Liisa Petäinen

The potential of convolutional neural network in the evaluation of tumor-stroma ratio from colorectal cancer histopathological images

Master's Thesis in Information Technology

April 26, 2022

University of Jyväskylä

Faculty of Information Technology

Author: Liisa Petäinen

Contact information: petainenliisa@gmail.com

Supervisors: Sami Äyrämö, Ilkka Pölönen, and Pekka Ruusuvuori

Title: The potential of convolutional neural network in the evaluation of tumor-stroma ratio from colorectal cancer histopathological images

Työn nimi: Kasvain-strooma suhdeluvun arviointi suolistosyövän histopatologisista kuvista konvoluutioneuroverkkoja käyttämällä

Project: Master's Thesis

Study line: Mathematical computer science

Page count: 49+4

Abstract: In this Master's Thesis, the ability of convolutional neural networks in the evaluation of tumor-stroma ratio from histopathological images, is studied. The goal is to find out, whether pre-training with domain-specific data brings more accuracy to the convolutional neural network model. Tumor-stroma ratio is predicted with the trained model and the predicted values are compared with visual tumor-stroma estimations made by pathologist. When domain-specific data was used in the pre-training of the convolutional neural network, a slight improvement in the validation accuracy of the model was observed. Correlation between the predicted and visual values was also found. Further analysis is needed to study what is the connection of these computationally predicted values to other clinicopathological factors and overall survival of the patient.

Keywords: Convolutional neural networks, computer vision, artificial intelligence, medical image analysis, digital pathology, colorectal cancer, histopathology, tumor-stroma ratio

Suomenkielinen tiivistelmä: Tässä Pro gradu-työssä tutkitaan konvoluutioneuroverkkojen käyttömahdollisuuksia histopatologisista kuvista tehtävässä kasvain-strooma suhdeluvun arvioinnissa. Tarkoituksena on selvittää, mikä on siirto-opettamisen vaikutus, kun opettamisessa käytetään kohdealuespesifistä dataa. Mallin ennustamaa kasvain-strooma suhdelukua

verrataan patologin visuaalisesti tekemään arvioon.

Tutkimuksesta selvisi, että kohdealuespesifisen datan käyttö esiopetuksessa lisää konvoluutioneuroverkkomallin tarkkuutta. Myös korrelaatiota ennustetun ja visuaalisen arvion välillä oli havaittavissa. Tulevaisuudessa olisi hyvä tutkia kasvain-strooma-suhdeluvun yhteyttä muihin kliinispatologisiin tekijöihin ja potilaan elinaikaan.

Avainsanat: Konvoluutioneuroverkot, konenäkö, tekoäly, lääketieteellisten kuvien analysointi, digitaalinen patologia, suolisyöpä, histopatologia, kasvain-strooma suhdeluku

Preface

First of all, this process has been a fun one. Full of interesting, problematic questions to solve after the other. Figuring out some curly problems is in a weird way such annoyingly satisfying.

There are bunch of people to thank for this process. Thanks to my supervisors Sami Äyrämö, Ilkka Pölönen and Pekka Ruusuvuori for giving me guidance in both scientific and technical questions. Huge thanks to pathologists Teijo Kuopio and Juha Väyrynen, without your effort, knowledge and practical data issues this master's thesis would not have been possible. I would also like to thank professor Jukka-Pekka Mecklin and Central Finland Hospital District for the extensive, high-quality data which made this thesis interesting and special to me.

Thanks to my husband Petteri, daughter Saana and son Vili. They have been so patient throughout this process, while wife and mom has spent a lot of time on the computer being enthusiastic about some coding, microscopic images, Python, numbers, nets etc., which might not seem so interesting from their perspective.

Jyväskylä, April 26, 2022

Liisa Petäinen

Glossary

CNN	Convolutional neural network.
CRC	Colorectal cancer.
Histopathology	Field of pathology, in which the samples from living tissue are processed in a pathology laboratory into thin slices which can be visualised on a microscope.
H&E	Haematoxylin & Eosin, a common staining chemical in histopathology, which makes the tissue samples visible in different shades of pink and purple.
Stroma	Connective tissue. In this case, the stroma indicates the connective structures within the tumor sites.
TSR	Tumor-stroma ratio, proportion of stroma within the tumor site.
WSI	Whole-Slide Image, a digitized image from a microscopic tissue sample.

List of Figures

List of Tables

Contents

1 Introduction

Within the last five years, the ability of convolutional neural networks (CNNs) to learn features have been noted also in the field of pathology. CNNs have become the most widely used technique when developing algorithms for pathology tasks (Litjens 2017). Automating some of the tasks pathologists work with on a daily basis, would bring huge relief to the rising workload pathologists struggle with. This is particularly important, as the amount of pathologists is decreasing all the time.

This field of studying and analysing digitized histopathological data is called digital pathology. It is one of the many application fields of computer aided medical image analysis. Today, digital pathology is not only about digitizing tissue sections as it was before, it is more about the effort to automate at least some of the pathologists' routine tasks. The challenge lies in the specificity of the problems: one experienced pathologist can solve multiple different tasks from multiple different tissue and cancer types, but one computer vision algorithm does not have the ability to perform the way human brain does.

Another challenging feature is the lack of annotated data. CNNs require huge amounts of data in order to learn the most essential features in an image. Even though the volume of medical data is huge these days, due to the privacy issues, it is not easily available. This leads to the situation, where data is not widely shared between facilities, it is rather utilized on everyone's own needs.

Data for this study comes from colorectal cancer (CRC), which is the second most most death causing cancer in the world, causing over 900,000 deaths every year (World Health Organization 2020). When pathologist takes a look at a histopathological sample, there are multiple morphological features to observe when making diagnosis. One of these features is the amount of stroma within the tumor site, i.e. tumor-stroma ratio (TSR), which has been shown to correlate with survival of the patient in many solid cancer types (Huijbers 2013; Ma 2012; Mesker 2007).

Stroma is a connective tissue which does not have any special role or functions within normal tissue. In the case of stroma within the tumor, debates are ongoing whether the role of stroma

is to help the tumor to proliferate or to suppress the tumor growth (Colangelo et al. 2017). Nevertheless, for some reason, the prognosis seems to be poorer, if TSR is over 50 %.

There are two main goals in this Master's Thesis. The first goal is to find out, whether pre-training the CNN with domain-specific data would bring more accuracy to the model, when classifying different histopathological structures. The second main goal is to find out, whether TSR can be reliably defined with the help of CNNs. The predicted TSR values are compared with the TSR values estimated visually by a pathologist.

The structure of this study goes as follows: first a literature review is presented in chapters 1-4. After that, data and methods are described in chapter 5. The results of this study are presented in chapter 6, after which discussion is followed in chapter 7. Bibliography and appendices are listed at the end of this study.

2 Convolutional neural networks

CNNs have become the state-of-art method in computer vision tasks within the last 10-20 years. Origin of neural networks dates back to 1962 when David Hubel and Torsten Wiesel studied a cat's visual cortex. They classified cells responsible for pattern recognition into two groups: "simple cells" and "complex cells" (Hubel and Wiesel 1962). The first corresponding computer model was presented twenty years later, when Kunihiko Fukushima published a self-organizing network. This network had layers that were connected with mathematical operations between the cells of different layers (Fukushima and Miyake 1982). These were the first steps towards modern visual pattern recognition techniques.

Few years later, in 1989 LeCun et al. (1989) trained CNN that predicted handwritten digits, the model was based on Fukushima's "neocognitron" architecture. The final breakthrough of CNNs happened in 2012 with ImageNet image classification challenge where Krizhevsky, Sutskever, and Hinton (2012) presented the best performing model called AlexNet. Since then, ImageNet challenge has been running on a yearly basis and has been won by CNN architectures like 19-layer VGG19 and 22-layer GoogLeNet (Russakovsky 2015; Simonyan and Zisserman 2014; Szegedy et al. 2015).

2.1 Typical structure

CNN is constructed from different types of layers, which typically are convolutional layers, pooling layers and fully connected layers (Figure 1). The main task of a convolutional layer is to produce feature maps by applying convolutional operations to the input. Pooling layers can be added to "squeeze" the amount of information, which is passed on to the next layer. Fully connected layers are the ones solving the final classification problem with the data they have from the previous layer (Goodfellow, Bengio, and Courville 2016).

When CNNs are applied to color image data, each layer is a grid-like, three dimensional structure, which has height, width and depth. For example, if the input is 32 x 32 RGB-image, height and width of the first layer are both 32, depth is 3 which is the number of color channels in the input image (Aggarwal et al. 2018; Goodfellow, Bengio, and Courville

2016).

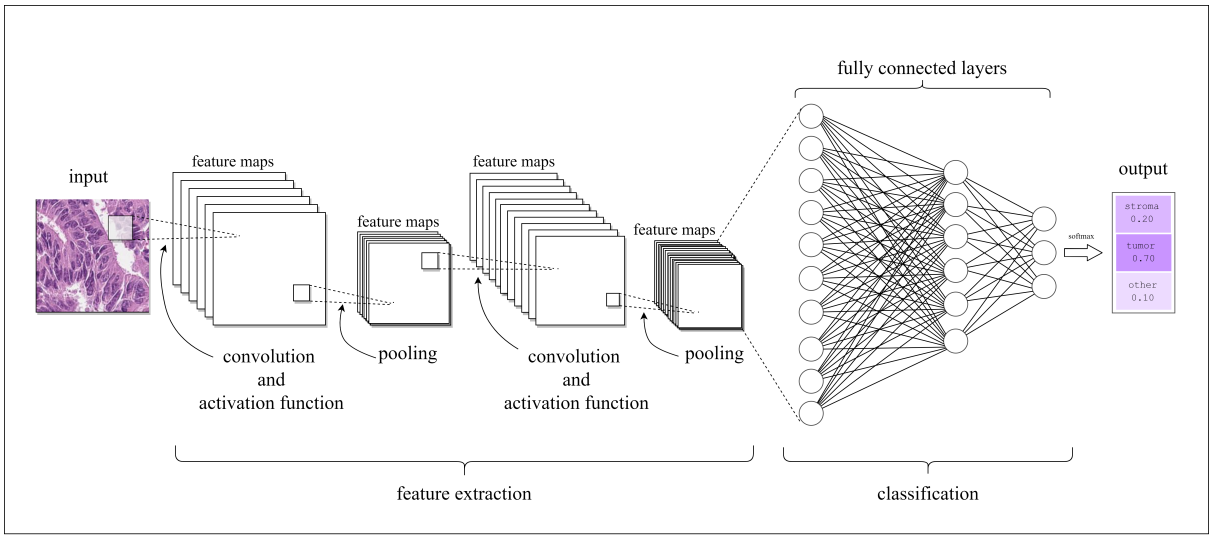

Figure 1: Structure of a CNN. In this example, every convolution layer is followed by a pooling layer. Final classification is performed after fully connected layers with softmax function.

