was done to compare an algorithm to practicing radiologists Rajpurkar et al. (2018). This study by Pranav Rajpurkar et al. developed a convolutional neural network called CheXneXt that could identify the presence of 14 different pathologies like pneumonia, pulmonary masses, and pleural effusion. The results were encouraging. On 10 out of 14 pathologies, the algorithm performed evenly or better than specialized radiologists.

	Pathologist AUC	Algorithm AUC
Cardiomegaly	0.888	0.831
Emphysema	0.911	0.704
Hiatal Hernia	0.985	0.851
Atelectasis	0.808	0.862
Other	No significant difference	No significant difference

4.2 Machine learning in histopathology

As seen in the preceding section, networks can classify images reasonably accurate in various medical fields. However, the critical question is not whether an algorithm can classify images; it is whether the algorithm classifies images better than a pathologist. We will discuss some studies that made this comparison between algorithms and pathologist to discover whether the current techniques could potentially replace the expertise of a pathologist.

The research of Cruz-Roa et al. (2017) focused on a deep learning approach to identify the extent of an invasive tumor on digitized whole-slide images. The study compared the results to the manually annotated ground truth in the Cancer Genome Atlas. Invasive breast cancers are those that spread from the original site into the surrounding tissue. The origin of this tumor is either the milk ducts or the lobules. These tumors comprise roughly 70% of all breast cancer cases Dillon et al. (2010). An automated and reproducible methodology for the detection of invasive breast cancer on tissue slides could potentially reduce the total amount of time required to diagnose a breast case and lessen the inter-and intra-observer variability Van Baardwijk et al. (2007).
The detection performance of the ConvNet, trained with data from the Hospital of the Unversity of Pennsylvania and University Hospitals Case Medical Center is

measured in terms of mean Dice coefficient, positive predictive value (PPV) and negative predictive value (NPV). They compared their CNN to 195 cases from The Cancer Genome Atlas. The research delivered the following results:

Statistic	Result
Dice	75.86%
PPV	71.62%
NPV	96.77%

The classifier has a high degree of agreement in the prediction of the presence and extent of invasive tumor regions.

A recent study of Wei et al. (2019) did a comparison study between pathologists and a deep neural network. They used a deep convolutional neural network to automatically classify the histologic patterns of lung adenocarcinoma on surgical resection sides. Their model was evaluated on an independent set of 143 whole-slide images. It achieved a Kappa score and an agreement with three pathologists for classifying the predominant patterns that were slightly higher than the inter-pathologist scores.

	Kappa	Agreement
algorithm-pathologist	0.525	66.6%
inter-pathologist	0.485	62.7%

These results show a higher agreement among the used algorithm and a pathologist than between pathologists. The results are positive since the algorithm reduces inter observability.

A study by Bejnordi et al. (2017) compared an algorithm to a pathologist in a simulated clinical setting. Articles like this are essential since all the papers discussed above conclude that clinical research needs to be done before we can determine the functionality of automated classifiers. This paper is the first step in accomplishing this. For this purpose, a dataset was acquired by the Radboud University Medical Center and the University Medical Center Utrecht. This dataset contains 399 slides of sentinel node metastasis. A sentinel node is the first lymph node the breast tissue drains its lymph-fluid in and is an important indicator of metastasis. A retrospective study showed that expert pathologists changed the nodal status in 24% of patients Vestjens et al. (2012). Therefore, it would be helpful if an algorithm could take over this job and thereby reduce this statistic. The common used gold standard for determining whether a sample contains malignant tissue is IHC staining [1].

The panel of pathologists consisted of two groups. One group without time constraints (WOTC) and another group with time constraints (WTC). This is an important distinction to make since a group of pathologists with unlimited time will perform better than a single pathologist bounded to a time limit. In a clinical

[1]Immonohistochemistry (IHC) staining is used to exclude a human bias. IHC is a staining process where specific antigens are targeted in tumor cells. Using this technique, all metastasis is seen. In the article discussed above, no IHC is used by the pathologists. However, using IHC is a normal procedure in the process of classifying images in normal care Veta et al. (2015).

setting, however, no time constraint would be infeasible because of the massive amounts of tissue that needs to be diagnosed. To assess the slides more realistically, the pathologists got 2 hours to classify 129 images, the number of images the test set contains. The other group of pathologists was not time-constrained. We show the results of the best performing algorithm on WSI classification from Wang et al. (2016), which used a GoogLeNet architecture Szegedy et al. (2015).

	Algorithm	Pathologist WTC	Pathologist WOTC
AUC	0.994	0.810	0.966
Specificity		98.5%	98.7%
Sensitivity		62.8%	93.8%

Measured by the AUC, the algorithm outperforms the pathologist WTC and is comparable with a pathologist WOTC. An important thing to notice is the difference in sentivity of a pathologist WTC and WOTC. This sensitivity means a lot of false negatives judgments were made. It is precisely because of such statistics, that there is theoretical huge area of improvement by using automated classification tools.

