University of Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER THESIS

Study programme/specialisation: Masters of Science and Engineering / Signal processing, Robot -and Health Technology	Spring 2023 Open
Author: Mari Amdalsrød Hognestad	
Course responsible: Trygve Chrsitian Eftestøl Supervisor(s): Trygve Christian Eftestøl, Dordi Lea	
Title: Classification of histological images of colorectal cancer using deep learning Credits: 30	
Keywords: Colorectal Cancer, Deep Learning, Classification, Digital Pathology, Convolutional Neural Network	Pages: 51 + appendix: 6, bibliography: 5 Stavanger 15. juni 2023

Abstract

MSc project within biomedical image analysis of histological images of colorectal cancer in collaboration with Stavanger University hospital (SUS).

Colorectal cancer is one of the most common forms of cancer in Norway and the incidence is increasing. In cases of cancer, the surgical preparation (tumour) will be assessed by a pathologist to determine the degree of severity that governs further treatment. To facilitate the workflow of pathologists, it is possible to develop machine learning models that can recognize tissue types and structures in digitized histological images, referred to as whole slide images (WSI).

This master's thesis comprises two sub projects; pre-processing and model training. The pre-processing part consists of manually delineating the tumour from normal tissue in whole slide images with an annotation tool. The algorithm is designed to take a whole slide image (.ndpi) with the corresponding XML file (.ndpa) and then return tiles within the annotated region of interest (ROI). By defining the size and level of the tiles to to be extracted, they will be collected in separate folders for each tissue type. These folders forms the basis for the dataset that will be utilized for model training.

The training process will use the constructed dataset to train a model based on machine learning algorithms. Both binary and multi classification models were carried out. The results demonstrated room for improvement with an overall f1-score of 0.74 for the binary classification and 0.38 for the multi classification.

Preface

This thesis was conducted during the spring semester of 2023 and it signifies the completion of the Master of Science degree in Cybernetics and Signal Processing from the Department of Electrical Engineering and Computer Science at the University of Stavanger.

Initially, the task was to analyze immune cells and the immune response in patients with colorectal cancer. Challenges such as a new slide scanner and thus a new file format and coordinate system, gigapixel images and problems with GPU resources led to more work in the pre-processing part than first assumed. For that reason, the task had to be restructured and tissue classification became the alternative.

I am grateful for the opportunity to work with technology within the health sector. Contributing to improving the healthcare system using innovative solutions is truly inspiring and a significant part of my motivation.

I would like to express my deepest gratitude to my head supervisor, Prof. Trygve Christian Eftestøl from the University of Stavanger, for his support and guidance during this semester. I would also like to thank my co-supervisor Dr. Dordi Lea from Stavanger University hospital for her time and medical knowlege. Their expertise and insight have been essential in shaping the direction of my work.

Stavanger, 15. June 2023
Mari Amdalsrød Hognestad

Contents

Glossary

ACROBATICC Assessment of clinically related outcomes and biomarker analysis for translational integration in colorectal cancer.

AI Artificial intelligence.

AJCC American Joint Committee on Cancer.

AUC Area Under the Curve.

CEA Carcinoembryonic antigen.

CNN Convolutional neural network.

CPU Central Processing Unit.

CT Computed Tomography.

FN False negative.

FP False positive.

FT Fatty tissue.

GI Gastrointestinal tract.

GPU Graphics Processing Unit.

HE Hematoxylin and Eosin.

IM Invasive margin.

JAMA Journal of the American Medical Association.

MP Muscularis propria.

MRI Magnetic Resonance Imaging.

PET Positron Emission Tomography.

ROI Region of interest.

SUS Stavanger University Hospital.

TC Tumor center.

TN True negative.

TNM Tumour-Node-Metastasis.

TP True positive.

VGG16 Visual Geometry Group 16.

WSI Whole slide image.

Chapter 1

Introduction

Cancer represents an enormous health burden and is one of the leading causes of death globally. In Norway, 38.265 new cancer cases were reported by The Norwegian Cancer Registry in 2022, and among these cases, 4.652 were colorectal cancer [74].

Colorectal cancer is a common term for cancer that occurs in the colon and rectum. In recent decades, the incidence of colorectal cancer has more than doubled. Research shows that the earlier the cancer is detected, the greater the probability of survival. This is because over time the cancer will potentially spread and develop into worse stages (I-IV). In advanced disease, the cancer will have spread so deep into the intestinal wall and thus more extensive surgery is needed to remove it.

The world is constantly developing as a result of technology, and in the field of pathology microscopes have been replaced with digitally scanned slides. This development has facilitated what is commonly referred to as digital pathology. Instead of pathologists having to examine a sample through a light microscope, the tissue slide is scanned and converted into a digital whole slide image (WSI). These WSIs are high-resolution images with an optical magnification of up to 40x, meaning they have detailed information extending to the cellular level. In this way, it is possible to zoom in and out on the digital slide to evaluate a sample. In order to process a high-resolution WSI, you have to consider memory limitation. Therefore, each WSI needs to be divided into several sub-images, known as tiles, to reduce the machine's memory usage.

Lack of health personnel is a big issue that faces the modern healthcare system. After a sample has been taken and converted to a WSI, a longer response time may occur due to the lack of pathologists. In the hope of reducing the response time, artificial intelligence (AI) based algorithms are used to ease the workflow of pathologists. Automated image analysis, screening and prognostic predictions are some examples where artificial intelligence can be used.

The former involves training a model to recognize and quantify different cell structures and changes in tissues. As this applies to image recognition, a convolutional neural network (CNN) would be a preferred machine learning method. There are several well-known pre-trained CNN models, one example of such a model being VGG16.

1.1 Project description

The aim of this thesis is to classify different tissue types in a WSI using deep learning. 174 whole slide images were provided to be annotated before undergoing pre-processing. The annotation will be done in different colours, where each color represents a type of tissue. After the annotation has been made, the metadata is put together with the corresponding WSI where it will then extract smaller sub-images from the annotated ROIs. Each tile will then contain one of the tissue types; tumour centre (TC), invasive margin (IM), muscularis propria (MP) or fatty tissue (FT). The tiles will then be separated into different folders for each class. Furthermore, each folder is divided into training, test and validation sets.

After the dataset has been processed, a deep learning algorithm for multi and binary classification will be constructed. Since the data is labeled with the correct output, the model will be trained with supervised learning. The pre-trained VGG16 model was used as a feature extractor for transfer learning, with some few layers added to make the structure even deeper. The fully trained classification models will then be experimented on an unknown test set to produce results. The result will eventually be evaluated by various metrics.

Chapter 2

Medical background

This chapter will provide a deeper understanding of basic pathology, specifically knowledge required to cancer in colon and rectum.

2.1 Cancer

Cell division is a natural process in the body that replaces cells that have died. In some cases, mutations can occur in the cell's genetic material that disrupts this normal regulation. This leads to abnormal cells that grow rapidly and uncontrollably. Eventually a cancerous tumour is formed [50].

Cancer accounted for 9.6 million deaths worldwide in 2018, making it the second leading cause of death globally [66]. Since then, cancer has continued to be a health burden in society. The Norwegian Cancer Registry reported that 38,265 new cancer cases occurred in 2022, which is a larger annual increase than normal [64]. But although the number of cancer cases is increasing, the number of cancer survivors is also increasing. Between 2017 and 2021, the five-year relative survival total was 77.1% for men and 76.3% among women. The fact that more and more people recover from cancer can be explained by a strong health system that offers prevention, earlier diagnosis, and more adapted treatment methods for the various forms of cancer [52].

2.1.1 Colorectal cancer

Colorectal cancer, as the name suggests, is a type of cancer that occurs in the colon or rectum. Both colon and rectum share similar tissue structure, function and the possibility of cancer. Therefore, cancers that arise in these areas will have similar characteristics, risk factors and treatment methods. Hence given the common term colorectal cancer [73].

Colon and rectum are important components of the gastrointestinal (digestive) tract and together they make up the large intestine. The large intestine is located in the lower abdomen, between the small intestine and the anal opening. It is responsible for the last stage of digestion and absorbs water and salts from the intestinal contents, approximately 0.3-0.5 liters per day, and can store the intestinal contents (chymus) temporarily as a waste product or excrement (faeces), before emptying via the anal opening as faeces [72].

Colorectal cancer is a serious disease that affects millions of people worldwide. It is one of the most common forms of cancer in the world. It is therefor important that knowledge of colorectal cancer improves in order to prevent cancer related death through better diagnostics and treatment.

Epidemiology

Colorectal cancer is the second most frequent form of cancer in Norway [75]. In 2022, 3,385 people was diagnosed with colon cancer and 1,267 people rectal cancer. According to the 2022 annual report from the Norwegian Cancer Registry, the incidence of colon cancer was evenly distributed, while for rectal cancer the occurrence was highest among men.[74]

What particularly stands out about the epidemiology of colorectal cancer is the increase in cases over the last 50 years. The largest increase in cases of colorectal cancer occurred in the late 1960s, early 1970s, and since then the incidence has doubled [12]. However, the survival has also increased in recent decades [53]. This may be due to detection of precursors at an early stage and better treatment. Recently a screening program have been rolled out nationally for people over 55 years in Norway. The effect of this program is still unknown, but studies have shown that it reduces the risk of colorectal cancer related death [13].

It cannot be determined with certainty the specific causes of colorectal cancer, but epidemiological studies indicate that lifestyle plays a role [12]. This can apply to both diet, smoking, high alcohol consumption and obesity. In addition, people with inflammatory bowel disease or a family history of colorectal cancer will also have an increased risk of colorectal cancer [86].

2.2 Tissue types

The tissues that make up the gastrointestinal tract (GI-tract) include the mucosa, submucosa, muscularis propria and serosa, see figure 2.1. Each tissue has a different morphology and plays an important role in the function of the colon and rectum [32].

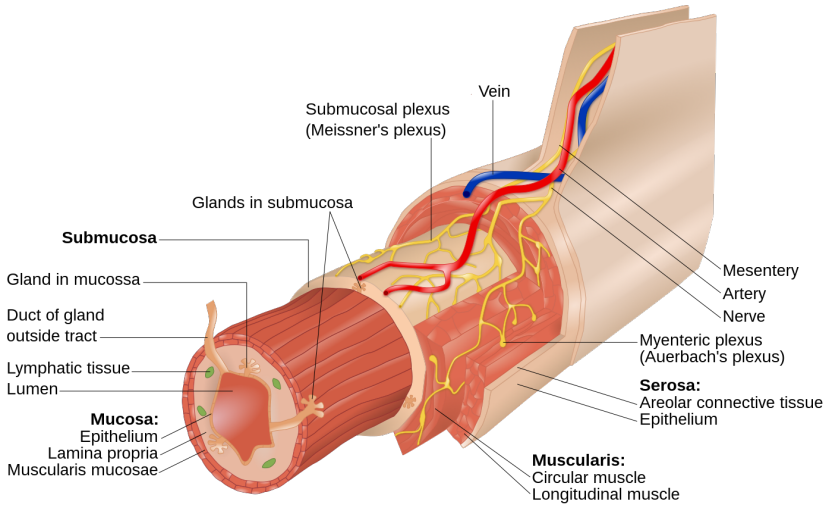
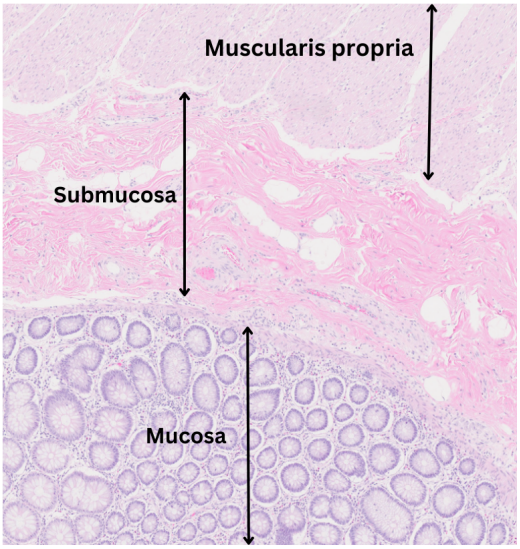
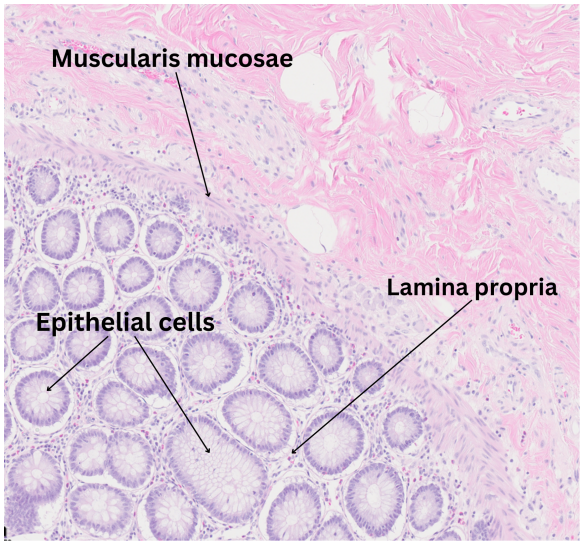


Figure 2.1: Illustration of the GI-tract. Image by Goran tek-en, licensed under the Creative Commons BY-SA 4.0 license [88].

The inner lining of the intestinal wall is the mucous membrane, also known as mucosa, and it consists of three layers: epithelium, lamina propria and muscularis mucosae, see figure 2.2b. The epithelium forms the luminal surface and is the part that comes into contact with the contents of the intestine. The epithelium consists of specialized cells that form a protective barrier and prevent unwanted substances from penetrating the intestinal wall. Just below the epithelium is the connective tissue, lamina propria, which contains glands, lymphatic vessels and some lymph follicles. Its purpose is to support the epithelial cells. Muscularis mucosa is the thin muscle layer in the mucosa. It is located under lamina propria and consists of smooth muscle cells, see figure 2.2 [18].