To extract the essential features, the spatial relationships between pixels are passed to the next layer with convolution operations. Kernel, or a convolutional filter, goes through each channel in the each input image and produces feature maps for the next layer. The kernel can be e.g. size 1 x 1, 3 x 3 or 5 x 5, usually an odd number is preferred. For a two-dimensional image, convolution S is a sum of feature maps from all convolution operations of image I and kernel K:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{2.1}$$

Where (i,j) is the size of the input image I and (m,n) is the size of the kernel K. Activation function is applied to each convolutional feature map. The purpose of activation function is to add the non-linearity to the CNN, as convolution is a linear operation. As an activation function, CNN models often use Rectified Linear Unit (ReLU) (Nair and Hinton 2010):

$$f(x) = max(0,x) \tag{2.2}$$

Also other types of ReLU-functions have been developed since the introduction of the orig-

inal ReLU in equation 2.2, such as Leaky ReLU (Maas, Hannun, Ng, et al. 2013). In Leaky ReLU, x is multiplied with a constant a (equation 2.3). In the original presentation of Leaky ReLU, it was suggested that parameter $a = 0.01$. Increasing the value of parameter a was shown to have smaller test error compared to ReLU or LeakyReLU with $a = 0.01$ (Xu et al. 2015).

$$f(x) = max(ax, x) \tag{2.3}$$

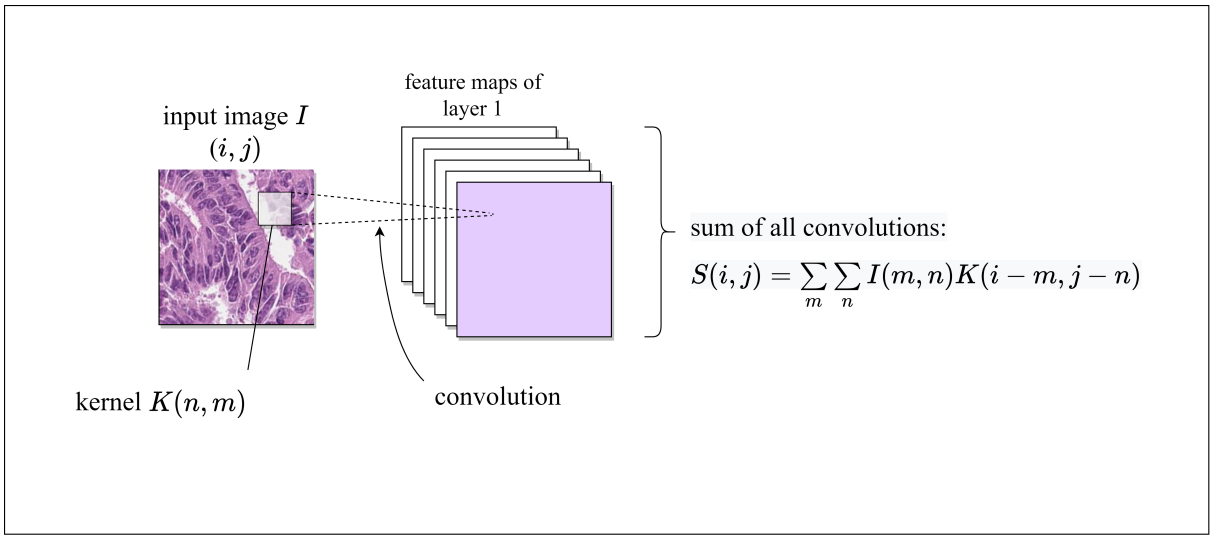

Figure 2: Generation of feature maps from layer 1. Input image I goes through convolution operations throughout the image with kernel K. After the convolution, an activation function (e.g. ReLU) is applied to each feature map.

Usually, CNNs have one or multiple pooling layers. As mentioned earlier, the goal of the pooling layer is to "squeeze" the information, and at the same time the resolution of the feature maps decreases. Typical types of pooling are max-pooling, average-pooling and stochastic pooling, max-pooling being the most common one (Zeiler and Fergus 2013; Kevin, Hayit, and Dinggang 2017). In max-pooling, a small rectangular region (e.g. 3 x 3 pixel^2) of the feature map is replaced by the maximum value of the region. Pooling layer decreases the number of parameters needed in the next level, thus decreasing the computational cost of the network as well (Kevin, Hayit, and Dinggang 2017). Stride is the number of pixels the filter moves at a time over the input matrix. In the case $n = stride$, the size of the activation for the next layer is reduced by n^2.

2.2 Optimization of the loss function

As the goal of the convolutional operation is to minimize the loss function, training a convolutional neural network is basically a global optimization problem. In classification tasks, the loss function is either softmax or the sigmoid cross-entropy function. Minimizing the loss function is done using backpropagation algorithm, which utilizes the information of previous weights when calculating the gradient of the loss function (Werbos 1990). Figure 3 sums up the typical learning process of a CNN.

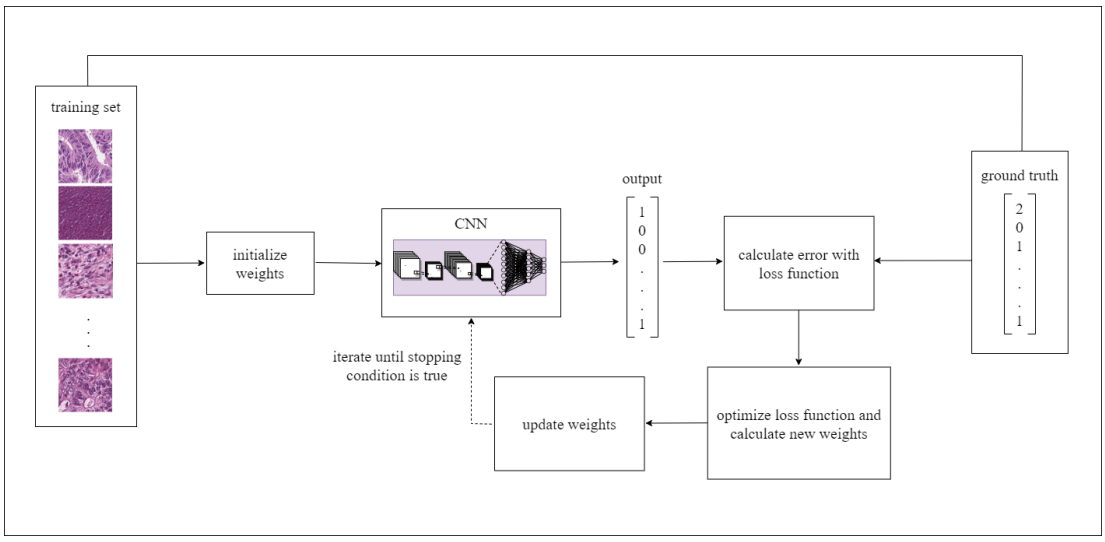

Figure 3: Typical learning algorithm of a CNN.

Softmax function for classification probabilities is the most common one when dealing with multiclass-problems. Softmax function turns the feature maps activations of the last fully connected layer into probabilities:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_i}} \tag{2.4}$$

Where z_i is i^{th} activation of a feature map, N is the number of classes. Cross-entropy loss uses the softmax probabilities and is calculated as follows:

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i) \tag{2.5}$$

Where t_i is the true label and p_i is the softmax probability of the i^{th} class.

2.3 Methods to improve accuracy of the convolutional neural network

CNNs can be constructed in countless numbers of ways. In order to improve accuracy some additional layers or structures, as well as different setups and hyperparameter tuning can be applied. One example of the improving accuracy is adding dropout-layers, which was originally introduced by Hinton et al. (2012). Dropout means randomly removing some of the connections of the last fully connected layer. This has been shown to increase accuracy, because dropout reduces overfitting (Srivastava et al. 2014). Overfitting is a common problem with neural networks, as networks learn such precise connections in the training phase which cannot be generalized to the data at hand. In addition to the original dropout, also other improved methods for dropout has been developed. One of these is DropConnect (Figure 4), in which some of the connections between, not only the last fully connected layer, but also from other fully connected layers, are dropped (Wan et al. 2013).

There are also other options to improve accuracy. These include methods like parameter tuning by tuning, e.g., the learning rate, as well as testing different optimization algorithms and loss functions. Also data augmentation is one option, which is rather easy to implement with computer vision libraries such as PyTorch. Data augmentation means increasing the number of input images by rotating, flipping or rescaling the original image. In addition, transfer learning is widely applied method when training the model and one method to consider especially when there is a minor or major lack in the amount of data at hand (Mikołajczyk and Grochowski 2018).

2.3.1 Transfer learning

Probably the most effective method to increase accuracy is transfer learning. Transfer learning means training a neural network with a pre-trained neural network that has already learned to recognise patterns. There are roughly two types of transfer learning methods: fine-tuning and feature extracting (Litjens 2017). In the case of medical images, fine-tuning means medical images are used only to fine-tune the hyperparameters of the final classifier

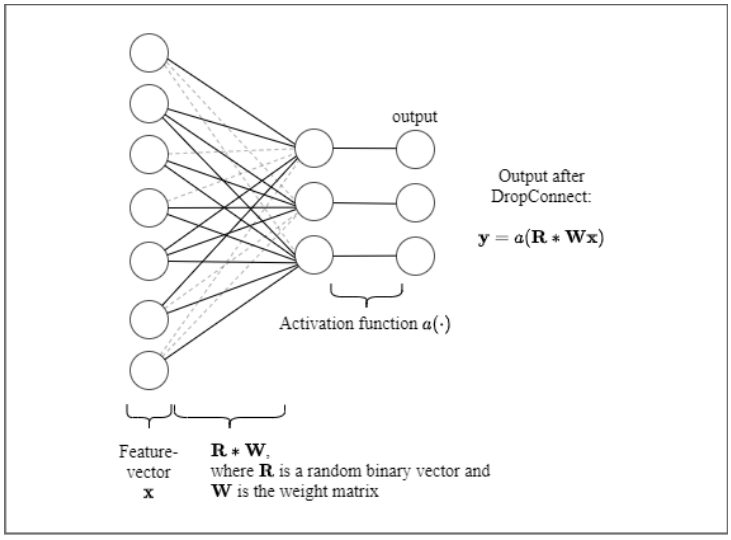

Figure 4: DropConnect drops connections between two fully connected layers (dashed lines). Dropped connections are defined by randomized binary vector **R**.

of the pre-trained network. When using pre-trained CNN as a feature extractor, a classifier is trained with medical image data, but the features origin is the pre-trained network. In practice, this means initializing the network with weights from the pre-trained network. (Goodfellow, Bengio, and Courville 2016)

Teaching a CNN requires a huge amount of training data, which is the problem when the origin of data is in the medical field. Even though ImageNet has images of dogs, planes and houses which are far from histopathological ones, initializing weights from a CNN trained with ImageNet increases accuracy and is often applied in the case of medical images (Litjens 2017; Mormont, Geurts, and Marée 2018).

Lack of data is a common problem also in digital pathology. Whole Slide Images (WSI) are large in size and when tiled to smaller image tiles, one WSI could easily produce 10,000 - 20,000 image tiles. But the problem is the lack of samples. One sample, one WSI, is from a tissue sample from one patient. To train a model, which generalizes to other images tiles from the same domain, image tiles should be derived from different samples. Training data is tiled from annotations made by pathologists, and annotation sites from one sample do not produce massive number of image tiles. This is why pre-training with ImageNet or other natural images dataset is a common approach also when training a CNN with histopathological images (Bayramoglu and Heikkilä 2016; Kieffer et al. 2017).

In the histopathology field, pre-training a neural network model with data from the same domain has been shown to increase accuracy of the final model. Initializing weights from the same domain, in this case histopathology, brings more accuracy when comparing to mere ImageNet-initialization (Mormont, Geurts, and Marée 2018; Sharmay et al. 2021).

Considering the wide heterogeneous range of medical images, transfer learning from any other medical domain to another medical domain might not always work. Study by Menegola et al. (2016) was a good reminder of this. They used images from diabetic retinopathy to pre-train a network to classify images from melanoma to malignant and benign. In their experiment, pre-training with ImageNet brought higher accuracy than pre-training with images from diabetic retinopathy. They did mention that the differences between the diagnosis processces of these two diseases might explain these results.