In the table below, a short overview is presented of the research discussed above.

	Benjordi et al.	Wei et al.	Cruz-Roa et al.
Test used	AUC	Kappa	Dice
Result pathologist	0.810	0.485	
Result algorithm	0.994	0.525	75.86%

5

Current issues and future solutions

5.1 Black Box

One of the most substantial obstacles in deep learning networks is the problem of the black box. The black box problem in deep neural networks is the inscrutability of its functioning by humans. The lack of understanding does not lie in the architecture of those networks; we do understand that. The difficulty is that we do not understand how a deep neural network generates its output. What is a network 'thinking' when classifying an input image as a tumor? As already briefly addressed in the introduction; this is not always a problem. Nevertheless, in medical care, where the lives of patients depend on the expertise of a doctor, it unquestionably is. A neural network is intended to detect regularities in data, not the underlying causal relationship. This problem seems insurmountable. However, London (2019) argues the following: 'The knowledge of underlying causal systems is in its infancy; the pathophysiology of the disease is often uncertain, and the mechanisms through which interventions work is either not known or not well understood. Therefore, decisions that are atheoretic and opaque are commonplace in medicine. Modern clinicians prescribed aspirin as an analgesic for nearly a century without understanding the mechanism through which it works. Lithium has been used as a mood stabilizer for half a century, yet why it works remains uncertain. Large parts of medical practice frequently reflect a mixture of empirical findings and inherited clinical culture. In these cases, even efficacious recommendations of experts can be atheoretic in this sense: they reflect experience of benefit without enough knowledge of the underlying causal system to explain how the benefits are brought about.'

5.2 Implementation

'While digital pathology has substantial implications for telepathology[1], second opinions, and education, there are also significant research opportunities in image computing with this new source of "big data" Madabhushi & Lee (2016)'. As discussed in the theory of image classification, one of the reasons why these networks work so good is the presence of large amounts of data. If all pathology centers shift to a digital workplace, much more data becomes available to train and optimize the models. This will, in turn, further reduce the variability and hence increase accu-

[1]The practice of pathology at a distance

racy. Nonetheless, their sill remains several important technical, and computational challenges that need to be overcome before computer-assisted image analysis of digital pathology can become a part of the routine clinical diagnostic workflow. One of the central problems in the computational interpretation of digital slide images has to do with color variations in the tissue induced by differences in slide preparation, staining, and even whole-slide scanners. It is, however, possible that with more data, resulting in a better-trained model, this problem will be resolved.

5.2.1 Cost of automated classifying

Another reason for the absence of the application of these algorithms is the price tag that comes along with it. To use predictive algorithms in daily routine, the whole pathology department needs to be digitalized. Each tissue slide needs to digitized through whole-slide image scanners, and these are not cheap. Since not much clinical research has been done, it is hard to calculate whether the investment in these expensive machines will pay off by the effectiveness of automated classifying.

6

Conclusion

6.1 Research Aims

This thesis aimed to explore and give an overview of the status quo of histopathological analysis. This field is still in its infancy, and a lot more research needs to be done. However, the earliest studies done in this field show promising results.
Deep convolutional neural networks can classify tissue fairly accurately and even outperform human pathologists in some areas. These encouraging results are not integrated into hospitals yet, mainly because of the problems discussed. However, at this very moment, the University Medical Center Utrecht is integrating the algorithm discussed in the paper of Bejnordi et al. (2017) in its pathology department. The first moves on the path to revolutionizing the field of pathology, are now taken.

6.2 Future Research

The most import future research needs to be done in the implications of letting a computer diagnose patients. This is more of an ethical discussion, but people do not feel comfortable with the idea of letting a computer decide about their lives. These kinds of revolutions require some time to adapt to the concept. The feeling when boarding on the first airplanes must have been the same. With the revolutionization that is ongoing in this field, there is a probability that the role of the pathologist nowadays will differ from the one in the future. Moreover, future research must be conducted in a clinical setting. Nearly all papers released till now are theoretical, and all argue that clinical research needs to take place before we can tell whether the role of automated classiffcation can be a part of medical care. One of the most important things for future research is that the computer vision specialists will need to work intimately with pathologists to construct new and innovative solutions to the decisive image analysis challenges in digital pathology.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H.-T. (2012). *Learning from data* (Vol. 4). AMLBook New York, NY, USA:.

Afshine Amidi, S. A. (2018). *Cnn.* `https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks`. (Accessed: 2019-06-30)

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... others (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, *318*(22), 2199–2210.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N., ... Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, *7*, 46450.