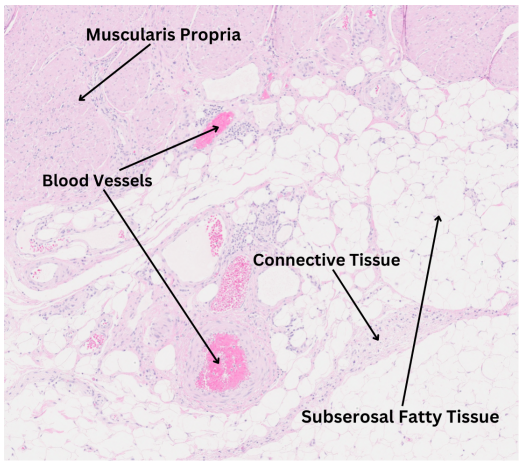
(a) Three innermost layers; mucosa, submucosa and muscularis propria, respectively.



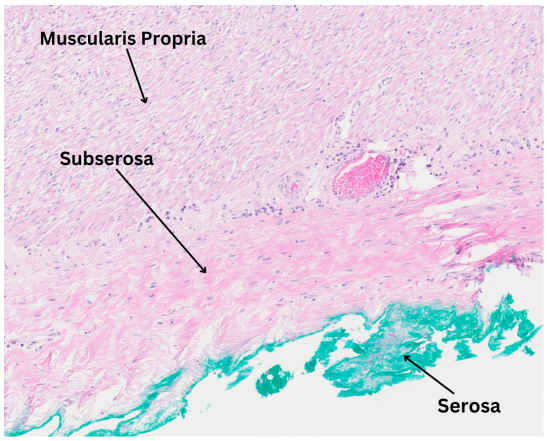
(b) The mucosa is composed of three layers; epithelium, lamina propria, and muscularis mucosae.

Figure 2.2: Sections from a scanned histological whole slide image stained with Hematoxylin and Eosin (HE).

Submucosa is an irregular connective tissue that binds the mucosa to the muscularis propria. It contains blood and lymph vessels as well as nerves that take care of secretion and motility. Muscularis propria lies as a thick layer of smooth muscle around the submucosa. It is responsible for peristaltic activity, which means the rhythmic muscle contractions that send the contents in a specific direction [18]. The outermost layer of the intestinal wall is the serosa, which consists of connective tissue and a single layer of epithelium [32]. Between muscularis propria and serosa is subserosa. Subserosa is composed of blood vessels, nerves and lymphatic vessels, and in some cases it may contain an ample amount of fatty tissue, know as subserosal fatty tissue, see figure 2.3.



(a) Subserosa with subersosal fatty tissue.



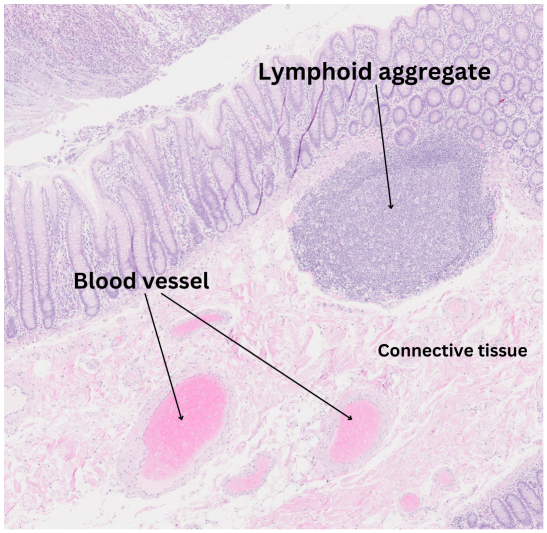
(b) Serosa and subserosa without subersosal fatty tissue.

Figure 2.3: Sections from a scanned histological whole slide image stained with Hematoxylin and Eosin (HE). Note that the serosa is inked in green, it does not have that color originally.

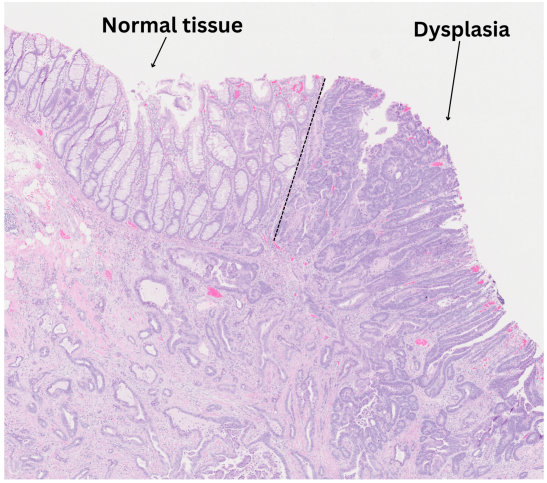
When analyzing histological images, see figure 2.4a, it can be useful to understand and recognise the characteristics of different tissue types in the image. Blood vessels transport blood around the body and ensure the supply of oxygen and nutrients to tissues and organs [87]. Lymphoid aggregate are accumulations of inflammatory cells. They perform an important function to prevent infections in the

gastrointestinal tract.

Dysplasia is abnormal structural changes in the epithelial cells and this is seen in precursors to cancer and invasive cancer. Figure 2.4b shows the difference between normal tissue and dysplasia. The severity of dysplasia is graded between high-grade and low-grade dysplasia. In the case of high-grade dysplasia, there is a higher risk of developing cancer. In cases where cancer has appeared and spread to surrounding tissue, dysplastic epithelium will also exist in the invasive cancer [29].



(a) Lymphoid aggregate and blood vessels with surrounding connective tissue.

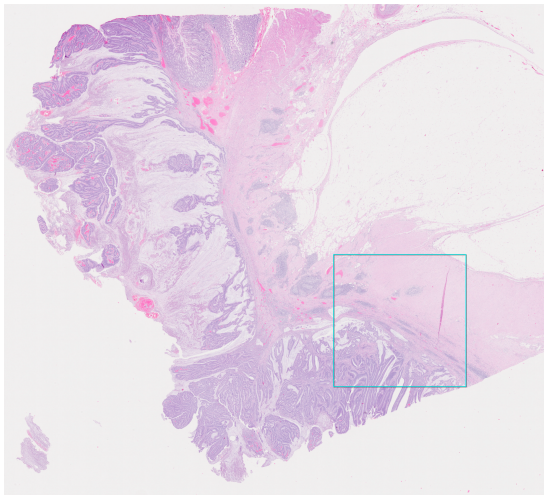


(b) Comparison of normal tissue and dysplasia. The dashed line is where the dysplasia starts to form.

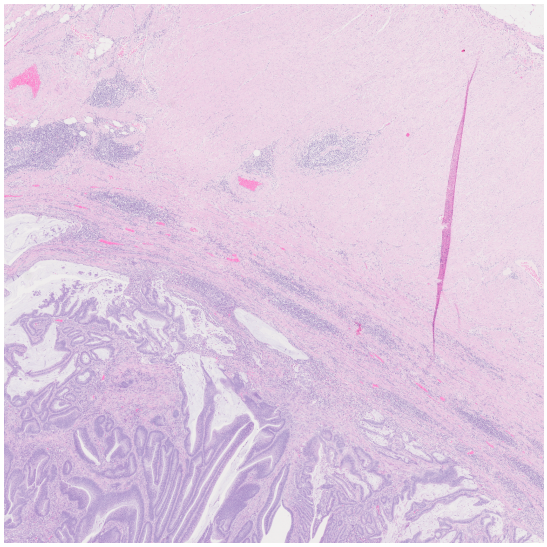
Figure 2.4: Sections from a scanned histological whole slide image stained with Hematoxylin and Eosin (HE).

A potential tumour will arise in the epithelial layer of the mucosa and over time develop and spread deeper into the other tissue types in the colon and rectum. Identifying which tissue layer the tumour has reached will help determine the aggressiveness of the tumour, which in turn may have an impact on treatment strategies and prognosis.

Tumours that have grown through the tissue layers and into the muscularis propria will require more extensive surgery to remove the tumour and thus affect the patient's prognosis. The area where the cancer infiltrates surrounding tissue is known as the invasive margin (IM). Which tissue layer this outer edge of a tumour is located in is of great importance. Cancer that has spread to the muscle tissue in the intestinal wall, may indicate a more advanced form of cancer that has the potential to spread to other parts of the body. Figure 2.5 shows a sample of cancer spread to the muscle tissue. As the muscle tissue is responsible for the peristaltic movements, these will be disturbed and lead to obstruction in the intestine.



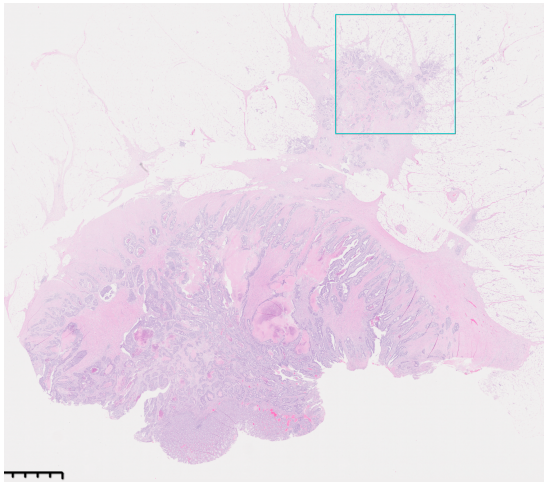
(a) Overview of the WSI. The cyan square shows where the invasive margin borders to muscularis propria.



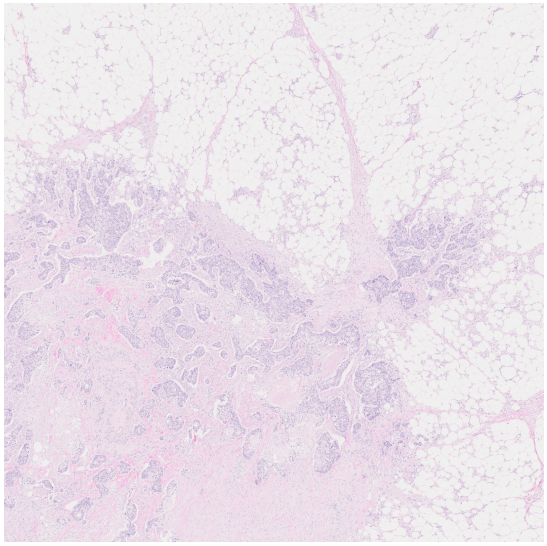
(b) A close-up of the invasive margin marked in the cyan square.

Figure 2.5: Whole slide image (WSI) where the invasive margin borders to muscularis propria.

The same applies if the cancer has spread to the subserosal fatty layer, see figure 2.6. The fatty layer is located in the outer serosa layer of the GI tract. Spread of cancer to the serosa is associated with reduced survival and may influence further treatment decisions, like the need for adjuvant treatment such as chemotherapy [23].



(a) Overview of the WSI. The cyan square shows where the invasive margin borders to the subserosal fatty tissue.



(b) A close-up of the invasive margin marked in the cyan square.

Figure 2.6: Whole slide image (WSI) where the invasive margin borders to subserosa, specifically the subserosal fatty tissue.

Knowledge of blood vessels in the intestinal wall can be important to assess the potential for the spread of the tumour to other parts of the body through the bloodstream. It can also be decisive for planning surgical interventions and choosing treatment methods. In addition, the cancer cells can invade the blood and lymph vessels and spread to other parts of the body, such as the liver and lungs. This can make treatment more challenging and reduce the prognosis for survival [92].

In general, cancers that spread to deeper layers of the GI-tract, and especially to lymph nodes, are associated with a poorer prognosis that often need more treatment like adjuvant chemotherapy. It is therefore crucial to identify the different tissue types when diagnosing tumours. Then, a pathologists can assess the spread to nearby lymph nodes and how deep the cancer has grown in the intestinal wall. This is determinant for the prognosis of the patient. Correct identification of tissue types is important for this classification.

2.3 Diagnosis

Diagnosing colorectal cancer involves a combination of clinical examinations. First and foremost, cancer is detected when patients present with symptoms such as blood in stool, change in stool, pain, etc. Then radiological examinations and colonoscopy are carried out to investigate further. Recently, screening has also been started, which aims to detect the disease at an earlier stage [84].

If cancer is suspected, a colonoscopy is carried out. That is when a flexible tubular device with a camera on the end is inserted into the intestine to inspect the intestinal walls. Here, the doctor looks for discoloration, unevenness, tumours or polyps as signs of cancer [51]. During a colonoscopy, tissue samples must be taken of suspected tumours. A small piece of tissue is then removed from the tumour which is then evaluated by a pathologist. During microscopic examinations of the tissue samples, the pathologist looks for characteristic features that indicate cancer. It can be cells that show abnormal shape and structure and cells that divide uncontrollably and rapidly, i.e. dysplasia. If dysplasia is detected in the sample, the pathologist checks for infiltration and spread to surrounding tissue, blood vessels or lymphatic vessels.

Blood tests can also give indications of cancer. In the presence of cancer, a protein called CEA (carcinoembryonic antigen) will give high levels. If CEA shows significantly high values, it may be an indication of widespread disease with a poor prognosis. CEA is also used as a tumour marker after surgery to see that the levels fall back to normal. If they do not, this indicates a recurrence of the cancer. Image examinations such as CT (computed tomography) scan, MRI (magnetic resonance imaging) and PET (positron emission tomography) scan give the doctor an overview of the spread and size of the tumour [51].

2.3.1 Staging

What stage the disease has reached at the point of diagnosis is vital for both prognosis and treatment. The diagnosis is based on preliminary examinations, findings during surgery and the pathologist's evaluation of the surgical specimen. Today, colorectal cancer is graded by a combination of Tumour-Node-Metastasis (TNM) classification and staging[85].

TNM classification describes the growth of the tumour through the intestinal wall (T) and whether there is spread to lymph nodes (N) or distant metastases (M) [46]. Table 2.1 shows how the TNM system applies to colorectal cancer.

Description	**T stage**
Tis	Tumour is located in the lamina propria and has not invaded through the muscularis mucosa.
T1	Tumour has grown into submucosa.
T2	Tumour has grown into muscularis propria.
T3	Tumour has grown through muscularis propria to subserosa.
T4	Tumour has grown directly through serosa or into other organs or structures.
Description	**N stage**
N0	No cancer in regional lymph nodes.
N1–2	Cancer in one or more nearby lymph nodes.
Description	**M stage**
M0	Cancer has not spread to other organs.
M1a-c	Cancer has spread to other organ(s) or abdominal cavity.