3 Digital pathology

Acquiring, managing and analyzing information from histopathological samples in a digital environment is called digital pathology. It started developing in the 1980s enabling pathologists to use remote consultation, which is called telepathology. In addition to the ability to share histopathological images through the internet with other pathologists, digital pathology brought many advantages to the field. When digitized, also the quality of tissue samples remain constant over time and they are easy to access later if needed (Al–Janabi, Huisman, and Van Diest 2012). At present, digitizing tissue slides is a routine procedure in most pathology laboratories (Pallua 2020). The digitized file form of the histopathology sample is the multilayered WSI, which is described in more detail in chapter 3.1.

When digital pathology had started to develop, the main focus was on the digital information it holds or how the histopathological information is stored. Nowadays, digital pathology is mainly concentrating on new ways of utilizing the growing digital information. The most popular technique to achieve this is through deep learning (Litjens 2017). This would not be possible without the recent improvements of computing capacities as WSIs are huge in size, even though the common practice is to tile the image to smaller patches.

3.1 Whole Slide Image

Before a living tissue is a static sample and can be visualized on a computer screen, it has gone through many different processing steps. At first, the tissue sample is processed in a histopathology laboratory into thin sections, which are placed on small microscope glasses, stained and then scanned with a special WSI-scanner. Examples of the WSIs from CRC sections are presented in figure 5. The first commercial WSI-scanner was introduced back in 1994 and today, there are plenty of different scanner manufacturers. Modern WSI-scanners have practical features such as robotics and barcode readers, which ease the slide scanning process (Pantanowitz et al. 2018).

When digitizing the microscopic glass slide, the main goal is to maintain the accuracy of the specimen, so that the diagnostic performance remains even if the specimen is viewed on a

computer screen instead of on a microscope. Several studies have shown the diagnoses made from conventional microscope glass are in good concordance with the diagnoses made from WSIs (Azam et al. 2021; Farahani, Parwani, and Pantanowitz 2015). Despite the numerous validation studies and the fact, that the technology for Whole Slide Imaging developed in the 1990s, it was not until 2017, when WSIs were approved for primary diagnosis by the US Food and Drug Administration (Pantanowitz et al. 2018). Within European Union, WSIs are also allowed for in vitro diagnostics (Azam et al. 2021).

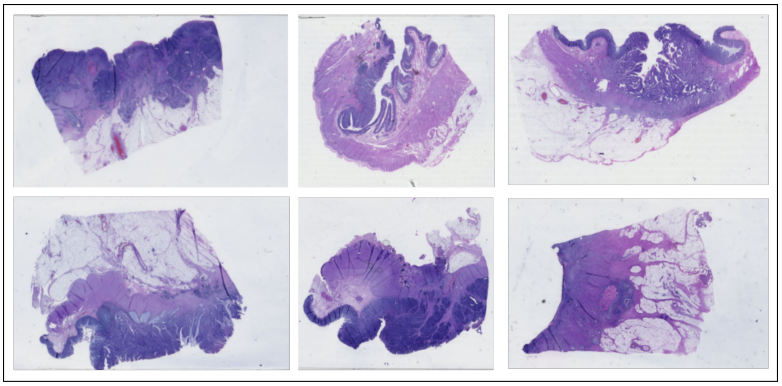

Figure 5: Example WSIs from CRC sections. Images origin: Central Finland Health Care District, "Suolisyöpä Keski-Suomessa 2000-2015" -project.

Depending on the scanner, image resolution is in the range of 0.25 - 0.50 μm per pixel. To give some perspective on how detailed the WSI is, the typical diameter of an animal cell is about 10 µm. From these zoomable WSIs, pathologists are able to navigate the image as with a conventional light microscope. To make the microscope-like zoomability possible, WSIs have a pyramidical structure. Biggest resolution level could have, e.g., 100,000 x 200,000 px^2 (Figure 6). Total number of pixels in one WSI can be trillions, which makes the file size of one WSI large, ranging from 1 to 4 gigabytes. This leads to the problem of data storage and handling, which is one of the major challenges in pathology institutes when digitizing samples (Pantanowitz et al. 2018).

In addition to the storage challenge, one challenging feature from developers' perspective, is the unstandardized file format of the WSIs. The main formats in use are TIFF, BigTIFF (TIFF variant, which supports larger file size), MRXS (compatible with Mirax-microscope's digital scanners) and JPEG2000. Image formats vary in size and can be hard to share as they require specific software (Farahani, Parwani, and Pantanowitz 2015). Standardizing the

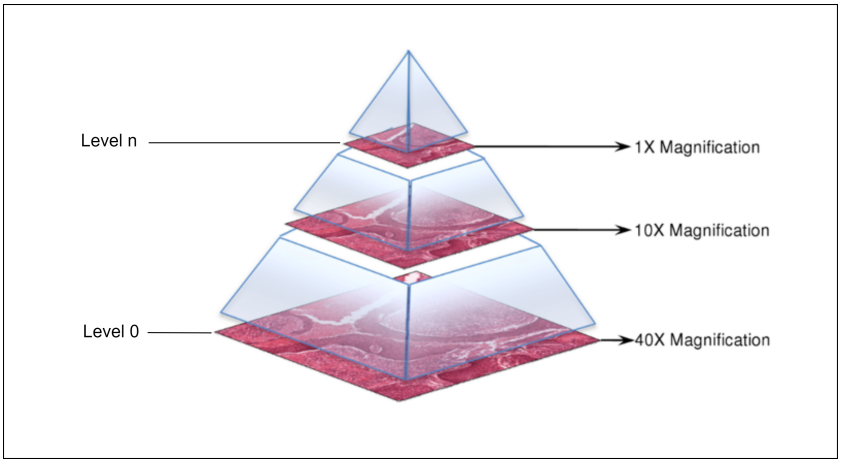

Figure 6: Pyramidal structure of a WSI. The highest resolution on level 0 can consist of billions of pixels. WSI can be comprised from n number of resolution levels (e.g. 7-9) to make the zoomability possible without losing accuracy (original image from CAMELYON-17 challenge, edit Liisa Petäinen).

WSI file format would help not only pathologists, but also professionals who are developing digital pathology applications and software. Some ease was brought few years back when Goode et al. (2013) introduced open-access OpenSlide-library. OpenSlide is written in C, but it enables reading and manipulating WSIs with applications written also with C++, Java and Python. OpenSlide-library has been very beneficial tool also in the development of bioimage analysis software, e.g., QuPath (Bankhead et al. 2017).

3.1.1 Preprocessing for deep learning

When training a CNN with WSIs, there is some preprocessing to be made. As WSIs are huge in size, they are tiled into smaller squares, e.g., 32 x 32 px^2 or 224 x 224 px^2, which makes them easier to deal with. If model development is based on existing CNN architecture, one needs to check which input-size is used in the current CNN architecture. Input-sizes of CNNs varies a lot depending on the architecture used.

In order to properly visualize the tissue under the microscope, it is stained. Haematoxylin & Eosin (H&E) stain is the most common method used. This makes structures visible in different shades of purple and pink. H&E stain makes the cell nuclei appear darker purple, as other structures appear in different, lighter shades of purple and pink. Histopathological samples are very vulnerable to inter and intra laboratory stain variations (Roy et al. 2018).

These variations are due to, e.g., differences in WSI scanners, H&E staining procedures and the manual tissue processing nature. Pathologists can handle these color variations with ease, but for CNN-models, the task is more complicated.

Color normalization is one solution to make computers concentrate on structures, even if the images are from different institutes or different laboratories. Color can be normalized in many different methods, some of them are unsupervised and computationally cheaper methods, some supervised, but most of them need a target image. Most common methods are Reinhard-method, Macenko-method and Vahadane-method (Reinhard et al. 2001; Macenko et al. 2009; Vahadane et al. 2016). Supervised methods are computationally complex and slow, which sometimes leaves them out, when considering which method to use.

Also one thing to keep in mind before the tiling phase is the zoom-level. WSIs comprise from different resolution levels. If CNN is pre-trained with some other histopathological dataset, one should use the exact same zoom-level, as the images in the pre-trained CNN.

3.2 Deep learning in digital pathology

Deep learning has been applied to many kinds of medical images within the past decade. These include medical images such as X-ray, ultrasound, magnetic resonance images (MRI) as well as lung computed tomography (CT). Main tasks deep learning algorithms have been taught to solve, are classification or sub-classification (e.g. tumor subtype), detection of disease (e.g. knee osteoarthritis) or tissue segmentation (Ker et al. 2017; Shen, Wu, and Suk 2017). Not to forget recent, still ongoing, coronavirus pandemic: with the help of deep neural networks, computer vision algorithms have been able to successfully diagnose COVID-19 disease from lung X-ray images and CT scan (Li et al. 2020; Rahimzadeh and Attar 2020; Wang et al. 2021; Xu et al. 2020). Compared to most medical images, the image file sizes in microscopic imaging are enormous. Nowadays, as the computing capacity of modern processors has grown, deep learning has been applied also to the field of pathology.

In digital pathology microscopic images are from histopathological sections. These can be H&E-stained or immunostained. Immunostaining exposes some particular protein of interest and this field of pathology is called immunohistochemistry. When a pathologist takes a

look at the slides on the microscope or on a computer display, he/she makes his/her own interpretation of the sample. This interpretation can be, e.g., a diagnosis of a disease or, in the case of cancer surgery, an analysis whether the surgical margin is clear from cancerous cells. When automating these kinds of tasks with the help of deep learning solutions, the main goal is to sustain the quality of the results and give pathologists extra time to focus on the most complex cases.

Tasks, which deep learning has been applied in digital pathology, include classification (normal, tumor or tumor subtype), recognition (amount of dividing cells) and segmentation (patch- or pixel-level segmentation) (Litjens 2017). Also more specific tasks are included, such as training deep learning models to predict certain genetic changes from H&E-stained histopathological sections (Coudray et al. 2018; Schmauch et al. 2020; Schaumberg, Rubin, and Fuchs 2018; Chang et al. 2018). This is important, as these analyses in a laboratory environment are time consuming and rather expensive (Echle et al. 2021). Predicting overall survival from histopathological sections is also gaining popularity in the field. The typical tasks in digital pathology regarding CRC are presented more precisely in chapter 4.2.

Main challenges in digital pathology lie in the insatiable data hunger of deep learning algorithms. In order to train a model, that generalizes well, deep learning needs a loads of training data. In the pathology field, to gain training data means hours of work from pathologist to annotate the regions of interest. Lack of annotated data is probably the biggest challenge in digital pathology. Also, as in the field of medical image analysis in general, the privacy issues of patient data cause some challenges. But there are ways to artificially increase the amount of data, e.g., data augmentation discussed briefly in chapter 2.3.

At the moment and in the future, constant development of new digital pathology applications is bringing and will bring plenty of benefits to the field of pathology. Hopefully it will also ease the work in practice by automating some of the daily tasks pathologists deal with. Combining other clinical data (e.g. medical history) with the information from the images, can make deep learning algorithms generate hypotheses we humans never had even thought of. All of these things are studied with one goal in mind: to improve diagnostics and healthcare in general. In addition, the opportunities of deep learning in digital pathology does not only cover research and diagnostics. It benefits also, e.g., the education of pathologists and

quality assurance of histopathology laboratory (Pantanowitz et al. 2018).

4 Colorectal cancer

Colorectal (colon) cancer (CRC) is one of the most malignant epithelial tumors and it is the second most common death-causing cancer in the world (World Health Organization 2020). 5-year survival rate in CRC is 58-65 %, even though there has been major improvement within the last few decades (Siegel, Miller, and Jemal 2020). As with many other types of cancer, also in the case of CRC, the early detection increases the chance of survival.