Dillon, D., Guidi, A., & Schnitt, S. (2010). Pathology of invasive breast cancer. *Diseases of the Breast. 4th ed. Philadelphia, Pa: Lippincott-Williams & Wilkins*, 374–407.

Edge, S. B., Byrd, D. R., Compton, C., Fritz, A., Greene, F., & Trotti, A. (2010). *American joint committee on cancer cancer staging manual.* New York: Springer.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, *25*(1), 24.

Géron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.".

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. (`http://www.deeplearningbook.org`)

Griffin, J., & Treanor, D. (2017). Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology*, *70*(1), 134–145.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... others (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, *316*(22), 2402–2410.

Gurcan, M. N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, *2*, 147.

Jackson, S. L., Frederick, P. D., Pepe, M. S., Nelson, H. D., Weaver, D. L., Allison, K. H., ... others (2017). Diagnostic reproducibility: what happens when the same pathologist interprets the same breast biopsy specimen at two points in time? *Annals of surgical oncology*, *24*(5), 1234–1241.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21.

Madabhushi, A., & Lee, G. (2016). *Image analysis and machine learning in digital pathology: Challenges and opportunities.* Elsevier.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

NCI. (2013). *Tumor grading.* `https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet#r1`. (Accessed: 2019-06-30)

Patterson, J., & Gibson, A. (2017). *Deep learning: A practitioner's approach.* " O'Reilly Media, Inc.".

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... others (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, *15*(11), e1002686.

Su, H. K., Wenig, B. M., Haser, G. C., Rowe, M. E., Asa, S. L., Baloch, Z., ... others (2016). Inter-observer variation in the pathologic identification of minimal extrathyroidal extension in papillary thyroid carcinoma. *Thyroid*, *26*(4), 512–517.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1701–1708).

Van Baardwijk, A., Bosmans, G., Boersma, L., Buijsen, J., Wanders, S., Hochstenbag, M., ... others (2007). Pet-ct–based auto-contouring in non–small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. *International Journal of Radiation Oncology* Biology* Physics*, *68*(3), 771–778.

Vestjens, J., Pepels, M., de Boer, M., Borm, G. F., van Deurzen, C. H., van Diest, P. J., ... others (2012). Relevant impact of central pathology review on nodal classification in individual breast cancer patients. *Annals of oncology*, *23*(10), 2561–2566.

Veta, M., Van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., ... others (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis*, *20*(1), 237–248.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Wei, J. W., Tafe, L. J., Linnik, Y. A., Vaickus, L. J., Tomita, N., & Hassanpour, S. (2019). Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific reports*, *9*(1), 3358.

Appendix A

Statistics

A.1 Cohen's Kappa

Cohen's Kappa is a measurement for inter-and intra-observer agreement. It measures the percentage to which degree two independent observers agree on a classification output. Applied to the topic of this thesis, observer one is the pathologist and observer two is the algorithm. Cohen's Kappa measures the chance of random agreement (AC), subtracts this chance from the observed agreement (OA) and normalizes the value. A Kappa value k of 1 resembles a perfect agreement between observer one and two whereas a Kappa value k of 0 resembles a total disagreement between the observers.

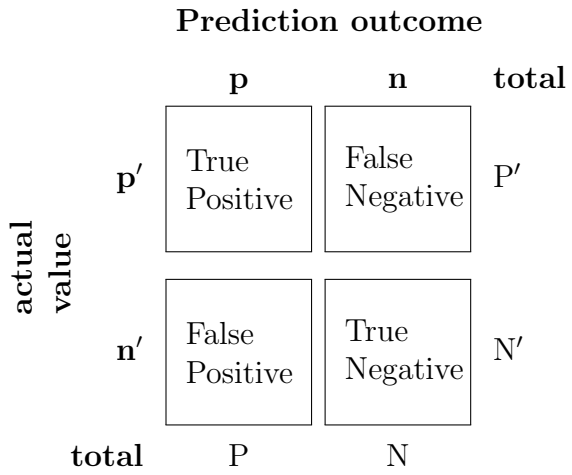
$$k = \frac{OA - AC}{1 - AC} \tag{A.1}$$

A.2 Confusion Matrix

The result of the in-and output of a simple classification algorithm can fall into the following four categories: false positive, true positive, false negative, and true negative. The result is false positive when the test, in this case, the prediction of the algorithm, says it is malignant while it is not. True positive is when the test classifies the sample as malignant, and it is. False-negative is when the test classifies as benign, but in fact, it is malign, this is the most harmful and should be avoided. Sensitivity is the number classified as true positives, and specificity is the number classified as true negative.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{false negatives} + \text{true positives}}$$

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$



A.3 Receiver Operating Characteristic

ROC curves provide a simple way to give all of the information. i.e., this curve shows the specificity and sensitivity for all threshold values. The Y-axis shows the true-positive rate (sensitivity). The X-axis shows the false-positive rate, which is the same as (1 - specificity); these are the values that are benign but are classified as malign. The ROC graph summarises all of the confusion matrices that each threshold produced.