Table 2.1: Tumour-Node-Metastasis (TNM) staging for colorectal cancer. Adapted from American Joint Committee on Cancer (AJCC) cancer staging manual [1].

Staging is based on the TNM classifiaction system and it describes the degree of malignancy. The more severe the stage of the disease, the worse the prognosis [34].

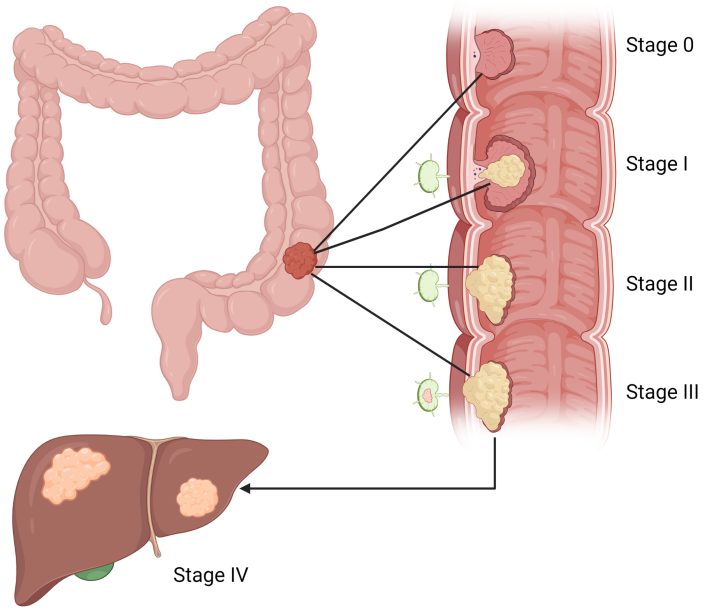


Figure 2.7: A cross-section of the GI-tract showing the stages of colorectal cancer, and how the tumour infiltrates the nearby tissue. Retrieved from Dordi Lea (2022). Use of quantitative pathology to improve grading and predict prognosis in tumours of the gastrointestinal tract. Created with BioRender.com [56].

There are four main stages of colorectal cancer, see figure 2.7. Stage 0 means that the primary tumour is premalignant lesion. Stage I shows that the tumour has spread to the first layer of the colon or rectum, i.e spread to muscularis propria. At stage II, the tumour has grown through the muscularis propria, but has not spread to other organs or nearby lymph nodes. Furthermore, stage III means that the cancer has spread to nearby lymph nodes, but not to other organs. By the time the cancer reaches stage IV it has spread to other organs, such as the liver or lungs [51].

After the cancer has been diagnosed and a stage has been given, the patients are then stratified into risk groups and tailored a suitable treatment and follow-up.

2.3.2 Treatment

Knowledge about the treatment of colorectal cancer is mainly based on cohort studies of varying quality. The Norwegian Directorate of Health has created guidelines for treatment in the various stages. Small tumours that are adenomas, i.e. precursors, or T1 can sometimes be removed locally through polypectomy, where the upper part of the mucous membrane is removed. For the majority of stages, surgery will be considered the most suitable option for treatment. For rectal cancer you can receive radiation before or after surgery and chemotherapy, but for colon cancer it is usually adjuvant therapy after surgery. In stage IV, life-prolonging treatment such as chemotherapy, immunotherapy, molecular therapy or surgery for metastases may be appropriate [30].

2.4 Digital pathology

Since the start of modern pathology, when pathologists carried out microscopic examinations, there has been a noticeable development. The concept of digital pathology has been around for decades, however it is in the last 20 years that the concept has really expanded. About two decades ago, in

the late 1990s, whole slide imaging was introduced [61]. Whole slide imaging is a process that converts tissue samples into a virtual slide. This means that pathologists can now observe color differences and structures digitally on a computer screen instead of under a microscope. This new application makes the process both faster and gives the ability to store large quantities of samples in searchable databases [56]. Pathologists can then share digital images for research and comparison of samples leading to improvement of accuracy of techniques and diagnosis [37].

2.4.1 Whole slide imaging

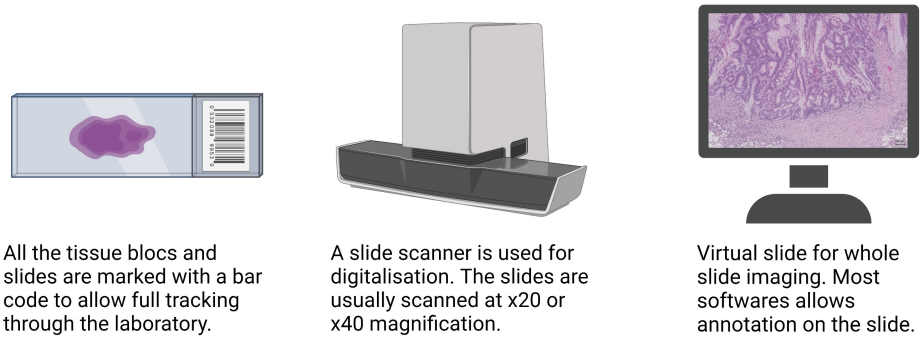


Figure 2.8: Whole slide imaging in a digital workflow in a pathology department. Image by Dordi Lea. Created with BioRender.com.

Figure 2.8 illustrates the workflow for whole slide imaging, often referred to as virtual microscopy [68]. Whole slide imaging covers the entire process from scanning a tissue sample until it is converted into a high-resolution digital image, known as a whole slide image (WSI). It is an extensive process that requires both specialized equipment and IT infrastructure. The equipment must meet the requirement for high-resolution images to ensure that the smallest details are captured. To achieve a sufficient level of detail, the virtual slide should have a minimum optical magnification of x20. Most scanners provide an optical magnification of 40x which leads to higher diagnostic accuracy but also requires more storage capacity.[56]

2.4.2 Use of deep learning in pathology

'Tidsskrift for Den norske legeforening', often referred to as 'Tidsskriftet', has published an article in which they state that the pathology of the future is digital. 'Tidsskriftet' address the problem that there is a critical shortage of pathologists in Norway, but in reality it is a global issue. The increase in the development of personalized medicine has led to more tests and thus increased work for pathologists [20]. In addition, the population is increasing and people live longer, hence more patients. Therefore, the use of digital tools to assist the pathologists is extremely useful.

The digitization of pathological work offers not only resource and time-saving benefits for pathologists, but it can also be used to enhance diagnostic accuracy. The use of artificial intelligence provides an objective and consistent assessment, while human visual perception has a natural limitation when it comes to interpreting data. Therefore, quantitative assessments will lead to a more reliable diagnosis as it removes the variation in uncertain cases. [56]

Due to the shortage of pathologists, it has become an area of focus to develop various AI algorithms that can facilitate the workflow. The department of pathology at St. Olav's hospital in Trondheim is in the process of using artificial intelligence to digitize cytology. The new cytology tool "GeniusTM digital cytologi diagnostisk system" (GeniusTM) analyzes and evaluates cytological preparations. Time saving is an important factor and compared to the manual method, time consumption will be reduced by 75% when using GeniusTM [27].

Studies have shown that artificial intelligence can make assessments better than pathologists. In 2017 an article titled "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer" was published in Journal of the American Medical Association, abbreviated to JAMA. In this study a deep learning model achieved better diagnostic performance than a panel of 11 pathologists. The algorithm achieved an area under the curve (AUC) of 0.994 compared to and average AUC of 0.884 for the pathologists [22].

Chapter 3

Technical background

This chapter will give a technical explanation of how neural networks work, mainly convolutional neural networks. The chapter focuses on the techniques used in this study as the overall topic is far too extensive.

3.1 Artificial intelligence

Technologies that are developed with the intention of performing tasks that require human intelligence is by definition artificial intelligence, abbreviated to AI. The goal of AI is to train an intelligent agent that can solve a specific task based on the environment presented.[94]

Machine learning is a specialization within artificial intelligence that uses statistical methods to find patterns in large amounts of data. Machine learning can be divided into three main categories, supervised, unsupervised and reinforcement learning [76]. More about supervised learning can be found in subchapter 3.7. All the learning techniques can be trained for various tasks such as regression, classification and segmentation. What differentiates them is their use of labels [91].

There are three factors that must be considered to create a successful machine learning system; a large set of training data, an algorithm that learns from the data and computational power to process the data. [94] For the latter, we have Graphics Processing Units (GPUs) which are designed to perform parallel processes. Because GPUs contain an additional amount of computational power, they can deliver remarkable acceleration in workloads like image recognition. Many of today's deep learning systems rely on GPUs working with CPU. There are several companies that manufacture GPUs, such as NVIDIA Corporation with its best-known GPU solutions GeForce, Quadro and Tesla [35].

3.2 Neural network

Neural networks are based on the structure of the human brain and its ability to solve problems. The brain is made up of several basic building blocks known as neurons. Neurons are nerve cells that function as information messengers in that they exchange information by forming connections between each other [19]. Similarly, neural networks consist of artificial nodes that mimic the biological neurons in the brain [62]. Figure 3.1 shows the structure of an artificial neuron.

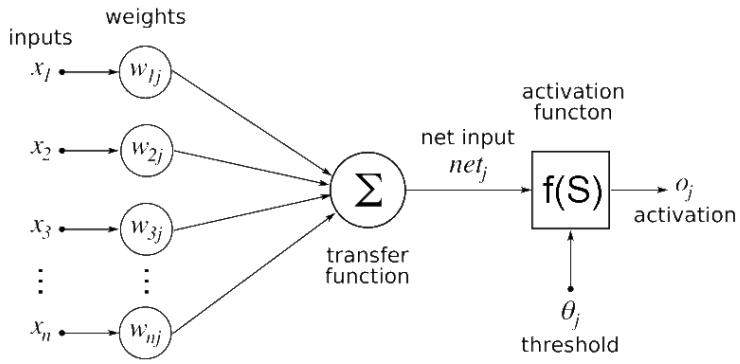


Figure 3.1: An illustration of an artificial neuron. Image by Geetika saini, licensed under the Creative Commons BY-SA 4.0 license [80].

Figure 3.1 illustrates that every jth artificial neuron includes several inputs x_i and weights w_{nj}, where $i = 1, ..., n$. Each input signal is combined with the connection weight by multiplication and then summed together [94].

Furthermore, the weighted sum net_j is fed through a non-linear function, known as an activation function. The purpose of the activation function is to check whether the value has reached a given threshold, and if so the data can be passed on to the next neuron to do the same. If the value is lower than the threshold, no data will be passed on [39].

3.3 Deep neural network

Generally, a neural network consists of an input and output signal and one hidden layer. A neural network with multiple hidden layers introduces deep neural networks, where the number of hidden layers increases with complexity. The more layers, the deeper the architecture. The name hidden layers comes from the fact that the layers are hidden in the weighted sum signals that are transmitted between the neurons [63].

The weights between the neurons are adjusted during the training of the neural network. They are optimized to learn how to solve a specific problem. At the start of a training session, all the weights and thresholds are initially set to random values. This means that the output value will be far from the expected value in the beginning of training.

A loss function is used to measure the performance of a model. The function returns the difference between the predicted output signal and the actual output signal. The loss function can be a simple Mean Squared Error (MSE) Loss or a more complex Cross Entropy Loss. Based on the value from the loss function, the model knows how much to adjust the weights to minimize the deviation [69].

A back propagation algorithm is used in order to minimize the loss function. As the name suggest, it goes backwards through the network to adjust the weights and biases in the different layers with the use of the chain rule. The chain rule computes the gradient of the loss function, which in turn shows how much the parameters needs to be adjusted [49].

3.4 Convolutional neural network

Convolutional neural network (CNN) is a specialized form of neural network favored for image recognition and classification. In general, a CNN consists of three types of layers; a convolutional layer, a pooling layer, and a fully connected layer [60].

Figure 3.2 shows the structure of a general CNN. It begins with an input image of size 8@128x128. Where 8 is the number of color channels and 128x128 are the dimensions of the image. After the input has passed through a maxpooling layer, the dimension is reduced to 64x64 but the color channels remain the same. Then comes the convolution layer, which consists of several filters. These filters are then convolved with the pixels in the image, resulting in more channels and a reduced dimension. The next

maxpooling layer outputs 24@16x16. Finally, the dense layer transforms the two-dimensional signal into a one-dimensional vector of size 1x256. Furthermore, it undergoes another dense layer to reduce the length of the vector. It is the length of this vector that refers to the number of possible classes that the CNN is trained to recognize, in case of figure 3.2 this is 12 different classes.

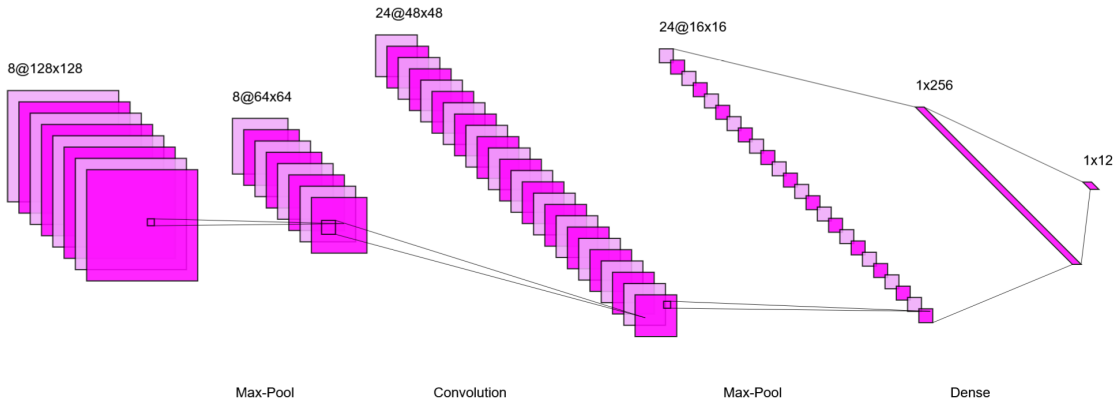


Figure 3.2: Illustration of a typical CNN structure. Created with NN-SVG, licensed under the MIT License [57].