CRC is epithelial cancer, which begins to develop in the epithelial cells of the colon. Epithelial cells are the cells, which line the inner surface of the colon. Figure 7 provides a generalized view of the colon structure and shows the part where the CRC histopathological sections come from. Figure 7 also presents examples of normal and tumorous epithelium. Normal epithelial cells are symmetric and homogeneous in their appearance. Tumor cells differ from normal epithelial cells and form unique, vague shapes, forming a somewhat heterogeneous group, what comes to their appearance. If tumor cells invade from the epithelium through smooth muscle layer which surrounds the colon, prognosis gets even worse (Klintrup et al. 2005).

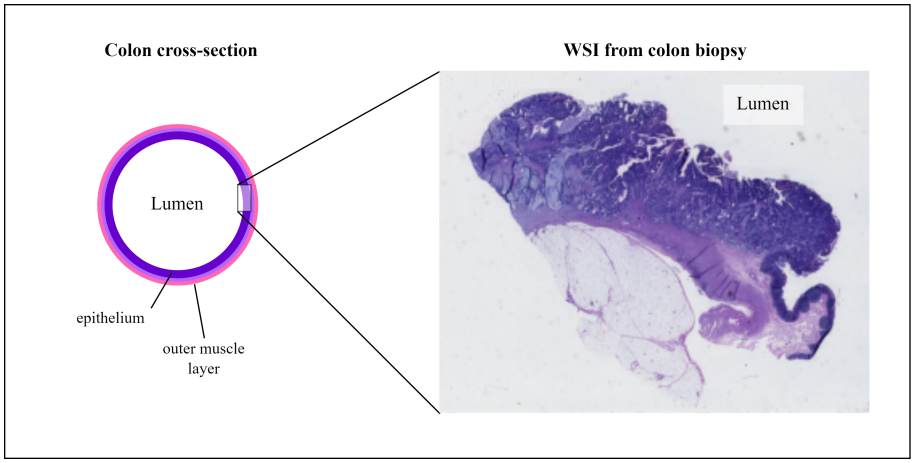

Figure 7: Colorectal cancer begins from the epithelial cells of the colon. Tumorous tissue can grow as polyps towards the lumen of the colon or invade deeper into the surrounding muscle of the colon. Image origin Central Finland Health Care District, "Suolisyöpä Keski-Suomessa 2000-2015" -project".

For prognosis being poor, some significant indicators predicting prognosis have been found. Finding and assessing prognostic factors is important as those factors is one piece of the

puzzle when clinicians are determining the best possible treatment for the patient. One of these factors is the tumor microenvironment, from which tumor-stroma ratio (TSR) can be evaluated.

4.1 Tumor-stroma ratio

TSR is the proportion of stroma within the tumor site, making it one of the indicators telling information about the tumor microenvironment. Tumor microenvironment is the immediate surroundings of the tumorous tissue and it has a potential of being a new target for tumor treatment. Whether the role of it (in this case the role of stroma) is to proliferate and help the tumorgenesis, or, whether the role is to suppress and make the life of a tumor as difficult as possible is still under a debate. Nevertheless, TSR tells much about the microenvironment and the significance of TSR as a prognostic factor has been shown in many types of cancers. The low amount of stroma (TSR $<$ 50 %) associates with better prognosis and multiple studies have shown TSR to be an independent survival predicting factor in many different cancer types (Huijbers 2013; Ma 2012; Mesker 2007).

As the signifigance of TSR to CRC prognosis is obvious, a protocol has been created for pathologists to follow when assessing TSR visually (Van Pelt 2018). This has increased the reproducibility of the evaluation task. Despite that, making the value of TSR reproducible and therefore fully comparable between samples and pathologists is challenging. The evaluation of TSR (visual and automated) is introduced in chapter 4.2.3.

4.2 Convolutional neural networks in colorectal cancer research and applications

In the case of CRC, the image data can be acquired, e.g., from colorectal examination (colonoscopy) or surgery, magnetic resonance imaging (MRI) or computed tomography (CT). In this section, the focus is on the CNNs trained with H&E-stained histopathological images.

As with many other cancer types, the development of image analysis methods regarding

CRC, have mainly been about classification, detection and outcome prediction. In addition, CNNs in CRC have been applied to inflammatory bowel diseases (IBD) detection. This is also very important, because IBDs, such as Crohn's disease, increase the risk of CRC (Kulaylat and Dayton 2010; Pacal et al. 2020). Next, CNNs applied to these tasks, as well as CNNs designed for automated evaluation of TSR, are introduced.

4.2.1 Detection and classification

In CRC image analysis, tasks in detection and classification have been, e.g., about nuclei or tumor nuclei detection, cancer staging or differentiating the detected tumor between malign and benign. Detecting tumor budding is also one detection task to which CNNs have been applied to, because visual detecting of tumor budding is time consuming. Tumor budding means the presence of a small cluster or multiple small clusters of tumor cells in the invasive front of CRC. (Pacal et al. 2020)

Some approaches use the information from nuclei detection to classify the sample into different tissue types. Sirinukunwattana et al. (2015) constructed CNN model to detect and classify four nuclei types. Their CNN-model, Spatially Constrained Convolutional Neural Network (SC-CNN), consists of two 9- and 8-layered networks. The other is designed for detection of nuclei utilizing the spatial information of the detected nuclei. The precision- and recall-scores for this CNN in two different validation folds were 0.758/0.781 and 0.823/0.827, respectively. The best weighted average F1 score of the CNN designed for classification was 0.784. Combined performance of their model was 0.692 (weighted average F1 score for nucleus detection and classification). The architecture of SC-CNN was customized, based on Alexnet-architecture and mitosis-detecting CNN by Cireşan et al. (2013).

Javed et al. (2020) utilized the architecture of the SC-CNN designed by Sirinukunwattana et al. (2015) and a custom 10-layer CNN by Tofighi et al. (2019) and trained their model to detect and classify seven nuclei types instead of four. They applied the information from cell-cell connections to compute a cellular communities in order to recognize the different tissue phenotypes. They compared their model with other approaches, their model gained the best classification performance on two different CRC datasets ($F_1 =$).

Classifying different parts of the specimen to specific categories is a crucial part of many processes in digital pathology. It affects other diagnostic metrics computed based on the classification performed in the first place.

4.2.2 Overall survival and outcome prediction

The prediction of survival from CRC has gained more popularity recently. Skrede et al. (2020) trained ten CNNs and developed a biomarker, which predicts outcome. More value to their work compared to the previous studies gives the easy way their marker is applied in practice. Their model needs no visual annotation, TMA-spots or other clinical information. Instead, it can be applied directly to the WSIs. Model uses two different zoom-levels for prediction, 10x and 40x, and their network is based on MobileNetV2 (Sandler et al. 2018). Their model was in high agreement with the true outcome of the patient (AUC of 0.713).

Kather et al. (2019) trained a CNN to recognize nine different tissue types from CRC WSIs. Best performing CNN architecture when classifying the image tiles, was VGG19, overall classification accuracy on a test set of 7,180 images was 94.3 %. They used the classification results to assess a "deep stroma score", a metrics based on the proportions of five tissue types (adipose tissue, debris, lymphocytes, muscle and stroma). Those tissue types had highest prognostic value and "deep stroma score" was found to have prognostic value in both recurrence-free survival and overall survival (hazard ratios 1.92 and 1.63, respectively).

4.2.3 Automated evaluation of tumor-stroma ratio

When TSR is evaluated visually, pathologists determine it from H&E-stained tissue sections by choosing the part of the specimen containing the largest amount of stroma. In practice, visual evaluation of TSR begins using low magnification on a light microscope. At first the area, which has the largest amount of stroma within the tumor site, is chosen. The final estimate of TSR is performed with higher magnification. TSR is scored by 10 % intervals from 10 % - 90 %, or as “stroma-high” or “stroma-low” (TSR $<$ 50 % or TSR $\geq$ 50 %, respectively) (Van Pelt 2018).

Computational evaluation of TSR is quite a new concept. At first, image tiles extracted from

the histopathological whole-slide image (WSI) have to be classified. After that, TSR can be predicted from the whole image or, e.g., using a particular spot selected by a pathologist.

Regarding CRC, only a few approaches have been introduced from this subject. In general, TSR is more or less additional part of the model outcome, as the first goal in these studies have been to segment the whole tissue sample. In the study, mentioned earlier about nuclei detection by Sirinukunwattana et al. (2015), they were also able to automate the estimation of TSR (Sirinukunwattana et al. 2018). They trained a VGG19-model to classify nine tissue types and the accuracy for detecting stroma and tumor were 90.4 % and 96.0 %, respectively. Controversary to other TSR-related studies, their study did not show prognostic value for TSR.

Zhao (2020) trained a CNN-model to recognize nine tissue types from H&E stained WSIs from CRC. Equally to previous models by Sirinukunwattana et al. (2015) and Kather et al. (2019), the architecture of the model was VGG19. Overall classification accuracy on two test sets were 95.7 % and 97.5 %. Classification accuracies for tumor and stroma were 92.8 % and 70.9 % on test set 1, which was the test set published by Kather, Halama, and Marx (2018). Randomly selected images from Yunnan Cancer Hospital were utilized to establish test set 2. Classification accuracies for test set 2 were 97.2 %, and 89.1 %. Pathologists annotations on a 126 image blocks of size 1 *mu*m x *mu*m areas were segmentated with the model. The segmentation results were in high agreement with pathologists' annotations (Pearson r = 0.939, 95% CI 0.914 - 0.957).

Instead of using the whole slide, the other approach is to calculate TSR from the same circular spot, where the visual TSR evaluation has been estimated. The downside of this is the manual effort needed, making the procedure less automatic. Geessink et al. (2019) developed a CNN based on a 11-layer VGG-architecture and trained it to classify nine tissue types. Using a 50 % cutoff-value between "stroma-high" and "stroma-low", there was a quite big disagreement between the model and pathologist (Cohen's kappa $\kappa = 0.239$). When a median of the TSR estimates computed by the model was used as a cutoff-value for "stroma-high" and "stroma-low", Cohen's kappa-value was slightly improved $\kappa = 0.521$. Also, "Stroma-high"- and "stroma-low"-groups had a strong prognostic value when the median was used as a cutoff-value. The sample size in this current study was quite small (129 patients).

In conclusion, automated TSR evaluation is a challenging task. The prognostic value it brings is mostly dependent on the classification accuracy of the CNN and whether TSR is predicted from the whole specimen or whether particular circular spot is used. If the entire slide is not utilized to calculate the TSR, one challenge yet to conquer is to find a well-performing algorithm that imitates the visual evaluating procedure. This is particularly challenging in the computation time point-of-view as the WSIs are massive in terms of number of pixels.

5 Data and methods

There are two main steps in the practical part of this study: training the CNN-models and predicting the TSR-values. At first, 12 different CNN-models are trained from annotated WSIs to classify image tiles into three different tissue groups: stroma, tumor and other. After this, TSR-values are predicted from the WSIs, which were not annotated and therefore left out of the training data. Those WSIs are tiled and tiles are classified with the models trained earlier. TSR-values are calculated based on the predictions of the model and the predicted TSR-values are then compared with the visually estimated TSR-values by a pathologist. Overall studyflow is presented in figure 8.