A.4 Area Under the Curve

The Area Under the Curve (AUC) is the area under the ROC-curve. It represents the accuracy of a diagnostic test and the scores can be interpreted in the following fashion:

.90-1 = excellent
.80-.90 = good
.70-.80 = fair
.60-.70 = poor
.50-.60 = fail

A.5 Dice similarity coefficient

The Dice similarity coefficient is used to quantify the performance of image segmentation. It uses a ground truth and compares this to an automated image segmentation tool. The Dice score measures the similarity between the objects. The Dice score can be described in terms of accuracy as described in the section about the confusion matrix:

$$D = \frac{2TP}{2TP + FP + FN} \tag{A.2}$$

The Dice score is not only a measure of how many positives you find, but it also penalizes for the false positives that the method finds. The Dice score is also

penalizing for the positives that the algorithm could not detect. When applied to detecting tumor extent, we have two 'masks.' Mask X is the ground truth that is labeled by an experienced pathologist. Mask Y is the labeling of the convolutional network. A pixel is denoted as one if it contains malignant tissue and as 0 if it does not. The number of positives is the total number of pixels that are labeled one by mask X. The number of true positives is the number of pixels that have value 1 in both X and Y. The number of false positives is the number of pixels labeled malign by Y but benign in X. The number of false negatives is classified as 1 in mask X but as 0 in Y.

Appendix B

Backpropagation

For simplification matters, the bias b_i^k for node i in layer k is added to the weights vector as w_{0i}^k with a fixed output $o_0^{k-1} = 1$ for node 0 in layer $k-1$. Note that

$$a_i^k = b_i^k + \sum_{j=1}^{r_{k-1}} w_{ji}^k o_j^{k-1} = \sum_{j=0}^{r_{k-1}} w_{ji}^k o_j^{k-1} \tag{B.1}$$

Backpropagation tries to minimize the cost function with respect to the weights by calculating $\frac{\partial C}{\partial w_{ij}^k}$ for each w_{ij}^k:

$$\frac{\partial C(X,\theta)}{\partial w_{ij}^k} = \frac{1}{N}\sum_{d=1}^{N} \frac{\partial}{\partial w_{ij}^k}(\frac{1}{2}(\hat{y}_d - y_d)^2) = \frac{1}{N}\sum_{d=1}^{N} \frac{\partial C_d}{\partial w_{ij}}^k \tag{B.2}$$

Furthermore, we need a cost function , C(X, θ), which defines the error between the desired output $\vec{\mathbf{y}}_i$ and the calculated output $\hat{\vec{\mathbf{y}}}_i$.

$$C(X,\theta) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{B.3}$$

To calculate the cost with respect to the weights, the chain rule is applied:

$$\frac{\partial C}{\partial w_{ij}^k} = \frac{\partial C}{\partial a_j^k}\frac{\partial a_j^k}{\partial w_{ij}^k} \tag{B.4}$$

The first term in the equation is often called the error and is denoted as:

$$\delta_j^k \equiv \frac{\partial C}{\partial a_j^k} \tag{B.5}$$

The second term in the equation can be calculated from the equation of a_j^k:

$$\frac{\partial a_j^k}{\partial w_{ij}^k} = \frac{\partial}{\partial w_{ij}^k}(\sum_{l=0}^{r_{k-1}} w_{lj}^k o_l^{k-1}) = o_i^{k-1} \tag{B.6}$$

Therefore, the partial derivative of the cost function w.r.t. a weight w_{ij}^k is:

$$\frac{\partial C}{\partial w_{ij}^k} = \delta_j^k o_i^{k-1}$$

Once this gradient is calculated, the weights can be updated using the following rule:

$$w_{ij+1} = w_{ij} + \alpha \frac{\partial C}{\partial w_{ij}} \tag{B.7}$$

If the error goes down

$$\frac{\partial C}{\partial w_{ij}} < 0$$

increase the weight.

If the error goes up

$$\frac{\partial C}{\partial w_{ij}} > 0$$

decrease the weight.

The gradient vector C represents the direction of the step that is taken to minimize the cost function.

$$-\nabla C = \begin{bmatrix} \frac{\partial C}{\partial w^{(1)}} \\ \frac{\partial C}{\partial b^{(1)}} \\ \vdots \\ \frac{\partial C}{\partial w^{(L)}} \\ \frac{\partial C}{\partial b^{(L)}} \end{bmatrix} \tag{B.8}$$