The convolution layer is the core building block of CNN. It passes a filter, also known as a kernel, over the image and performs a dot product between the filter and the overlapping part [60]. This generates a feature map that represents a specific characteristic. A problem with the feature map is that it is sensitive to the location of the features in the image. One way to address this sensitivity is through the use of pooling layers [59].

The pooling layer reduces the dimensionality by summing the feature maps within a patch and returning a smaller feature map with fewer dimensions. This also provides translational variance which is highly necessary in image recognition as objects can appear in different parts of an image with different orientations [44]. It is important to note that downsampling will limit the identification of features. Therefore, too many pooling layers will reduce the model's size to such a large extent that it leads to loss of information and details. There are two main methods of pooling layers, average- and maximum pooling. Average, as the name implies, takes the average of each feature map for each patch. While maximum pooling calculates the maximum value for each patch, causing it to extract the most salient features [59].

The last section of the CNN is what makes up the classification and it consists of a combination of flatten and fully connected layers. First, the input goes through the flattening layer to convert the multi-dimensional feature map into a one-dimensional feature map. This is because the fully connected layer only accepts one-dimensional arrays. [33]. The fully connected layer will predict the class of the image based on the features that have been extracted [26].

3.4.1 VGG16 convolutional neural network

VGG16 stands for "Visual Geometry Group 16" and is a network model that is particularly suitable for object detection and image classification. It is a convolutional neural network with a deep architecture. It was developed by researchers from the Visual Geometry Group (VGG) at the University of Oxford. The model was first introduced in a research paper by Karen Simonyan and Andrew Zisserman where the purpose was to explore the depth of neural networks and their ability to learn complex representations of images [42].

The VGG16 model is trained on the ImageNet dataset. The dataset includes over 15 million labeled high-resolution images belonging to approximately 22,000 categories. The ImageNet images have variable resolution, therefore all images were downsampled to a fixed size of 224x224 pixels before training the model. The model was trained for a week and used NVIDIA Titan Black GPUs to achieve its results [28].

VGG16 consists of 16 layers of which 13 of the layers are convolutional layers and 3 are fully connected layers. The convolution layers use a 3x3 filter size with stride 1 and padding to maintain the size of the image. A max-pooling of size 2x2 with stride of 2 is used to reduce the dimensionality. After each convolution layer, there follows a ReLU activation, which helps introduce non-linearity into the model [55]. Finally, it has two fully connected layers followed by a softmax as output [90].

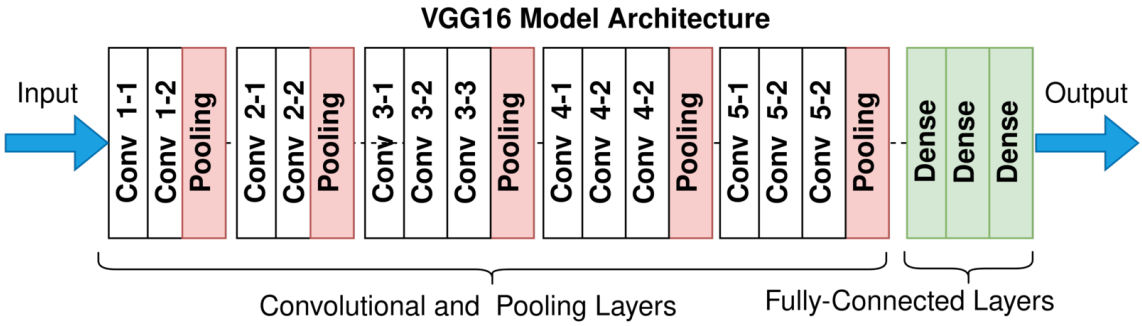


Figure 3.3: Structure of VGG16 model. Image by Gorlapraveen, licensed under the Creative Commons BY-SA 4.0 license [25].

Although Figure 3.3 may appear to have 21 layers, it is still referred to as VGG16 because there are 16 layers with weights being trained in the model. Note that VGG16 was trained on images of size 224x244 pixels [10]. Therefore, it can be useful to use the same size when training with VGG16. If not, the results may be biased.

What makes the VGG16 a good model is that, even though it has a deep architecture, the architecture is still relatively simple compared to other well-known models. After VGG16, there have been created models with even more layers, such as RestNet and Inception V3. These also give good, if not better, results on image recognition, but require more computing power and memory. Therefore, VGG16 is recommended if you do not have sufficient time, computing power and resources [78].

3.5 Regularization techniques

In machine learning, various regularization techniques are used to prevent overfitting and improve the generalization performance of a model [38]. This section will review some of the regularization techniques used in convolutional neural networks.

3.5.1 Early stopping

It is often difficult to choose the right number of training epochs. Too many epochs can lead to overfitting, while too few epochs can result in an underfit model. Early stopping is a method that stops training when the model stops improving on a validation set [16]. By choosing a number of epochs that will act as a patient, this will prevent the risk of overfitting as the model avoids becoming too specific to the training data.

3.5.2 Dropout layers

When large neural neworks are trained on small datasets, there is a risk of overfitting. This is because the model is trained on statistical noise that occurs during training [14]. Dropout layers randomly deactivates a certain number of nodes during training. They have no weights associated with them so

the nodes in the dropout layer is randomly set to zero during training. This forces the network to create a more robust model that does not depend on specific nodes, in addition reduce the sensitivity to noise [95].

3.5.3 Batch normalization

Training a neural network with many hidden layers can be a difficult and time-consuming process. A preferred method of accelerating training is the use of batch normalization. Batch normalization is a technique that calculates both mean and variance before normalizing the activation in each hidden layer [31].

Studies show that batch normalization makes the optimization landscape significantly smoother, which in turn introduces a more predictive and stable behavior of the gradients [81]. It turns out that this smoothing provides the opportunity for training with a greater learning rate, which means that the loss function will reach its minimum much faster, hence speeds up training [8]. In addition, the variation of batch normalization statistics introduces a random disturbance and helps to reduce overfitting. In that regard, batch normalization induces a certain regularization. However, it is still used together with dropout layers for better results [4].

3.5.4 Data augmentation

Data augmentation is a useful method for increasing the number of images in the dataset. It is a technique that applies transformations to original images and as a result you get several different transformed versions of the same image [5]. Transformation parameters that can be adjusted are rotation, shift, brightness, shear, zoom, horizontal or vertical flip [83]. As shown in Figure 3.4 and 3.5, a single image of the tumour center can be augmented by these transformations. This will prevent overfitting and cause the model to generalize better on unseen images [83].

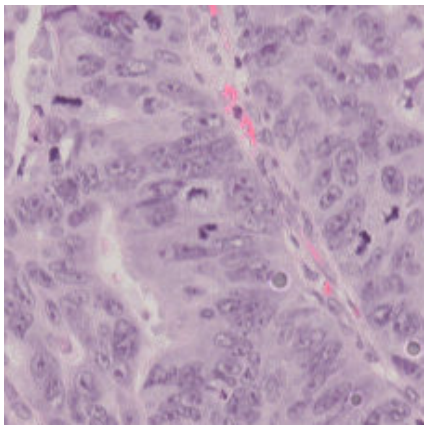


Figure 3.4: Original image of tumour.

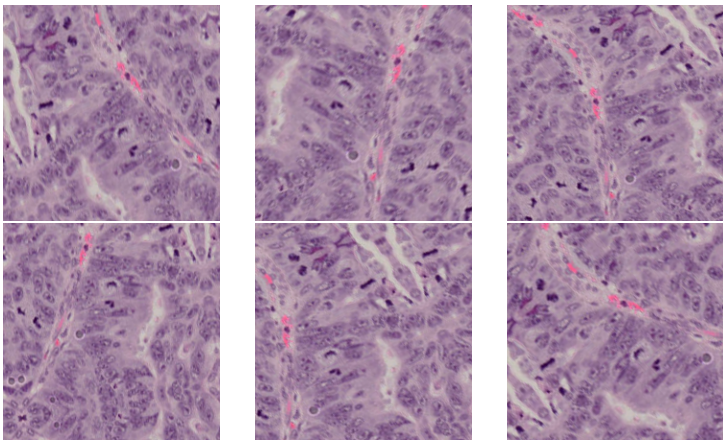


Figure 3.5: Augmented images of the tumour.

The TensorFlow library includes a tool called ImageDataGenerator that expands the dataset by making small changes to the existing dataset [89]. Training with ImageDataGenerator also comes with some drawbacks. Often the original dataset will contain biases, and when augmented, this data will also get biases. In addition, a simple training set will also lead to a form of bias as augmentation will limit the ability to recognize unknown data outside the training set. It is also important to note that increasing the number of images in the training set will lead to increased processing time for the model [2].

If variation in the training data is not desired, it is possible to only rescale with ImageDataGenerator. Rescaling means that the pixel values are divided by 255 to normalize the image data and bring the pixel values within the range [0,1] [15].

3.6 Evaluation metrics

Up until now, various techniques and layers that can be adjusted in a model have been reviewed. A challenge with machine learning is that there is no conclusion on what the optimal parameters are. Usually, models have to be trained on different parameter settings in order to then be compared with each other. Different types of metrics are used to determine how well a model has performed. A selection of different metrics will be presented here.[94]

3.6.1 Confusion matrix

A confusion matrix is an nxn matrix, where n is the number of categories in the output. Each row represents the true classification of the data and each column refers to the predicted classification. The easiest way to determine how well a model has performed is by looking along the diagonal of the confusion matrix. A reliable model will contain high values, and often a more intense coloring in the cell, along the diagonal. See figure 3.6. Confusion matrix can also easily show where the model struggles to predict by studying the highest values that are not along the diagonal [11].

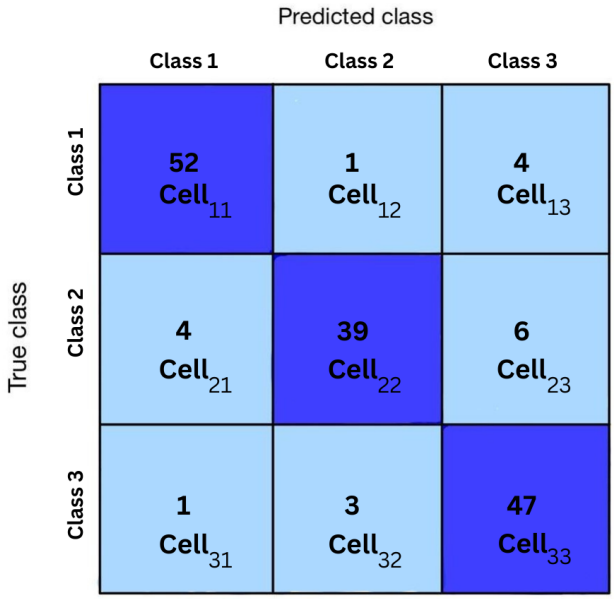


Figure 3.6: Illustration of an ideal confusion matrix with $n = 3$ classes. The diagonal presents the correct predictions.

Positive and negative classes

A confusion matrix is based on positive and negative classes in a model. In binary classification, where $n = 2$, there are four possible outcomes for predictions; True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [24]. Together, the four outcomes can form a binary version of a confusion matrix, see figure 3.7.

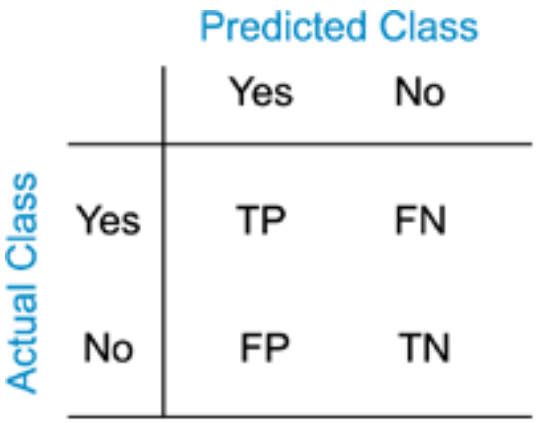


Figure 3.7: Description of TP, TN, FP and FN. Image by Oritnk, licensed under the Creative Commons BY-SA 4.0 license [67].

Similar to a confusion matrix, high values for TP and TN will indicate good performance. However, in a multi class setting, where $n \geq 3$ the values for TP, TN, FP and FN must be computed for each class [94].

True positive (TP) refers to the number of correct predictions. The TP for a class c is equal to the value in the cell at row c and column c in the confusion matrix, expressed in equation 3.1.

$$TP_c = cell_{c,c} \tag{3.1}$$

False positive (FP) is when the model wrongly predicts a negative class as a positive class. Equation 3.2 calculates the FP value for a class c.

$$FP_c = (\sum_{P=1}^{n} cell_{P,c}) - TP_c \tag{3.2}$$

False negative (FN) is when the model wrongly predicts a positive class as a negative class. Equation 3.3 calculates the FN value for a class c.

$$FN_c = (\sum_{Q=1}^{n} cell_{Q,c}) - TP_c \tag{3.3}$$

True negative (TN) is the samples that were correctly predicted as negative. Equation 3.4 calculates the TN value for a class c.

$$TN_c = (\sum_{P=1}^{n}\sum_{Q=1}^{n} cell_{P,Q}) - TP_c - FN_c - FP_c \tag{3.4}$$

3.6.2 Accuracy

Accuracy is an overall measure of the model's performance. It shows the ratio between the number of correct predictions and the total number of predictions [40].

$$\text{Accuracy} = \frac{\sum_{P=1}^{n} TP_P}{\text{total population}} \tag{3.5}$$

Where, the total population is the sum of all elements in the confusion matrix [94]. Accuracy makes it easy to compare models with each other. It is important to note that if there is an imbalance in the classes, the evaluation will not be very reliable. [94]

3.6.3 Classification report

A classification report is useful for checking the model's performance. The report measures the quality of the predictions of the trained classification model. An overview of typical metrics that are used in a classification report will be described here [54].