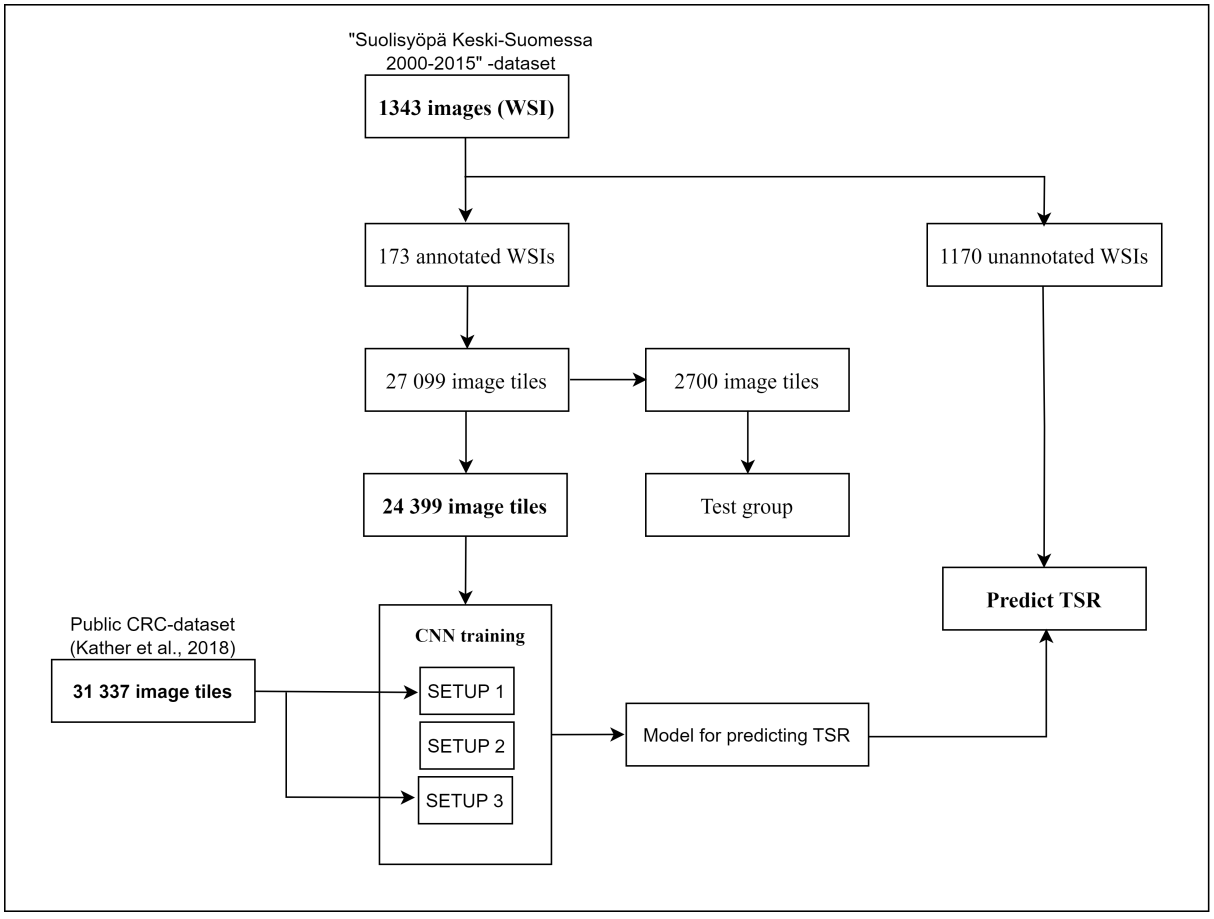

Figure 8: Diagram of the studyflow. The image tiles for the training of the CNN are tiled from 171 WSIs, the TSR-values are predicted for the remaining WSIs and compared with the TSR-values determined visually by the pathologist.

5.1 Data

Two different CRC-datasets are used: a CRC dataset from Central Finland Health Care District, "Suolisyöpä Keski-Suomessa 2000-2015" -project (referred further on as "KSSHP-dataset") and public CRC dataset (referred further on as "Kather-dataset") (Kather, Halama, and Marx 2018). KSSHP-dataset has 1343 anonymized WSIs, 171 of these have annotations and are used in the training of the CNNs. Annotations are made by an experienced pathologist with open source bioimage analysis tool QuPath (Bankhead et al. 2017). From these annotations, 27,099 tiles are derived. One tenth of the image tiles (2700 image tiles) are excluded from the training set into a test group, leaving 24,399 image tiles from KSSHP-data to the CNN training.

Kather-dataset has nine classes, but in this study we need to classify CRC images into three classes. Image groups from Kather-dataset were used as described in table 1. CNN will not be trained to recognize background or adipose tissue. Instead, those it will be clipped off when masking the image.

Original class	Description	Class in this study
ADI	adipose tissue	nan
BACK	image background	nan
DEB	debris (organic waste after cell dies)	Other
LYM	lymphocytes (one type of immune cells)	Other
MUC	mucus	Other
MUS	smooth muscle cells	Other
NORM	normal colon epithelium	Other
STR	stroma	Stroma
TUM	tumor	Tumor

Table 1: Descpription of the use of Kather-dataset, the original name of the class in Kather dataset in the first column. Adipose and background images will not be used in this study, they are removed in the image tiling phase. Classes DEB, LYM, MUC, MUS and NORM will be grouped as one class called "Other". Stroma and tumor classes will be used as such.

Size of each class in Kather-dataset is approximately 10,000 image tiles. Data will be balanced and equal amount of tiles from original classes DEB, LYM, MUC, MUS and NORM

are randomly chosen from original groups to form the group "Other".

Dataset	Image tile size	MPP	Number of image tiles
Kather	224 x 224 px^2	0.5	31,337
KSSHP	224 x 224 px^2	0.5	27,099

Table 2: Image tile size, accuracy (microns per pixel, MPP) and the total number of image tiles.

5.2 Preprocessing

Images in Kather-dataset are already preprocessed, images are 224 x 224 px^2 tiles and color is normalized. KSSHP-data needs to be preprocessed. At first, the annotated parts of the WSIs are tiled into 224 x 224 px^2 image tiles. The tile size was chosen to match the Kather-data and in addition, most of the CNNs in this study require the chosen size as an input-size. Image tiles were acquired from the highest resolution level, where accuracy is 0.5 microns per pixel. After tiling, color normalization by Macenko's method was applied as images in Kather-dataset are normalized with Macenko's method as well (Macenko et al. 2009). Examples of image tiles from each class are presented in figure 9.

If model input-size differed from size of the tiles, tiles were resized. CNN models pre-trained with ImageNet have their own requirements about the image normalization. Input images were normalized using given mean and standard deviation of $[0.485, 0.456, 0.406]$ and $[0.229, 0.224, 0.225]$. Images in SETUP 3 (without ImageNet-pretraining), were normalized according to mean and standard deviation calculated from the images of the Kather-dataset.

5.3 Convolutional neural network architectures

There are numerous architectures which have been successfully used in this kind of classification problem. The architectures are chosen based on their success on histopathology image classification tasks and accessibility. Many CNNs utilized in digital pathology are less or more customized, but as this thesis is rather focused on the bigger picture of automated

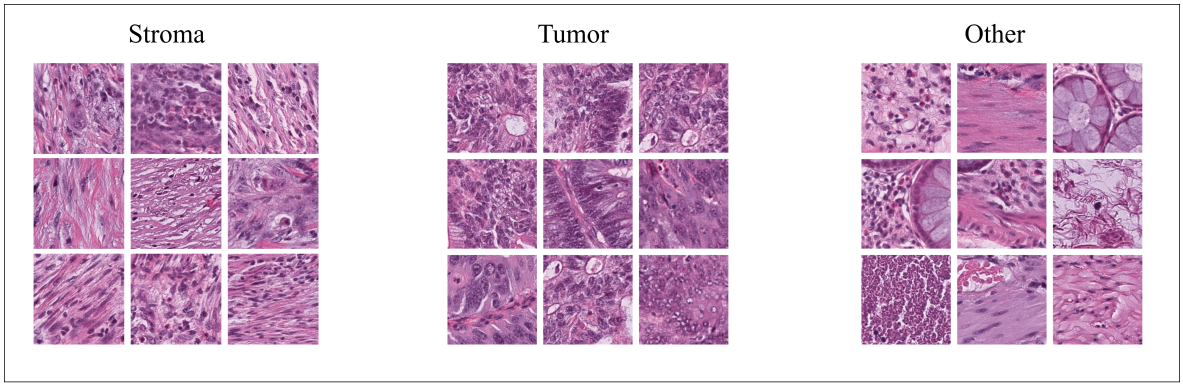

Figure 9: Normalized image tiles from each class from KSSHP-dataset.

TSR evaluation instead of CNN development, all architectures should be easily accessed with a ImageNet pre-trained-option via PyTorch-library (Paszke et al. 2019).

GoogleNet is a 22-layer network, which was the most successfull architecture in CAMELYON16-breast cancer challenge, ResNet was the winner in 2015 ImageNet-challenge and one of the most utilized architectures in CRC-related image analysis (Szegedy et al. 2015; He et al. 2016; Pacal et al. 2020; Litjens et al. 2018). VGG performed well in CRC tissue classification task by Kather, Halama, and Marx (2018). Alexnet is rather small network and known to be quite fast, as well as efficient, which is why Alexnet was included in this study as well (Krizhevsky, Sutskever, and Hinton 2012).

A lot of new, modified versions of these networks have been developed since the original architectures have been published. The exact versions of the architectures chosen in this study are: Alexnet, Googlenet, ResNet50 and VGG19. For detailed structures of the architectures, see references Krizhevsky, Sutskever, and Hinton (2012) (Alexnet), Szegedy et al. (2015) (Googlenet), He et al. (2016) (ResNet50) and Simonyan and Zisserman (2014) (VGG19).

5.4 Training and validating the convolutional neural networks

Three different pre-training approaches are used in this study (SETUP 1, SETUP 2 and SETUP 3). SETUP 1 and 2 are initialized with ImageNet-weights, in SETUP 1, the CNN is also pre-trained with Kather-data before training with KSSHP-data. In SETUP 2, CNN is trained merely on KSSHP-data after ImageNet-weights initialization. SETUP 3 is not

initialized with ImageNet-weights, it is trained "from scratch" with Kather-data and KSSHP-data. With three different training approaches and four different CNN architectures, the total number of different CNNs trained in this study is 12. Table 3 summarizes the training approaches.

Setup name	Training order				
SETUP 1	ImageNet	$\longrightarrow$	Kather	$\longrightarrow$	KSSHP
SETUP 2	ImageNet	$\longrightarrow$	KSSHP		
SETUP 3	Kather	$\longrightarrow$	KSSHP		

Table 3: All three setups were applied with four different CNN architectures, making the total number of different CNNs trained in this study 12.

Training of all setups goes as follows. At first, a test set of 2,700 image tiles was excluded from the image tiles derived from annotated images of the KSSHP-dataset, making the size of the training group 24,399 image tiles. The most suitable hyperparameters were validated using 5-fold cross-validation for the entire training group. Parameters chosen for each architecture and training approach are listed in appendix A.

Training data was then split into training and validation groups, using two-thirds of the data for training and one-third of the data for validation. At this point, the main focus was on the validation loss. All networks were trained for 30-40 epochs and training and validation loss were plotted. The number of epochs for the training of the final model was chosen visually, estimating the point where the validation error was the lowest before starting to increase again (see appendix C). The number of epochs for each model are listed in appendix B. The final model was trained using the complete training dataset of 24,399 image tiles.

5.5 Predicting tumor-stroma ratio

The 1170 unannotated WSIs of the KSSHP-dataset were tiled into 224 x 224 px^2 image tiles and the class of each tile was predicted with each CNN-model. For each WSI, the TSR was calculated based on the number of stroma- and tumor-tiles:

$$TSR = \frac{n_{stroma}}{n_{tumor}+n_{stroma}} \tag{5.1}$$

Where n_{stroma} refers to the total number of stroma tiles and n_{tumor} is the total number of tumor tiles.

5.6 Equipment

Microscope slides were scanned with Hamamatsu NanoZoomer-XR. Computing was performed in Linux GPU server Tesla P100, x 86_64 with Python-version 3.8.5.

6 Results

6.1 Validation results of the final networks

Table 4 lists validation losses and validation accuracies of all CNN-models. The most accurate network was VGG19 with SETUP 1 training approach.

	SETUP 1		SETUP 2		SETUP 3	
	loss	accuracy	loss	accuracy	loss	accuracy
Alexnet	0.132	95.4 %	0.141	95.2 %	0.286	90.6 %
Googlenet	0.102	**97.5** %	0.097	97.2 %	0.104	**97.7** %
ResNet50	0.114	97.3 %	0.092	97.0 %	0.158	94.5 %
VGG19	0.104	**97.8** %	0.095	97.4 %	0.109	96.5 %

Table 4: Validation loss and validation accuracy of all four CNN architectures and three training approaches. Three most accurate models are shown bold.

6.2 Loss and accuracy on the test set

Test set was excluded from the training images and it has 900 image tiles per class, total 2700 images. Table 5 presents the results on the test set.

Models had the most difficulties distinguishing between classes other and stroma as can be

	SETUP 1		SETUP 2		SETUP 3	
	loss	accuracy	loss	accuracy	loss	accuracy
Alexnet	0.113	96.5 %	0.106	97.1 %	0.209	93.4 %
Googlenet	0.071	97.9 %	0.065	**98.5** %	0.086	97.8 %
ResNet50	0.090	97.5 %	0.077	97.3 %	0.106	97.5 %
VGG19	0.068	**98.6** %	0.075	**98.3** %	0.104	96.8 %

Table 5: Loss and accuracy results on the test set. Three most accurate models are shown bold.

seen from confusion matrices plotted from the test set results (see figure 10). Other-class includes smooth muscle, which is easily mixed up with stroma. Example images of most incorrect predictions of VGG19 with training SETUP 1 are shown in appendice D.

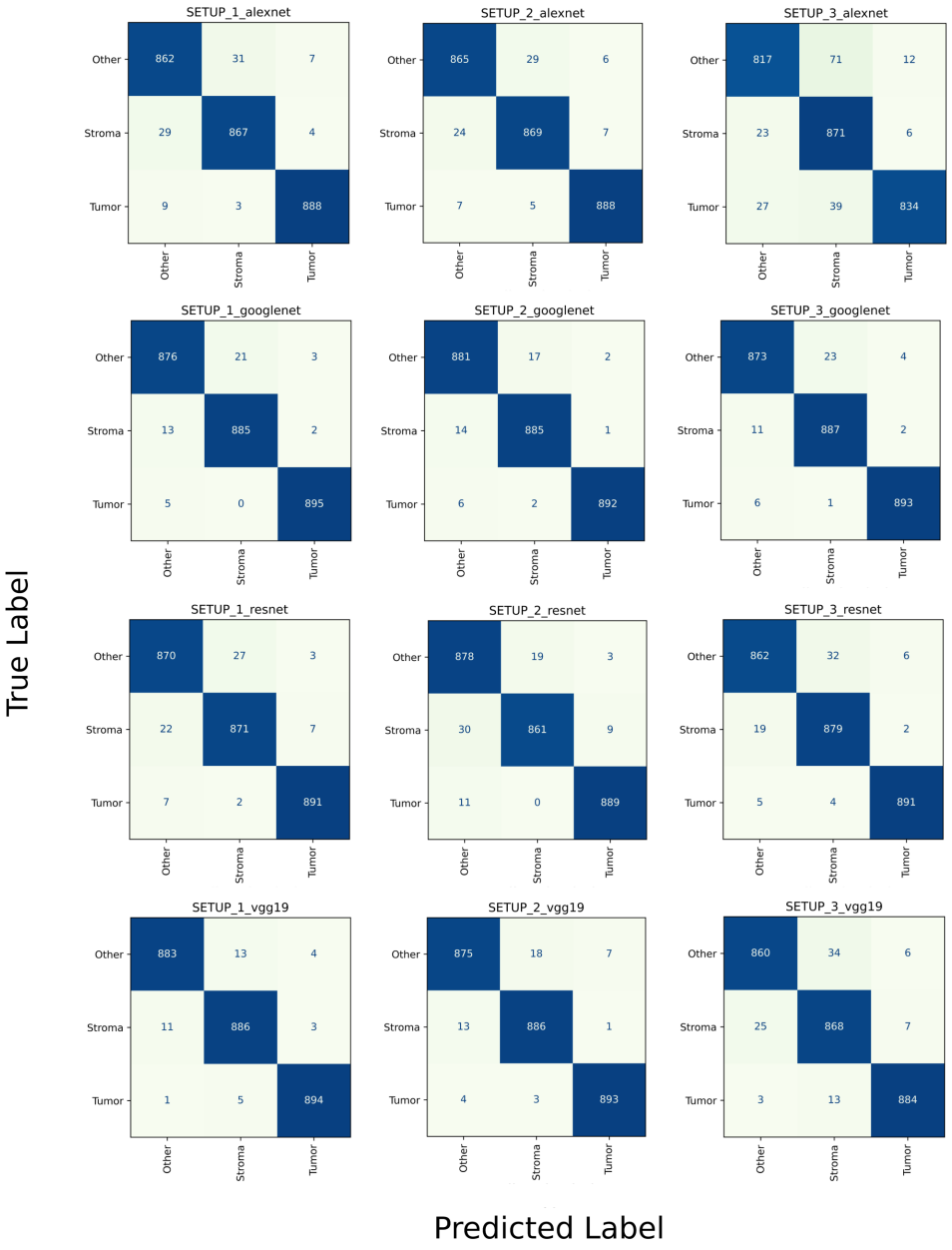

Figure 10: Classification performance of all 12 models on the test group presented as confusion matrices.

6.3 Tumor-stroma ratio predictions

Results from TSR predictions on the test set are shown from the top-3 models (see figure 11). Boxplots and statistics are shown groupwise, based on the true label (TSR-value ranging from 10 % to 90 %).

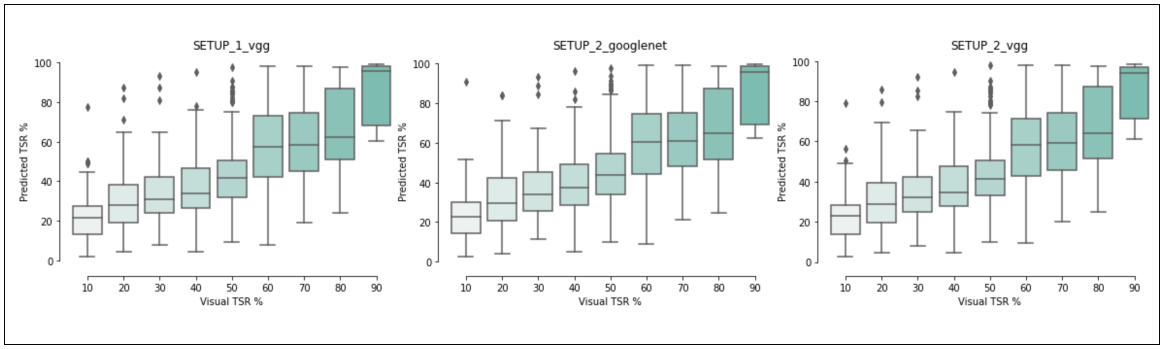

Figure 11: Boxplots showing the groupwise distributions of TSR predictions, results are grouped based on the true label (visually estimated TSR-value, ranging from 10 % to 90 %).

Mean squared error (MSE) and Pearson correlation of all 12 models are presented in figure 7. Mean, median, standard deviation and the number of WSIs in each group (n) are shown in figure 6. These statistics are also shown groupwise, such as boxplots in figure 11.

SETUP 1 / VGG19										
Visual TSR		10%	20%	30%	40%	50%	60%	70%	80%	90%
	mean	24.5	31.9	36.1	38.6	44.4	58.4	60.4	66.1	86.0
Predicted	median	23.8	29.7	31.7	35.8	43.0	58.5	59.2	64.7	95.2
TSR	std	14.7	16.6	16.1	15.0	16.2	19.3	19.1	20.7	15.5
	n	52	94	103	201	282	228	143	51	13
SETUP 2 / Googlenet										
Visual TSR		10%	20%	30%	40%	50%	60%	70%	80%	90%
	mean	25.1	32.8.2	37.0	39.8	45.8	59.6	61.8	67.0	86.4
Predicted	median	22.9	29.6	33.9	37.2	43.8	60.5	61.0	64.7	95.7
TSR	std	15.3	17.1	16.2	15.3	16.5	19.4	19.4	20.7	15.6
	n	52	94	103	201	282	228	143	51	13
SETUP 2 / VGG19										
Visual TSR		10%	20%	30%	40%	50%	60%	70%	80%	90%
	mean	23.8	31.3	35.0	37.5	43.4	57.6	59.7	65.6	85.6
Predicted	median	22.9	28.9	32.1	34.8	41.6	58.1	59.4	64.2	94.2
TSR	std	14.2	16.0	15.4	14.2	15.5	18.9	19.0	20.5	14.7
	n	52	94	103	201	282	228	143	51	13

Table 6: Main statistic metrics from top-3 models of the test results.

6.4 Performance of the final networks

Computing times were also subject of interest in this study. When training the model, training time per epoch is the one to pay attention to. Also, when TSR is calculated based on the predictions made by CNN, the other point of interest is the prediction time per image tile. One WSI can hold up to 20 000-30 000 image tiles (e.g. 224 x 224 px^2), so it's worth noting (See table 8).

	SETUP 1		SETUP 2		SETUP 3	
	MSE	Pearson	MSE	Pearson	MSE	Pearson
Alexnet	363.0	0.551	354.1	0.560	347.8	0.539
Googlenet	346.0	0.554	349.6	0.554	351.8	0.552
ResNet50	376.3	0.547	369.8	0.552	379.7	0.548
VGG19	348.8	0.562	332.1	0.567	348.6	0.564

Table 7: Mean squared error (MSE) of predicted TSR-values and Pearson correlation with visual TSR estimates. Results presented from all 12 models.

Architecture	Number of parameters	Model size (MB)	Training setup	Train time / epoch (s)	Predict time / tile (ms)
Alexnet	61,113,131	244	SETUP 1	182	26.41
			SETUP 2	474	26.82
			SETUP 3	254	11.42
Googlenet	5,602,979	22.6	SETUP 1	989	9.54
			SETUP 2	591	10.11
			SETUP 3	251	3.93
Resnet50	23,514,179	94.4	SETUP 1	370	4.44
			SETUP 2	354	4.56
			SETUP 3	175	3.69
VGG19	143,679,531	575	SETUP 1	738	4.27
			SETUP 2	487	5.56
			SETUP 3	541	6,92

Table 8: Number of parameters and model size of each architecture. Mean training times per epoch and mean predicting times per image tile are shown from all 12 models.

7 Discussion

Overall, the classifying accuracy of all 12 models was excellent and validation results were in good balance with the test results. The test results show, that ImageNet-initialization is highly recommended in the case of histopathological images. Pre-training the network with domain-specific data increased accuracy with ResNet50- and VGG19 -architectures. Although, considering the effort put on pre-training the network with additional dataset, differences in the test accuracies are rather minimal and some of those differences might be explained with randomness. Worth noting is, that it was not tested whether less amount of training iterations with domain-specific data would bring more accuracy to the final model or how data augmentation would increase accuracy.

Results on the test set confirmed that the most difficult part of the classification in all 12 models were between classes stroma and other. This might solely be due to the similarity of smooth muscle and stroma, as part of the other-class are image tiles from smooth muscle. For human eye, these two tissue types are challenging as well. Despite the minor difficulties, the accuracy of stroma classification was higher than those found in previous studies by Kather et al. (2019), Sirinukunwattana et al. (2015), and Zhao (2020).