Precision

Precision is defined as the ratio of true positives to the sum of true and false positives. It is also known as Positive Predictive Value, abbreviated as PPV [45].

$$PPV_c = \frac{TP_c}{TP_c + FP_c} \tag{3.6}$$

Equation 3.6 shows the expression for precision for each class c. PPV evaluates the models ability to correctly identify positive cases among the predicted positive cases. Ideally PPV should be close to 1 [96].

Recall

Recall, also known as sensitivity or True Positive Rate (TPR), is defined as the proportion of true positives compared to the sum of true positives and false negatives [59]. Expression for recall for each class c is:

$$TPR_c = \frac{TP_c}{TP_c + FN_c} \tag{3.7}$$

Recall measures how suitable the model is to recognize a positive class. Best performance implies a TPR value close to 1 [96].

F1-score

F1-score shows the balance between precision and recall. It uses the harmonic mean to avoid bias towards extreme values and give more weight to low values [41]. Equation 3.8 shows the expression for f1-score for each class c.

$$\text{F1-score}_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \tag{3.8}$$

With perfect precision and recall, f1-score will have a value of 1 [45].

Support

Support shows the number of actual occurrences of each class in the specified dataset. It shows how well each class is represented by providing information on how many samples belong to each class. Imbalance in the training data may indicate a need to adjust the class distribution [47].

3.7 Learning techniques

When training a model, there are several techniques to choose from. The two most common methods are supervised and unsupervised learning. They differ in the way the models are trained. Supervised means that the input data is labeled with the correct output, while unsupervised trains without lables and will learn by identifying patterns. The former is recommended if the task is to classify data or make predictions. Although supervised machine learning is more resource-intensive due to the need for labeled data, it provides better control and an overview of the training [9].

Supervised learning involves a series of functions that maps an input to an output based on a series of input-output pairs. An input set X contains a collection of samples x and a corresponding set Y will

contain a label y for each sample in the input set. The purpose is to find a connection between the input data and the corresponding labels [94].

This is achieved with a mapping function $y = f(x)$. The aim is to train a model that approximates this function by giving a predicted lable $\hat{y}$ to the input x. This yields the function:

$$\hat{y} = f_{model}(x) \tag{3.9}$$

Where f_{model} is the function the model uses to make the predictions. After a prediction is made it is compared to the correct label. Then a loss function is used to measure the distance between the prediction and the label in order to minimize the loss.

3.8 Data distribution

A machine learning algorithm learns from the dataset presented. Therefore, it is important to understand how the data can be distributed to give the model the best performance and reliability [94]. Typically, the dataset is divided into three parts; train, test and validation. The training set optimizes the parameters of the model during training, the validation dataset evaluates the model during training and the test set evaluates the final performance [3].

3.8.1 Train_test_split

The Scikit-Learn (or sklearn) package provides tools to perform common machine learning operations. One such tool is the train_test_split function. The function splits the dataset into separate sets for training and testing. Resulting in the possibility to distinguish between what dataset the model is to learn from and what it is to be tested on. This makes the process more transparent [82].

The most common way to use train_test_split is by specifying a percentage of the dataset to be assigned to the test set. One should also note that there is a random seed in the function that ensures variation each time the dataset is split. By setting this random state parameter to 42, you ensure that the dataset is the same every time the code runs [21].

3.8.2 Preventing cross-contamination

A common problem in deep learning is cross-contamination, which refers to the transfer of unwanted data from one source to another. Let's say you have one set of data to train your model and another set to test its accuracy. If some of the test data is included in the training data, the model can appear better than it actually is. This is because the model is tested on known data so it knows the answer in advance. It is therefore necessary to ensure that data used for training is not a part of the test dataset, as this can make the model less reliable [17].

Chapter 4

Data material

This chapter will review the data used in this study and the work that was done before the digital whole slide images were created. To make it easier to handle, the term "data material" refers to all available data, while "dataset" refers to parts of the data material that have been extracted into smaller subsets.

4.1 Hamamatsu slide scanner

Stavanger University Hospital (SUS) are using slide scanners to digitize tissue samples. The scanner that was utilized in this study was the slide scanner from Hamamatsu Photonics. Hamamatsu is a Japanese company that produces high-resolution slide scanners, where the scanner at SUS has a x40 optical lense [71]. The digital scanner takes in a tissue sample, scans the sample and turns it into a virtual slide that is an copy of the original tissue sample with x40 magnification. This virtual slide can then be used for whole slide imaging. [56]

4.1.1 NDP.View2

Hamamatsu Photonics also has a department that specializes in optical solutions. Within this area, they have developed NDP.view2, an image viewing software. The software is used to visualize and analyze microscopic images such as whole slide images [70]. NDP.View2 offers various functions, including an annotation tool. The annotation tool allows the user to view, create and modify annotations on images. Several annotations can be made on a single image, where each annotation can have a different color which in turn makes it easy to distinguish between the different types of annotations. Exporting annotations can be done in two formats; XML (.ndpa) or Excel (.csv). Figure 4.1, shows a fragment of an XML file where the metadata of the annotations is preserved.

```xml
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<annotations>
    <ndpviewstate id="2">
        <title/>
        <details/>
        <coordformat>nanometers</coordformat>
        <lens>2.244354</lens>
        <x>18072508</x>
        <y>-3856627</y>
        <z>0</z>
        <showtitle>1</showtitle>
        <showhistogram>0</showhistogram>
        <showlineprofile>0</showlineprofile>
        <annotation type="freehand" displayname="AnnotateFreehand" color="#ff0000">
            <measuretype>0</measuretype>
            <closed>1</closed>
            <pointlist>
                <point>
                    <x>15302468</x>
                    <y>-1881502</y>
                </point>
```

Figure 4.1: Metadata regarding annotations from the region of interest (ROI).

The metadata includes, among other things, the annotation's id and colour, the type of annotation and a pointlist with x and y coordinates of the annotation.

4.1.2 Histological whole slide images

The data were collected from the ACROBATICC biobank at the Department of Pathology, Stavanger University Hospital (SUS). The data material includes scanned histology images from colorectal patients. The material includes patients with stage T1 to T3 at the time of surgery, i.e. without distant metastases.

In total, the data material consists of 174 surgical sections stained with hematoxylin and eosin, also known as HE sections, which are further scanned and converted to digital whole slide images. To maintain the anonymity of each patient, each image is named 'TEST' followed by a study number identification (ID).

Magnification

In order to examine the stained sample, the information from each WSI is extracted from different resolutions. This is because different magnifications of a WSI will provide different information. A high magnification will reveal features like cell size and shape, while a lower magnification will give more context information from the surrounding tissue [94].

The scanner at SUS has an optical lens of x40. This means that the objective in the scanner provides an optical magnification of 40 times, i.e. the WSI appears 40 times larger than the sample. However, the total magnification may differ from the optical magnification depending on whether an eyepiece is used. The eyepiece contains an ocular lens, often with a 10x magnification power. Therefore, the total magnification will be the product of the ocular lens and the optical lens [94].

$$\text{Total magnification} = \text{Optical lense magnification} \cdot \text{Ocular lense magnification} \tag{4.1}$$

Multilevel gigapixel image

Even though the tissue samples are only a few millimeters, a corresponding whole slide image (WSI) with 40x optical lens would be categorized as a gigapixel image. Often the size of a WSI exceeds 100,000 x 100,000 pixels, making it challenging for computers to process. In order to handle these high-resolution images, it is necessary to divide the image into several sub-images, where each image can be stored separately and thus reduce the load on the computer's resources [94].

WSIs already consist of such a structure, known as a pyramid structure. Instead of storing a single image with an extremely high resolution, the WSI is divided into several smaller images, called tiles. Together, these tiles create a hierarchical pattern with different levels of resolution. See figure 4.2.

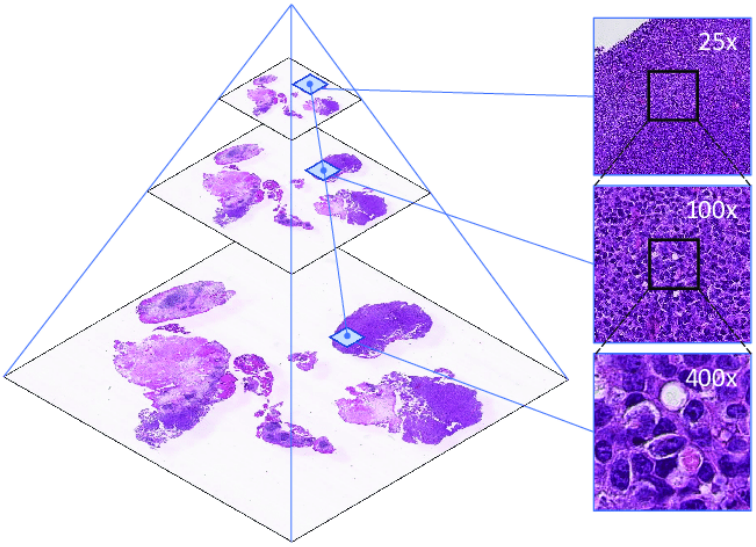


Figure 4.2: Illustrates how the WSI are stored in a pyramid format. Retrieved from Wetteland, R. (2021). Automated Grading of Bladder Cancer using Deep Learning [94]. Licensed under Creative Commons Attribution 4.0 International [36].

The image shows the total magnification view of each tile, i.e. including ocular lens magnification of 10x. See equation 4.1. At the top of the pyramid is a low-resolution image that provides an overview. The further down the pyramid, the more details are revealed.

Chapter 5

Method

This section will deal with the work behind processing the data material into a dataset that can be used in a classification model. Then the procedure behind the construction of the models will be presented as well as the justification for the choices that were made.

The structure and algorithm for the pipeline was inspired by Rune Wetteland's paper "A Multiscale Approach for Whole-Slide Image Segmentation of five Tissue Classes in Urothelial Carcinoma Slides" [93], while the content was based on Dr. Sreenivas Bhattiprolu's work on the use of OpenSlide for whole slide images [7].

5.1 Pre-processing

Due to the size of the data and storage requirements, the whole slide images must be pre-processed before they can be used to train the model. This sub-chapter will present the work behind the conversion from data material to the finished dataset. Appendix A provides the README file for the pre-processing source code.

5.1.1 Annotation

In this study, the annotation for each tissue to be classified was performed using the NDP.View2 software. The annotation involved marking regions of interest (ROI) on each WSI to highlight specific areas of interest. Annotations were made accordingly based on the type of classifications.

Among the 174 available WSIs, only 65 WSIs were annotated. After completing the annotations, they are exported in the form of an XML file (.ndpa), see example shown in figure 4.1. The export makes it easier to save the annotation data until they are used for further processing.

The annotation coordinate system

It turns out that the coordinate system of NDP.View2 is not of the same standard as other software, where the top-left corner is the origin. The coordinate system of NDP.View2 is based on the center of the slide scanner, which in turn depends on the location of the tissue sample. Therefore, the image center will vary in relation to the center of the slide scanner, also known as slide centre.

I so happens that OpenSlide has a specific function that retrieves the offset value from the slide center to the origin in both horizontal and vertical directions. This makes it possible to calculate the position of the top left corner of the WSI and specify this as the new origin.

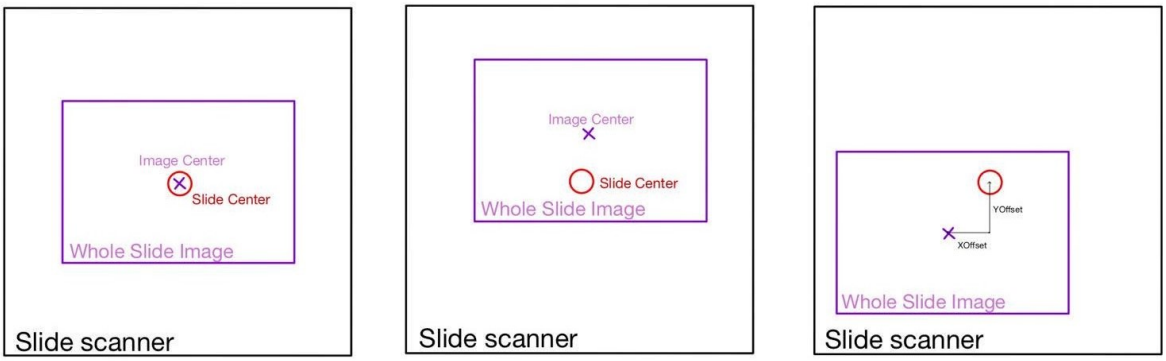


Figure 5.1: Illustrates how the image center changes with respect to the slide center.

Figure 5.1 shows how the center of the slide can vary in relation to where the sample is placed on the scanner. The annotations provided in the metadata are based on the slide centre. Therefore, to be able to take advantage of these for further processing, equation 5.1 and 5.2 had to be derived.

First, the offset in horizontal and vertical direction was obtained from Open slide's function [65]. This gives the distance from the center of the entire slide to the center of the main image in nanometers. In order to end up in the top-left corner you need to subtract this offset from half the dimension of the main image.