In this study, the most accurate model was VGG19 with domain-specific pre-training and ImageNet-initialization. When picking up the most suitable model, model performance is usually trade-off between the model accuracy and training time. When those are taken to account, the model with VGG19-architecture, initialized with weights from ImageNet and with or without domain-specific pre-training would be the model of choice at this point. Models using VGG19-architecture were also performing more efficient, in terms of training time and predicting time, when compared to the best Googlenet-model. Computation times are important to keep in mind, and to minimize those without losing the accuracy, would be worth studying.

Considering the differences between the visual TSR scoring and the automated version in this study, TSR-values predicted by the models were correlating rather well. Bringing together few pathologists to evaluate over 1000 WSIs might not end up in perfect correlation either.

Overview by Van Pelt (2018) showed, that the inter-observer kappa-values for visual TSR, using the scoring method recommended by Van Pelt (2018), ranged from 0.60 to 0.89. The TSR correlation results are a bit difficult to compare with any previous studies regarding TSR in CRC, as Zhao (2020) utilized the annotations as ground truth, Kather et al. (2019) and Sirinukunwattana et al. (2018) studied the TSR association with prognosis and Geessink et al. (2019) scored the samples into "stroma-low"- and "stroma-high" -groups. Differing from those studies, a pathologist's visual TSR estimate was the ground truth in this study. It would be interesting and valuable to study the correlation of the TSR predicted by the model with several pathologists' TSR estimates.

Nevertheless, some aspects from the visual evaluation should be brought to the automated model. For example, only those stroma tiles, which are tumor-related would be taken into account. One other option is to mimic the visual process by going through the image frame by frame and choosing one spot to make the final TSR prediction. Once again, computing times should be noticed in this case. In addition to these, using a smaller tile size might increase classifying accuracy since some tumor areas, as well as stromal areas in between, seem to be quite narrow. This might have a significant effect on the final TSR predictions.

Even though this automated version of TSR evaluation is not fully comparable with the visual one, the reproducibility of a computational method is a major advantage over the visual TSR evaluation. All models solved the classification problem without major problems and what comes to the TSR predictions, the correlations with visual TSR estimates were satisfactory. This kind of tool in daily practice would definately bring reproducibility to the TSR evaluation process and would be a handy tool, making the TSR evaluation in a fraction of a time compared to the visual method.

As TSR estimated visually has been shown to be an independent prognostic factor in solid cancer types, it would be valuable to study the correlation of TSR predicted by this model with other clinicopathological factors and overall survival of the patient.

Bibliography

Al–Janabi, Shaimaa, Andre Huisman, and Paul J Van Diest. 2012. “Digital pathology: current status and future perspectives”. *Histopathology* 61 (1): 1–9.

Aggarwal, Charu C, et al. 2018. “Neural networks and deep learning”. *Springer* 10:978–3.

Azam, Ayesha S, Islam M Miligy, Peter KU Kimani, Heeba Maqbool, Katherine Hewitt, Nasir M Rajpoot, and David RJ Snead. 2021. “Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis”. *Journal of Clinical Pathology* 74 (7): 448–455.

Bankhead, Peter, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. 2017. “QuPath: Open source software for digital pathology image analysis”. *Scientific reports* 7 (1): 1–7.

Bayramoglu, Neslihan, and Janne Heikkilä. 2016. “Transfer learning for cell nuclei classification in histopathology images”. In *European Conference on Computer Vision,* 532–539. Springer.

Chang, P, J Grinband, BD Weinberg, M Bardis, M Khy, G Cadena, M-Y Su, S Cha, CG Filippi, D Bota, et al. 2018. “Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas”. *American Journal of Neuroradiology* 39 (7): 1201–1207.

Cireşan, Dan C, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. 2013. “Mitosis detection in breast cancer histology images with deep neural networks”. In *International conference on medical image computing and computer-assisted intervention,* 411–418. Springer.

Colangelo, Tommaso, Giovanna Polcaro, Livio Muccillo, Giovanna D’Agostino, Valeria Rosato, Pamela Ziccardi, Angelo Lupo, Gianluigi Mazzoccoli, Lina Sabatino, and Vittorio Colantuoni. 2017. “Friend or foe?: The tumour microenvironment dilemma in colorectal cancer”. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1867 (1): 1–18.

Coudray, Nicolas, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. 2018. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning". *Nature medicine* 24 (10): 1559–1567.

Echle, Amelie, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. 2021. "Deep learning in cancer pathology: a new generation of clinical biomarkers". *British journal of cancer* 124 (4): 686–696.

Farahani, Navid, Anil V Parwani, and Liron Pantanowitz. 2015. "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives". *Pathology and Laboratory Medicine International* 7:23–33.

Fukushima, Kunihiko, and Sei Miyake. 1982. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In *Competition and cooperation in neural nets,* 267–285. Springer.

Geessink, Oscar GF, Alexi Baidoshvili, Joost M Klaase, Babak Ehteshami Bejnordi, Geert JS Litjens, Gabi W van Pelt, Wilma E Mesker, Iris D Nagtegaal, Francesco Ciompi, and Jeroen AWM van der Laak. 2019. "Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer". *Cellular Oncology* 42 (3): 331–341.

Goode, Adam, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. 2013. "OpenSlide: A vendor-neutral software foundation for digital pathology". *Journal of pathology informatics* 4.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press. http://www.deeplearningbook.org.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep residual learning for image recognition". In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 770–778.

Hinton, Geoffrey E, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. "Improving neural networks by preventing co-adaptation of feature detectors". *arXiv preprint arXiv:1207.0580.*

Hubel, David H, and Torsten N Wiesel. 1962. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. *The Journal of physiology* 160 (1): 106–154.

Huijbers, A. 2013. “The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial”. *Annals of oncology* 24 (1): 179–185.

Javed, Sajid, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. 2020. “Cellular community detection for tissue phenotyping in colorectal cancer histology images”. *Medical image analysis* 63:101696.

Kather, Jakob Nikolas, Niels Halama, and Alexander Marx. 2018. *100,000 histological images of human colorectal cancer and healthy tissue.* URL:.

Kather, Jakob Nikolas, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. 2019. “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study”. *PLoS medicine* 16 (1): e1002730.

Ker, Justin, Lipo Wang, Jai Rao, and Tchoyoson Lim. 2017. “Deep learning applications in medical image analysis”. *Ieee Access* 6:9375–9389.

Kevin, Zhou, Greenspan Hayit, and Shen Dinggang. 2017. *Deep Learning for Medical Image Analysis.* The Elsevier and MICCAI Society Book Series. Academic Press. ISBN: 9780128104088.

Kieffer, Brady, Morteza Babaie, Shivam Kalra, and Hamid R Tizhoosh. 2017. “Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks”. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA),* 1–6. IEEE.

Klintrup, Kai, Johanna M Mäkinen, Saila Kauppila, Päivi O Väre, Jukka Melkko, Hannu Tuominen, Karoliina Tuppurainen, Jyrki Mäkelä, Tuomo J Karttunen, and Markus J Mäkinen. 2005. “Inflammation and prognosis in colorectal cancer”. *European journal of cancer* 41 (17): 2645–2654.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “Imagenet classification with deep convolutional neural networks”. *Advances in neural information processing systems* 25:1097–1105.

Kulaylat, Mahmoud N, and Merril T Dayton. 2010. “Ulcerative colitis and cancer”. *Journal of Surgical Oncology* 101:706–712.

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

Li, Lin, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. 2020. “Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT”. *Radiology.*

Litjens, Geert. 2017. “A survey on deep learning in medical image analysis”. *Medical image analysis* 42:60–88.

Litjens, Geert, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 2018. “1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset”. *GigaScience* 7 (6): giy065.

Ma, Wei. 2012. “Tumor-stroma ratio is an independent predictor for survival in esophageal squamous cell carcinoma”. *Journal of Thoracic Oncology* 7 (9): 1457–1461.

Maas, Andrew L, Awni Y Hannun, Andrew Y Ng, et al. 2013. “Rectifier nonlinearities improve neural network acoustic models”. In *Proc. icml,* 30:3. 1. Citeseer.

Macenko, Marc, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. 2009. “A method for normalizing histology slides for quantitative analysis”. In *2009 IEEE international symposium on biomedical imaging: from nano to macro,* 1107–1110. IEEE.

Menegola, Afonso, Michel Fornaciali, Ramon Pires, Sandra Avila, and Eduardo Valle. 2016. “Towards automated melanoma screening: Exploring transfer learning schemes”. *arXiv preprint arXiv:1609.01228.*

Mesker, Wilma E. 2007. “The carcinoma–stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage”. *Analytical Cellular Pathology* 29 (5): 387–398.

Mikołajczyk, Agnieszka, and Michał Grochowski. 2018. “Data augmentation for improving deep learning in image classification problem”. In *2018 international interdisciplinary PhD workshop (IIPhDW),* 117–122. IEEE.

Mormont, Romain, Pierre Geurts, and Raphaël Marée. 2018. “Comparison of deep transfer learning strategies for digital pathology”. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops,* 2262–2271.

Nair, Vinod, and Geoffrey E Hinton. 2010. “Rectified linear units improve restricted boltzmann machines”. In *Icml.*

Pacal, Ishak, Dervis Karaboga, Alper Basturk, Bahriye Akay, and Ufuk Nalbantoglu. 2020. “A comprehensive review of deep learning in colon cancer”. *Computers in Biology and Medicine,* 104003.

Pallua, J.D. 2020. “The future of pathology is digital”. *Pathology-Research and Practice* 216 (9): 153040.

Pantanowitz, Liron, Ashish Sharma, Alexis B Carter, Tahsin Kurc, Alan Sussman, and Joel Saltz. 2018. “Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives”. *Journal of pathology informatics* 9.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. “Pytorch: An imperative style, high-performance deep learning library”. *Advances in neural information processing systems* 32:8026–8037.

Rahimzadeh, Mohammad, and Abolfazl Attar. 2020. “A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2”. *Informatics in Medicine Unlocked* 19:100360.

Reinhard, Erik, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. 2001. “Color transfer between images”. *IEEE Computer graphics and applications* 21 (5): 34–41.

Roy, Santanu, Alok kumar Jain, Shyam Lal, and Jyoti Kini. 2018. “A study about color normalization methods for histopathology images”. *Micron* 114:42–61.

Russakovsky, Olga. 2015. “Imagenet large scale visual recognition challenge”. *International journal of computer vision* 115 (3): 211–252.

Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 4510–4520.

Schaumberg, Andrew J, Mark A Rubin, and Thomas J Fuchs. 2018. “H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer”. *BioRxiv,* 064279.

Schmauch, Benoıt, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. 2020. “A deep learning model to predict RNA-Seq expression of tumours from whole slide images”. *Nature communications* 11 (1): 1–15.

Sharmay, Yash, Lubaina Ehsany, Sana Syed, and Donald E Brown. 2021. “HistoTransfer: Understanding Transfer Learning for Histopathology”. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI),* 1–4. IEEE.

Shen, Dinggang, Guorong Wu, and Heung-Il Suk. 2017. “Deep learning in medical image analysis”. *Annual review of biomedical engineering* 19:221–248.

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal. 2020. “Cancer statistics, 2020”. *Ca-a Cancer Journal for Clinicians* 70 (1): 7–30.