For example, you have a WSI with dimension ixj in pixels, and extracted OpenSlideOffsetX and OpenSlideOffsetY form the Open slide library. The resolution is 220 nm/pixel. The expression then becomes:

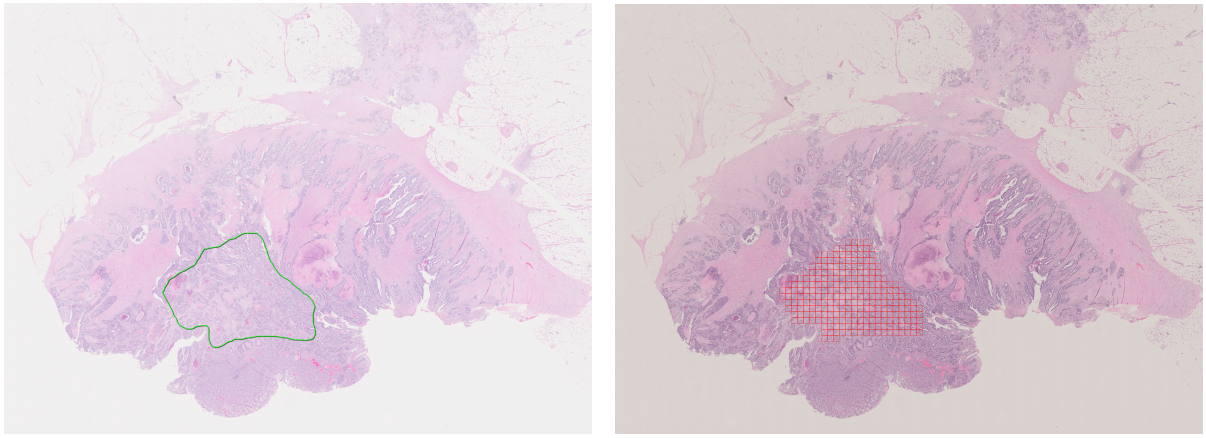
$$\text{XOffset} = i\,\text{pixels} \cdot \frac{220}{2}\frac{\text{nm}}{\text{pixel}} - \text{OpenSlideOffsetX} \tag{5.1}$$

$$\text{YOffset} = j\,\text{pixels} \cdot \frac{220}{2}\frac{\text{nm}}{\text{pixel}} - \text{OpenSlideOffsetY} \tag{5.2}$$

Furthermore, the offsets in x and y direction will be added together with the annotation coordinates in the metadata. Then the coordinates will be consistent with the image's origin. The same process must be carried out for each image as both the placement on the scanner and the dimensions of the image can vary.

5.1.2 Parameterized tile extraction

Instead of training the model with large gigapixel images, it is now possible to extract tiles from the annotated ROIs in the WSI, see figure 5.2 . This will reduce the amount of data by only extracting the interesting tissues of an WSI. Due to the different colors of the annotations, it is also easy to distinguish between the different tissue types. The annotation colors can further be used to assign an identifier to each tissue type. This way you can iterate over each color and label the tissue with the correct color identification.



(a) Annotation of tumour center.

(b) Tiles drawn from the annotation coordinates.

Figure 5.2: Extracting tiles from an annotated ROI.

Before extracting tiles you need to specify different parameters. First and foremost, the size of the tile must be defined, as well as the resolution you want to retrieve the tile from. In the beginning, an attempt was made to use tiles with size 512x512 pixels at the highest resolution. However, this resulted in too many samples in the dataset which in turn made the training process longer. Additionally, the highest resolution only provided information at the cellular level, see figure 5.3, and for tissue classification there is a need to see the structure, which is revealed at a lower magnification level.

An attempt was then made to reduce the resolution. Too low of a resolution resulted in poor focus of details in the image. However, good results were obtained by scaling down the image to an optical magnification of 10x as shown in figure 5.4 or 5x as shown in figure 5.5. As a result, it produced enough detail to highlight the important information.

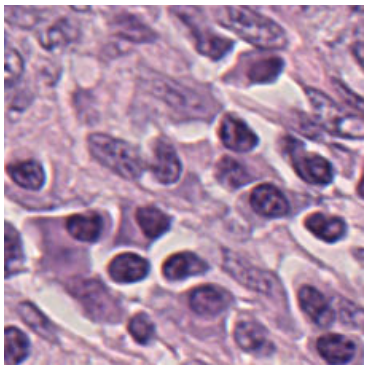


Figure 5.3: Tile at optical magnification 40x.

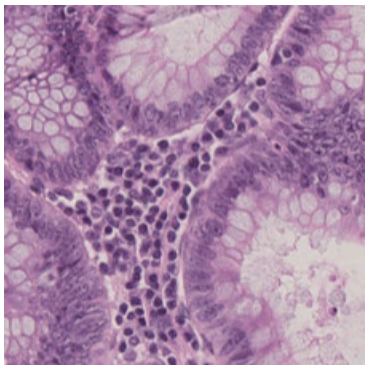


Figure 5.4: Tile at optical magnification 10x.

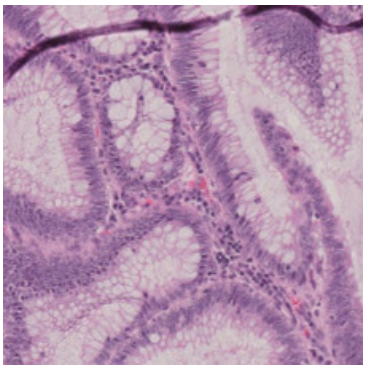


Figure 5.5: Tile at optical magnification 5x.

A challenge that arises now is that too much information comes in each tile. By halving the size of the tiles to 224x224, it is possible to capture more specified information in each tile. This adjustment will also optimize the dataset due to the new input dimensions will ensure compatibility and integration with the pre-trained model VGG16.

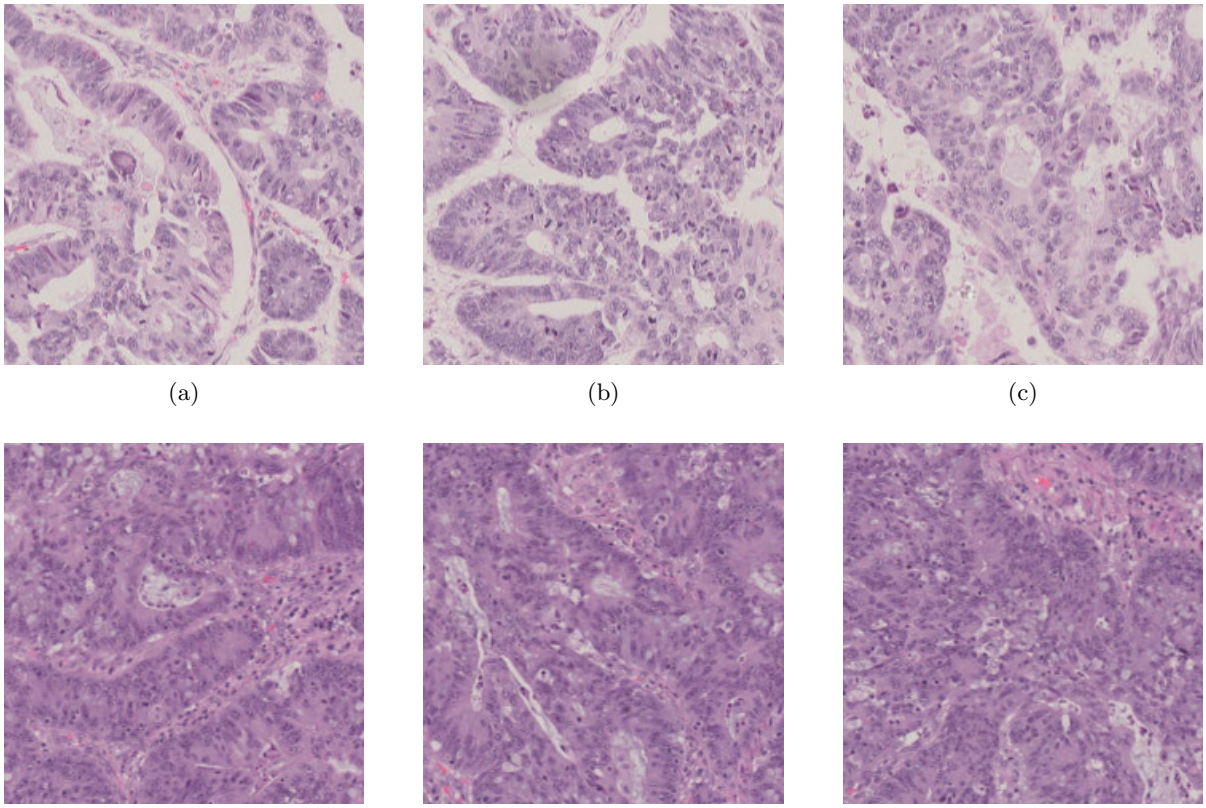
Each WSI also has a different number of levels in each tile. It is therefore necessary to take into account which level a tile belongs to before scaling down and extracting tiles from the whole slide image. Worst case scenario, you end up taking out tiles from the wrong level and thus miss important details. Therefore, the number of levels for each WSI is checked before extracting tiles so that the correct resolution is obtained for the dataset.

Based on this, you can adjust the size and level of the tiles to be extracted before saving all the tiles

in separate folders based on the color of each annotation. Thus, an organized dataset is formed, where each type of tissue is grouped together in its own folders.

5.1.3 Distribution of data material

The tiles that are now organized in categorized folders must further be grouped into test, training and validation sets. The function train_test_split was used to divide the data material into training and validation sets. However, the tiles to be used in the test set had to be processed separately to avoid cross-contamination. This is because the tiles belonging to the same WSI will be very similar to the neighboring tiles and will therefore cause a bias in the training which can affect the model's result.



(a) (b) (c)

(d) (e) (f)

Figure 5.6: Tumour center samples from two different surgical samples. (a), (b), and (c) represent Sample 1, while (d), (e), and (f) represent Sample 2. All taken at 5x optical magnification.

Figure 5.6 shows three tiles of tumour center taken from two different samples. The observation clearly shows that the tiles originating from the same sample are very similar to each other, while tiles from different samples can have a significant difference. Based on this, a manual selection of the 65 WSI was carried out as to what should be the basis for training the machine learning algorithm and what should be kept for testing. Generally, the test set should be around 10-20% of the training set. Therefore 7 WSI were used for testing the model. This manual selection will ensure that no bias occurs in the training and gives a reliable result.

5.1.4 Data analysis

In order to get a better overview of the dataset that has been created, information from each tile is stored in a table. This table can further be used to analyze the data.

slide_name	tissue_type	col	row
D:/wsi_multi/TEST28.ndpi	4	39.80016233766234	9.650365259740258
D:/wsi_multi/TEST28.ndpi	4	40.80016233766234	9.650365259740258
D:/wsi_multi/TEST28.ndpi	4	41.80016233766234	9.650365259740258
D:/wsi_multi/TEST28.ndpi	4	42.80016233766234	9.650365259740258
D:/wsi_multi/TEST28.ndpi	4	43.80016233766234	9.650365259740258
D:/wsi_multi/TEST30.ndpi	2	27.159496753246756	17.514285714285712
D:/wsi_multi/TEST30.ndpi	2	28.15949675324675	17.514285714285712
D:/wsi_multi/TEST30.ndpi	2	18.159496753246756	18.514285714285712
D:/wsi_multi/TEST30.ndpi	2	19.159496753246756	18.514285714285712
D:/wsi_multi/TEST30.ndpi	2	20.159496753246756	18.514285714285712
D:/wsi_multi/TEST30.ndpi	3	12.898579545454549	12.00349025974026
D:/wsi_multi/TEST30.ndpi	3	13.898579545454549	12.00349025974026
D:/wsi_multi/TEST30.ndpi	3	14.898579545454549	12.00349025974026
D:/wsi_multi/TEST30.ndpi	3	15.898579545454549	12.00349025974026
D:/wsi_multi/TEST30.ndpi	3	10.898579545454544	13.00349025974026

Figure 5.7: Table of extracted tiles from a whole slide image (WSI).

Figure 5.7 shows a section of a table after tiles have been extracted. The table shows the path and name of the tile as well as the type of tissue. In addition, you get information about which column and row in the whole slide image the tile is taken from. Here, columns and rows are defined by the fact that the WSI is divided into a grid where the size of each cell is the tile size.

The table serves multiple purposes, including determining how many tiles is associated with each tissue type. This gives the opportunity to get an insight on how the data is distributed. Consequently, informs if more data of a tissue type is needed in order to get a balanced dataset.

5.2 Model architecture

In this project, the VGG16 model was used as the pre-trained model for all experiments. The use of VGG16 as a feature extractor for transfer learning saves time and resources as you do not have to train a model from scratch. In order to adapt the VGG16 model for classification between different tissue types, some extra layers were added on top of the existing VGG16 architecture.

5.2.1 Transfer learning

By using transfer learning, it is possible to benefit from the knowledge and experiences from VGG16. To achieve the best possible performance when using transfer learning, it is necessary to experiment with different parameters and configurations [48].

The VGG16 model is trained on the ImageNet dataset, which means that it is initialized with weights that have already been trained on millions of images. To avoid weights and parameters not being updated during training, it is possible to freeze the layers in the VGG16 model. By freezing the layers, the model will focus on learning custom patterns specific to this classification. It is particularly important to freeze the layers if you have limited training data as this can increase the risk of overfitting [79].

As VGG16 is trained on images of size 224x224 pixels, the dimensions of the input images were specified to the same size. As a result, VGG16 gets the dimension it expects and the results become more reliable. To strengthen the dataset even further, data augmentation is used on the dataset. By doing so, the model gets a more diverse data and consequently becomes more robust.

5.2.2 Training procedure

In addition to the ones already mentioned, there are other parameters that can also be adjusted to optimize both the model and the processing time. Various regularization techniques were used to improve the training. Dropout layers were added to prevent overfitting. Batch normalization had the purpose of smoothing the optimization landscape to provide a more stable behavior of the gradients. Early stopping was added to stop the training if the model did not improve.

At first, the model was trained on 100 epochs with early stopping at 20 epochs. After studying the initial findings, it was clear that numerous epochs only increased the training duration and that the model's performance rarely improved beyond 20 epochs. It turns out that the dataset was not extensive enough to improve over 100 epochs. Thus, the number of epochs was reduced to 20. Early stopping had a patience of 10 epochs, if the model did not improve, the training was terminated.

Training with a deep model and a vast dataset will naturally require more memory. By reducing the batch size, memory utilization will improve, which makes it possible to train larger amounts of data. However, the disadvantage of this is that the smaller the batch size, the longer the the training duration. In this project, a batch consisted of 16 samples. This gave the best balance between memory use and training time in addition to the possibility of parallel processing on GPUs.

Resolution

By retrieving images with a lower resolution, you will be able to save time during training of the model. With the highest resolution, the training took far too long and it was basically not necessary to have such detailed information. The decision fell on extracting information with the best resolution without losing detail, which was an objective magnification of 10x. That is, the resolution of the WSI downsampled by a factor of 4. It may be useful to note that the downsampling of a WSI is exponential. This means that if you descend three levels down in resolution, this corresponds to an optical magnification of 5x. See equation 5.3

$$\text{Updated optical magnification} = \frac{\text{Original optical magnification}}{2^n} \qquad \text{where } n = \text{number of levels down} \tag{5.3}$$

5.3 Tissue classification

Both a binary and multi classification of the tissue types were carried out. This section will show the purpose of the various methods and what distinguishes the two classification types. Appendix B provides the README file for the model training source code [77] [6].

5.3.1 Binary classification

Initially, a model was created that would distinguish between the invasive margin (IM) and the tumour center (TC). Tumour center means the core of the tumour, while invasive margin is where the spread of cancer stops, i.e. the boundary area where it passes into normal tissue.

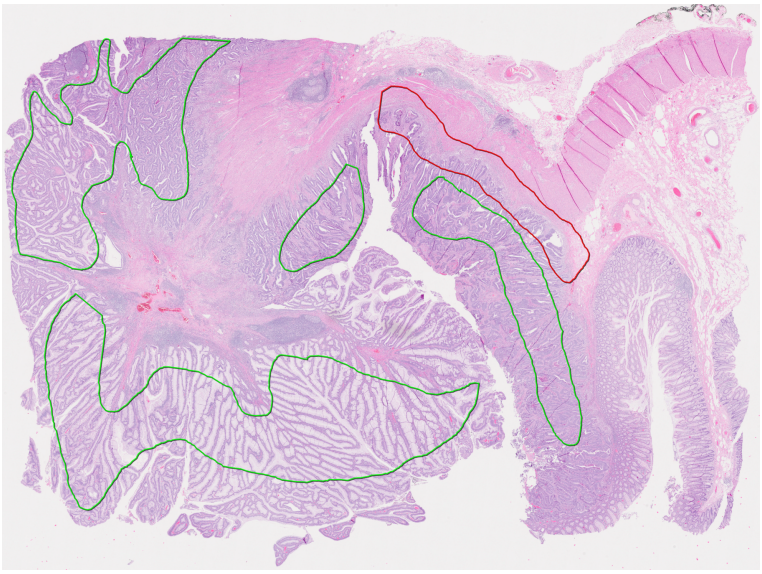


Figure 5.8: Annotated ROIs of TC (green) and IM (red), created with NDP.View2.

During the preparation of the dataset, both tumour center (TC) and invasive margin (IM) had to be annotated. Figure 5.8 shows an example of annotation of a WSI. Green annotation shows TC while red annotation is IM.

One already encounters a challenge in that each WSI contains a much larger proportion of TC than IM, therefore there is a risk of TC being overrepresented. In order to reduce the chance of overrepresentation, not all areas with tumour center were annotated. This gives a greater likelihood of balance during training. As IM includes cancer, a challenge may arise in binary classification. It is difficult to distinguish between something that is definitely cancer and something that is 50% cancer and 50% unknown. This dilemma is described more clearly in the results section, subsection 6.1.

Due to the rise of these problems, the focus was rather shifted to multi classification where one can assess several classes of both cancer and normal tissue and thus get a more nuanced understanding of the tumour and invasive margin.

5.3.2 Multi classification

Apart from the annotations performed in the binary model, two additional tissue types were added for the multi classification. As the invasive margin often occurs in muscularis propria (MP) and subserosal fatty tissue(FT), these two types were added as new classes.

The transition from binary to multi classification includes some modifications in the model, but the actual model architecture may be the same. Firstly, the last layer in the model must be changed to have four outputs instead of one. Moreover, the activation function must be updated to ”softmax”. The model will then be receptive to all four different types of tissues annotated on the WSI and give a probability for each classification.

Furthermore, to update the activation function, the loss function must also be changed from "binary_crossentropy" to "categorical_crossentropy" in order for it to be able to handle a multi classification. The last necessary change is to modify the class_mode parameter of flow_from_directory to "categorical"

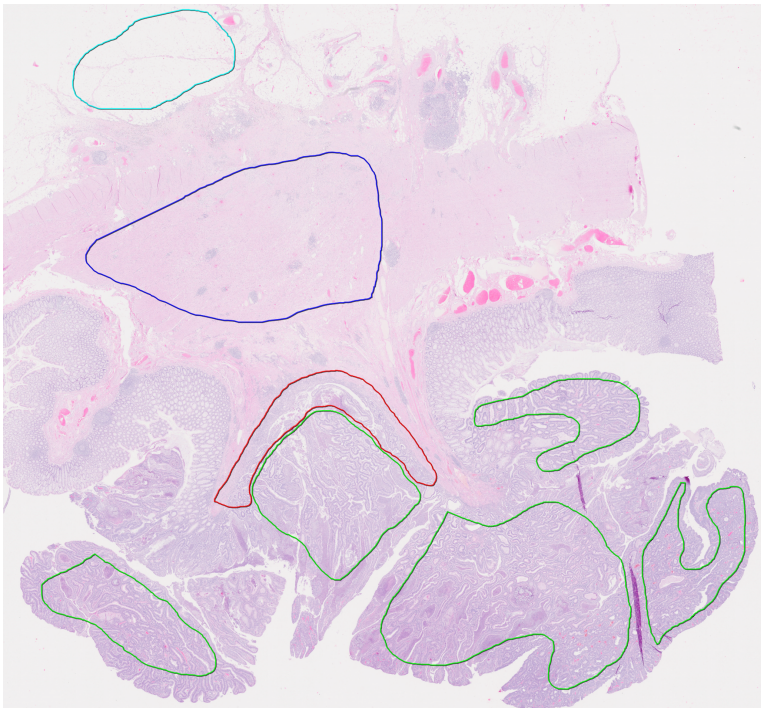


Figure 5.9: Annotation of the four tissues to be classified. IM (red), TC (green), MP (blue), FT (cyan), created with NDP.View2.

Figure 5.9 shows annotations from the 4 different classes; tumour center (TC), invasive area (IM), muscularis propria (MP) and subserosal fatty tissue (FT). Once again, it can be observed that IM is underrepresented.

5.4 Experiments

After the model for multi classification was designed, the model was explored using a set of experiments. These experiments will give more knowledge about the models performance, and consequently providing a stronger foundation for evaluating the model.

5.4.1 Multiple level classification

With suspicion of too weak and unbalanced dataset, the solution was to train models with tiles from different levels of magnification. In this experiment, tiles from three optical magnification levels were extracted; 10x, 5x and 2.5x. The reason the tiles were not taken from higher resolution like 40x or 20x is the way the annotations were carried out. If you extract tiles from the invasive margin with an optical magnification of 40x, there is a greater probability that the invasive border will not be included in each tile. Figure 5.11 attempts to illustrate this.

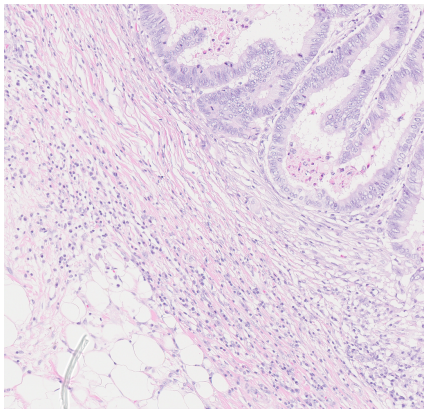


Figure 5.10: Original tile at 5x optical resolution.

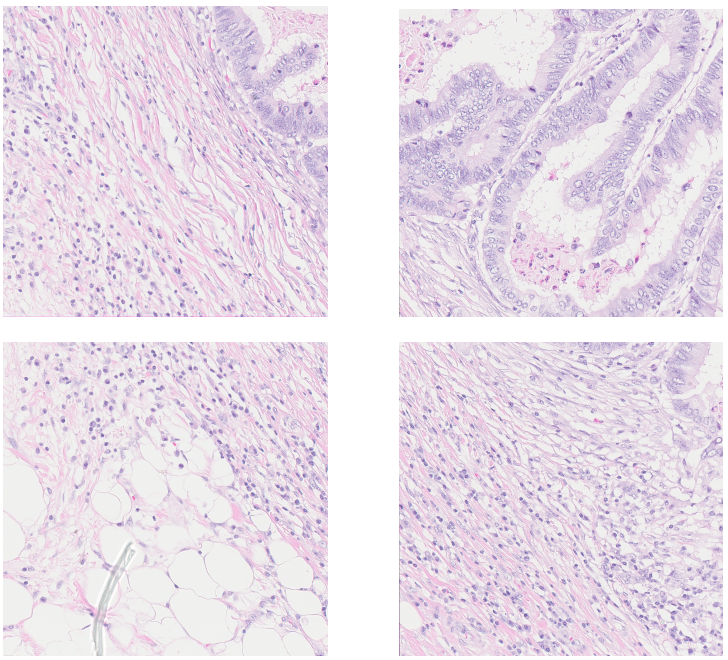


Figure 5.11: The original tile with a higher resolution, resulting in four tiles.

Figure 5.10 shows a tile extracted with optical magnification of 5x. While the tiles in figure 5.11 have a higher resolution, but the same size of the tile. In this instance, the tile at the top left and the tile at the bottom right show the invasive margin. However, the tile at the top right can easily be misinterpreted as tumour center as it contains almost only cancer, compared to the tile at the bottom left which does not contain any cancer and can therefore be interpreted as fatty tissue.

5.4.2 Unlabeled prediction

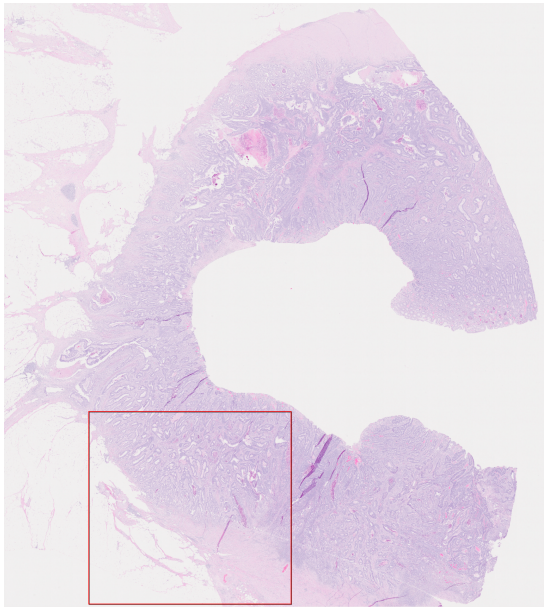
Even though the model is already tested with unknown data during training, this is a different approach. By conducting this experiment you will get a specific prediction for each image rather than a generalized prediction of the training. It is also easier to trouble shoot as you can physically see what the model is struggling to identify by going through each image. This experiment will be carried out using two approaches; specified tiles and unspecified tiles.

Specified tiles

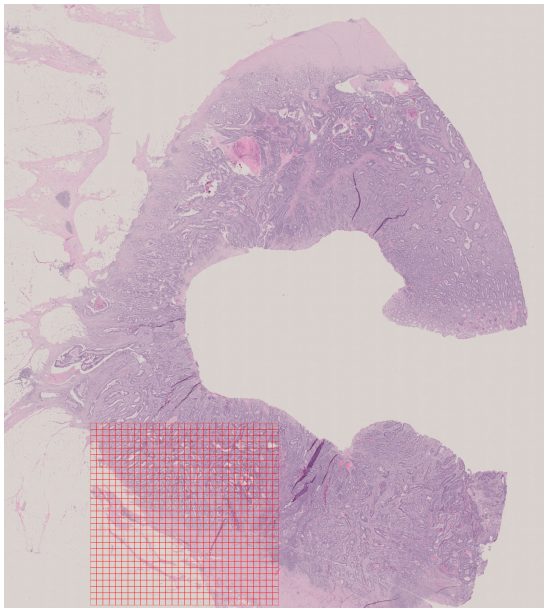
This experiment involves taking random data from annotations from the different tissue types and then collecting the tiles in the same folder, i.e. they are not labeled. This means that a directory of unlabeled tiles was used as input and as output you get what the model predicts and with what certainty by computing the confidence score. The confidence score is given as a percentage and will show the probability of the image being detected correctly by the algorithm [58]. Because the tiles are extracted from given annotations, such as in figure 5.9, each tile that is tested will belong to a class, hence given the name specified tiles.

Unspecified tiles

In order to get a more visual result, it was decided to perform inference in a new way. The experiment involves using the learned patterns and properties in the data to make predictions on a section from a WSI. Therefore, instead of feeding the model with tiles from given tissue annotations, one rather takes a section from a WSI, see figure 5.12a, and extracts tiles from this, see figure 5.12b.



(a) Overview of the annotated area in the WSI



(b) Extracted tiles from the annotated area in the WSI.

Figure 5.12: Annotation and extraction of tiles from a random section of the whole slide image (WSI).

Then each of these tiles will proceed through the model and given a confidence score. The difference is that the tiles will not be specified this time, i.e. they will not be extracted from a class. This allows for a comprehensive observation of how the model behaves when it encounters unknown data, which can be a mixture of various tissue types.

Then each tile will be predicted. If the tile is predicted with more than 70% certainty, it will be placed in a folder with the associated class. If the tile is predicted with less than 70% certainty, it will be placed in a separate folder. To visualize the result, the images in each folder are placed together based on which column and row they had in the original image. This approach makes it possible to see where the results are lacking.

Chapter 6

Result

Throughout this chapter, the result will be described and discussed. Each classification will be evaluated by various metrics to get the best understanding of the result.

6.1 Best model binary classification

The binary model was designed to distinguish between the tumour center (TC) and the invasive margin (IM). The table in figure 6.1 shows the classification report from the best binary classification model.

	precision	recall	f1-score	support
0	0.14	0.06	0.08	67
1	0.79	0.91	0.85	266
accuracy			0.74	333
macro avg	0.47	0.48	0.46	333
weighted avg	0.66	0.74	0.69	333

Figure 6.1: Binary classification report, 0 = invasive margin and 1 = tumour center.

From the overall accuracy, the model appears to perform well with an accuracy of 74%, meaning that it correctly classified 74% of all the test data. Upon closer examination of each class, this suggests something else. The prognosis for the tumour center is good. The precision is 0.79 indicating that 79% of the output that were classified as tumour center were correct. The recall shows an even better result with 91% of the actual tumour center samples were correctly classified.