Simonyan, Karen, and Andrew Zisserman. 2014. “Very deep convolutional networks for large-scale image recognition”. *arXiv preprint arXiv:1409.1556.*

Sirinukunwattana, Korsuk, Shan E Ahmed Raza, Yee-Wah Tsang, David Snead, Ian Cree, and Nasir Rajpoot. 2015. “A spatially constrained deep learning framework for detection of epithelial tumor nuclei in cancer histology images”. In *International Workshop on Patch-based Techniques in Medical Imaging,* 154–162. Springer.

Sirinukunwattana, Korsuk, David Snead, David Epstein, Zia Aftab, Imaad Mujeeb, Yee Wah Tsang, Ian Cree, and Nasir Rajpoot. 2018. “Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer”. *Scientific reports* 8 (1): 1–13.

Skrede, Ole-Johan, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. 2020. “Deep learning for prediction of colorectal cancer outcome: a discovery and validation study”. *The Lancet* 395 (10221): 350–360.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: a simple way to prevent neural networks from overfitting”. *The journal of machine learning research* 15 (1): 1929–1958.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. “Going deeper with convolutions”. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 1–9.

Tofighi, Mohammad, Tiantong Guo, Jairam KP Vanamala, and Vishal Monga. 2019. “Prior information guided regularized deep learning for cell nucleus detection”. *IEEE transactions on medical imaging* 38 (9): 2047–2058.

Vahadane, Abhishek, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. 2016. “Structure-preserving color normalization and sparse stain separation for histological images”. *IEEE transactions on medical imaging* 35 (8): 1962–1971.

Van Pelt, G.W. 2018. “Scoring the tumor-stroma ratio in colon cancer: procedure and recommendations”. *Virchows Archiv* 473 (4): 405–412.

Wan, Li, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. 2013. “Regularization of neural networks using dropconnect”. In *International conference on machine learning,* 1058–1066. PMLR.

Wang, Shuai, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. 2021. “A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)”. *European radiology,* 1–9.

Werbos, Paul J. 1990. “Backpropagation through time: what it does and how to do it”. *Proceedings of the IEEE* 78 (10): 1550–1560.

World Health Organization, WHO. 2020. *Cancer,* https://www.who.int/news-room/fact-sheets/detail/cancer.

Xu, Bing, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. “Empirical evaluation of rectified activations in convolutional network”. *arXiv preprint arXiv:1505.00853.*

Xu, Xiaowei, Xiangao Jiang, Chunlian Ma, Peng Du, Xukun Li, Shuangzhi Lv, Liang Yu, Qin Ni, Yanfei Chen, Junwei Su, et al. 2020. “A deep learning system to screen novel coronavirus disease 2019 pneumonia”. *Engineering* 6 (10): 1122–1129.

Zeiler, Matthew D, and Rob Fergus. 2013. “Stochastic pooling for regularization of deep convolutional neural networks”. *arXiv preprint arXiv:1301.3557.*

Zhao, Ke. 2020. “Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer”. *EBioMedicine* 61:103054.

Appendices

A Parameter tuning

	SETUP 1		SETUP 2		SETUP 3	
	LR	Optimizer	LR	Optimizer	LR	Optimizer
Alexnet	5e-3	SGD	1e-3	Adam	4e-3	Adam
Googlenet	4e-3	Adam	4e-3	Adam	2e-3	SGD
ResNet50	2e-3	SGD	2e-3	SGD	4e-4	Adam
VGG19	1e-2	SGD	9e-3	SGD	8e-3	SGD

Table 9: Chosen parameters in the final training phase of all models. LR refers to learning-rate and optimizer is the optimization-function used. Parameters were chosen using 5-fold cross-validation. Loss-function was cross-entropy for all setups.

B Number of epochs

	SETUP 1	SETUP 2	SETUP 3
Alexnet	13	10	4
Googlenet	23	35	35
ResNet50	7	5	24
VGG19	14	13	37

Table 10: Number of epochs when training the final model.

C Training and validation loss

Figure 12: Training and validation losses for all setups and CNNs. Vertical line represents the point where training was stopped when training the final model.

D Incorrect predictions

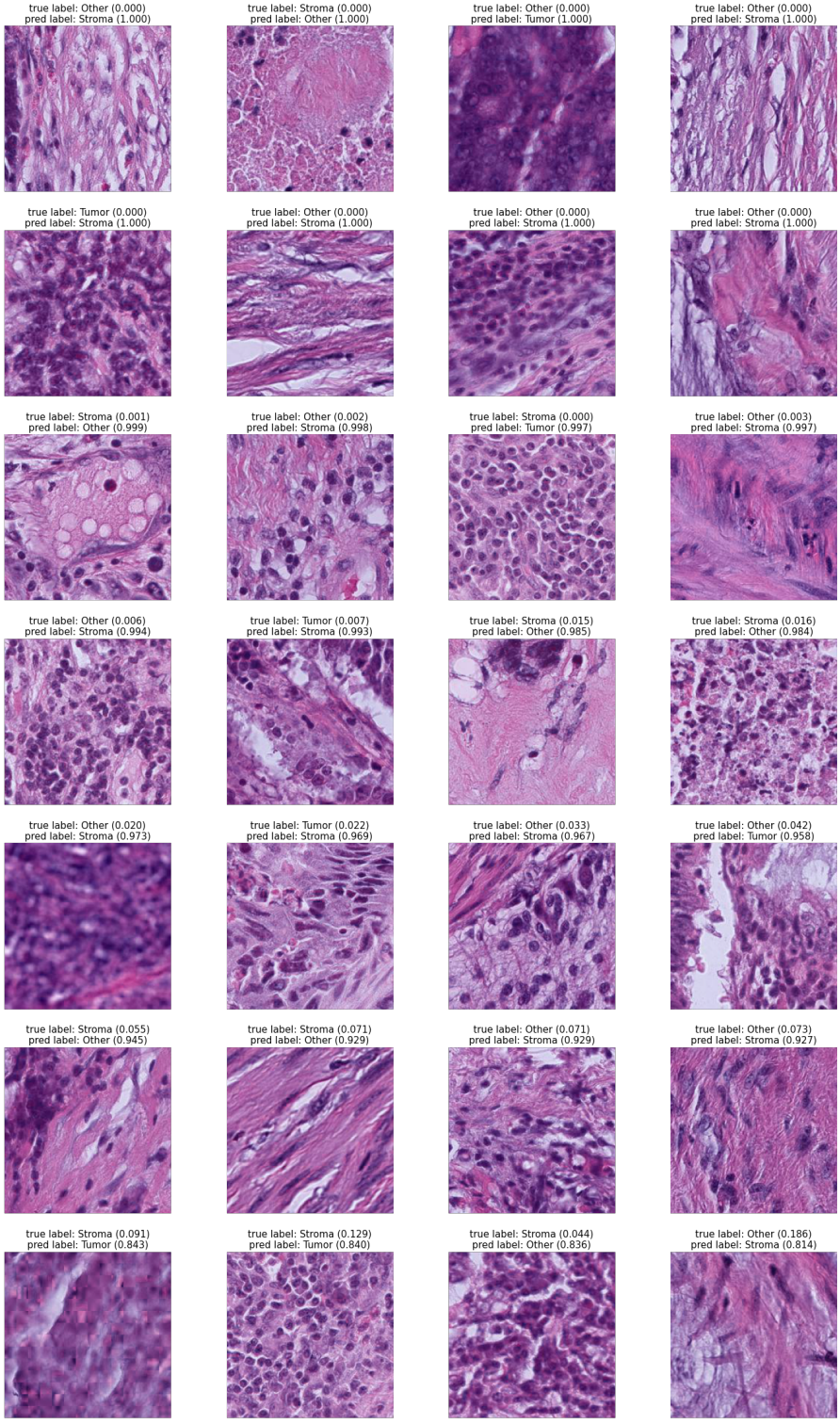

Figure 13: Example images of incorrect predictions of the most accurate model VGG19 pre-trained with ImageNet and domain-specific data. Model gets often confused between smooth muscle (from class other) and stroma.

E Example of tumor-stroma prediction

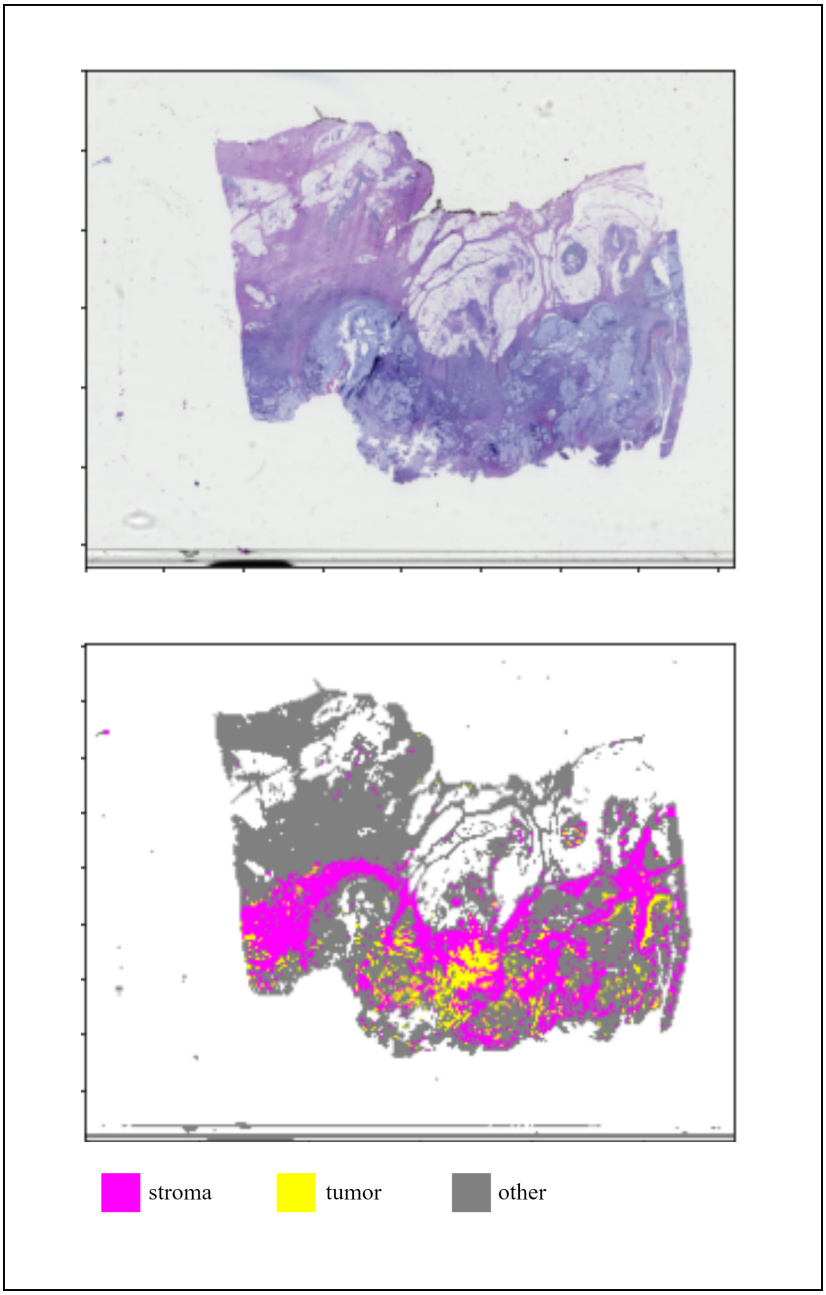

Figure 14: TSR predicted for this WSI were in a range from 81.5 % to 87.2 % and within the most accurate three models the range was from 84.9 % to 85.4 %. Visually estimated TSR was 50 %. This is mainly due the differences within the manual and automated versions. Adding a distance metrics to the algorithm might bring these two methods closer to each other, , ignoring those stroma-tiles which are far from the main tumor site.