However, for invasive margin the classification is exceptionally poor. This shows the recall is only 0.06 suggesting that only 6% were correctly classified. Furthermore, only 14% of the invasive margin classifications were correct.

Macro average shows the average result, where each class is weighted equally [43]. For IM, the f1-score is 0.08, while for TC it is 0.85, which is a big difference. Therefore, the macro average becomes the average of the two results which is 0.46. Weighted average takes into account that TC is overrepresented and therefore the performance of each class is calculated with the proportion of samples in each class. This results in a higher f1-score of 69%.

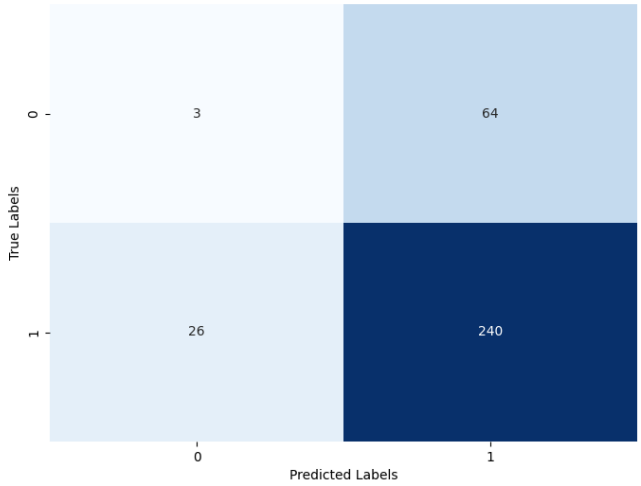
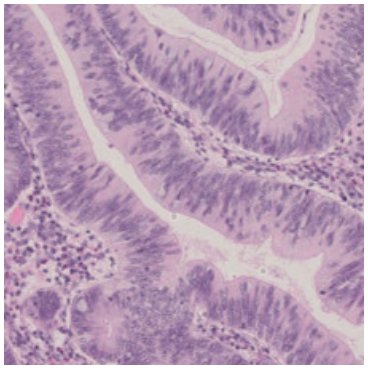


Figure 6.2: Binary classification confusion matrix.

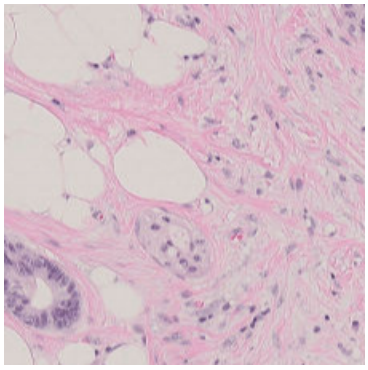
Another way to interpret the result is with a confusion matrix. Figure 6.2 shows the confusion matrix of the best model for binary classification, i.e. consistent with the classification report from figure 6.1. The matrix does not conform to the ideal confusion matrix. It shows that class 1 (TC) is correctly predicted 240 times, while for class 2 the predicted label only matches the true label in three cases.

The poor results may be due to TC being overrepresented, which creates an imbalance in the training. The support tells how much a class is represented, and in the classification report you can see that the model trains with 266 samples from TC versus only 67 samples from IM.

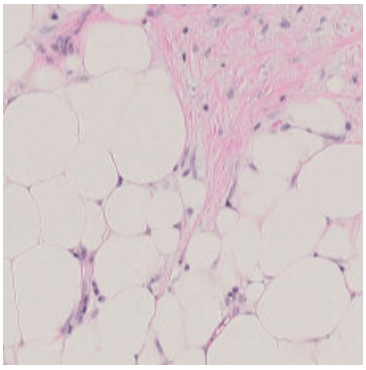
Furthermore, an experiment was carried out in which the model was tested on several unlabeled tiles. The result, see figure 6.3, includes the name of each tile, the model's prediction for each tile, and with what confidence score. When the binary model was used on the unlabeled experiment, it predicted with up to 99% certainty on tiles with tumour center. But because TC is overrepresented, the model would only predict this class. With cases where IM should be predicted, it rather got a certainty of less than 50% for predicting TC.



(a) Predicted TC with 99% certainty.



(b) Predicted TC with 47% certainty.



(c) Predicted TC with 0.5% certainty.

Figure 6.3: Predicted TC images at different certainties.

Figure 6.3 shows with what certainty it predicts that each tile contains tumour center. 6.3a is 99% sure that the tile belongs to class TC, which is correct. The output from 6.3b shows only a 47% certainty that the tile is from TC. In this case, the tile is not from TC, but it belongs to class IM. Figure 6.3c contains

no cancer and has been assumed to be TC with 0.5% certainty. Based on those results, it may indicate that the model looks at how much cancer there is in each tile instead of predicting the class.

6.2 Best model multi classification

The classification report for multi classification is shown in figure 6.4. Following the assumptions from binary classification, the result is not as good as expected and suspicions of a weak and unbalanced dataset arose.

	precision	recall	f1-score	support
0	0.13	0.11	0.12	56
1	0.38	0.40	0.39	288
2	0.23	0.21	0.22	185
3	0.39	0.40	0.39	280
accuracy			0.34	809
macro avg	0.28	0.28	0.28	809
weighted avg	0.33	0.34	0.33	809

Figure 6.4: Multi classification report.

The classification report shows an even worse overall accuracy than for binary classification. Again, this model is more complex as it handles multiple classes. Compared to the binary model, the multi classification identifies the TC class worse. In addition, recall for TC is higher in binary than for multi classification. In summary, binary classification shows better performance on class TC, while multi classification handles several classes and is therefore more complex.

Although the binary model has a higher overall accuracy, it can be noticed that precision, recall and f1-score are all better with multi classification for the invasive margin class. Thus, multi classification gives a more reliable and realistic result as it is weighted over several classes. Again, support shows that invasive margin is underrepresented, which can also give poor results due to an imbalance in the dataset.

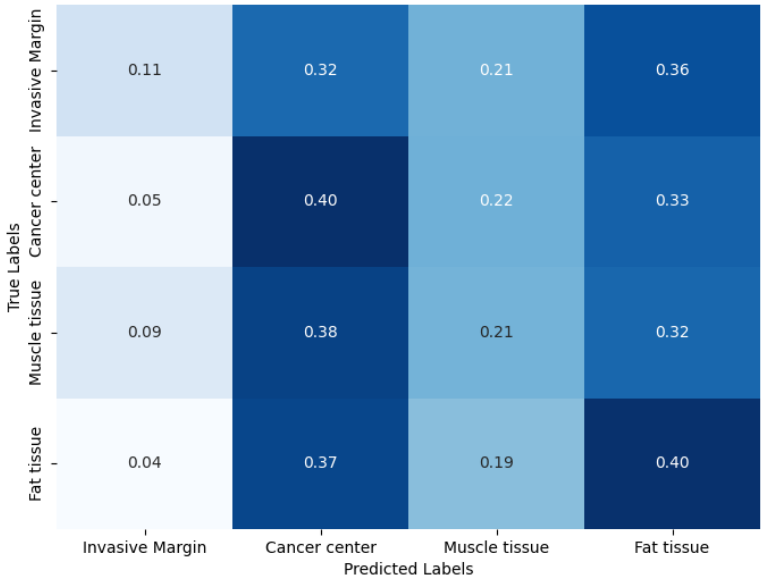
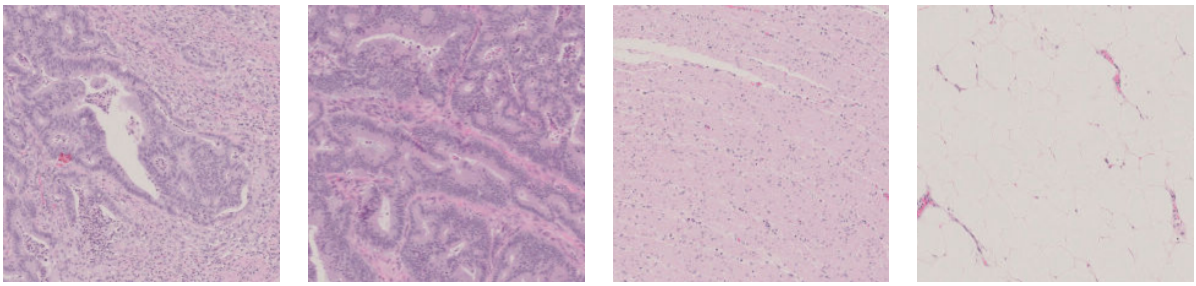


Figure 6.5: Multi classification confusion matrix.

Confusion matrix from figure 6.5 shows that the classes that were predicted the most are tumour center and fat tissue. Again, it predicts rarely the invasive margin. If you compare the confusion matrix with the corresponding classification report in figure 6.4, you see that tumour center and fat tissue have better support than the other two classes and this is reflected in the result. This suggests that the dataset is unbalanced and more samples from the invasive margin and the muscularis propria class are needed.

If the model is run through the unlabeled prediction experiment, you end up with the result shown in figure 6.6.



(a) Predicted TC with 77% certainty. (b) Predicted TC with 97% certainty. (c) Predicted MP with 85% certainty. (d) Predicted FT with 99% certainty.

Figure 6.6: Unlabeled prediction of classes.

The results shows that it predicts both TC, MP and FT very well, but again there were no predictions on IM. Figure 6.6a shows a tile that is predicted to be a tumour center with 77% certainty, but it seems to fit more into the IM category.

6.2.1 Multiple level classification

The idea that more levels would give better results was also weakened when the results were presented. Up until now, it has been trained on around 10,000 tiles. But with several levels, the dataset was over 65,000 files. This is a significant rise from before. Naturally, this will require more storage space and processing time. Unfortunately, the result was not as good as expected. It has improved but not enough to align with the processing time. Figure 6.7 shows the classification report from training with 3 magnification levels; 10x, 5x, 2.5x.

	precision	recall	f1-score	support
0	0.10	0.04	0.05	236
1	0.50	0.54	0.52	1767
2	0.16	0.14	0.15	571
3	0.31	0.32	0.31	1058
accuracy			0.38	3632
macro avg	0.26	0.26	0.26	3632
weighted avg	0.36	0.38	0.37	3632

Figure 6.7: Multi classification report from training with three levels.

The classification report shows worse results for classification of IM, MP and FT but shows a slightly better overall accuracy with an f1-score of 38%. This can be explained by looking at the support, where TC is overrepresented. Even if the model performs worse in all other classes, the overall accuracy will increase as TC has more support. The classes in the model have thus become more imbalanced.

If you look at the Macro average, which does not consider support, this has gone down to 0.26 and thus

worse than the macro average in figure 6.4. But if you take into account that TC is more overrepresented, you get a weighted average of 37%.

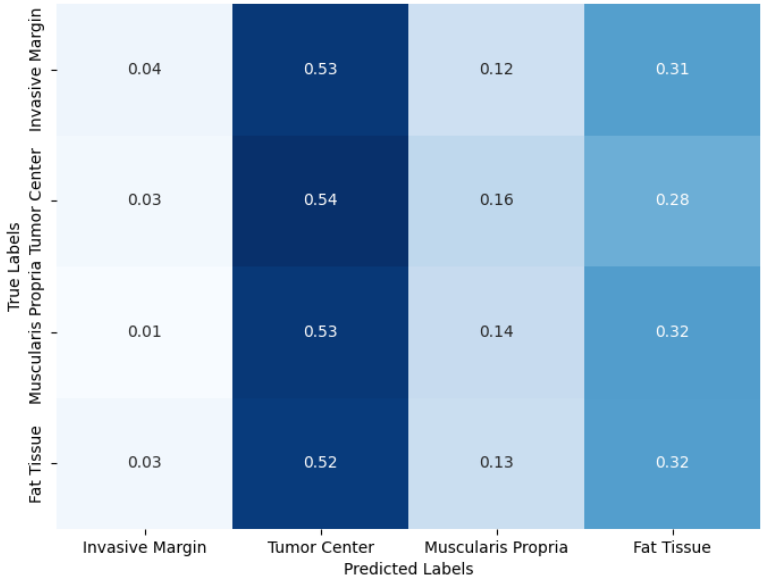


Figure 6.8: Multi classification confusion matrix.

6.3 Experimental results

The output of the unspecified tiles experiment gave better visual results. Figure 6.9 shows the section from the WSI in figure 5.12. This section was further divided into tiles of size 224x224 pixels where each tile was classified. The algorithm was designed so that if a tile was classified with a confidence score of less than 70%, it will not be considered as trustworthy results and will not be included.

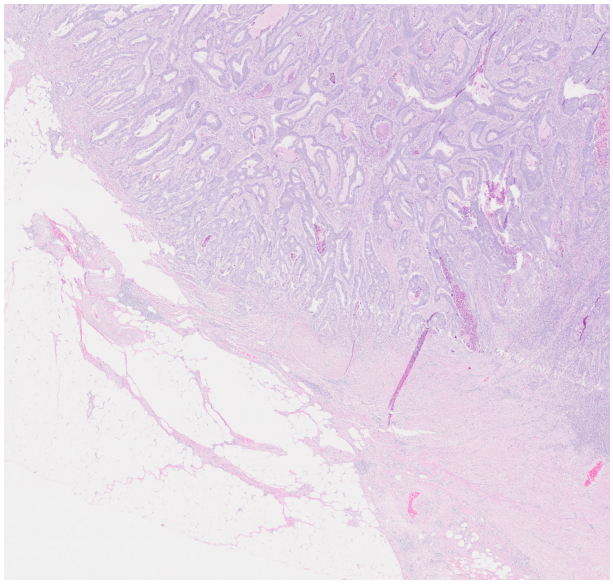
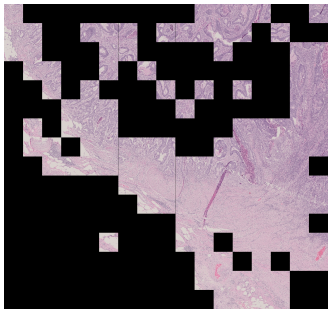


Figure 6.9: Original image.

Firstly, the model trained on tiles with only one level was tested. Figure 6.10 shows the result from

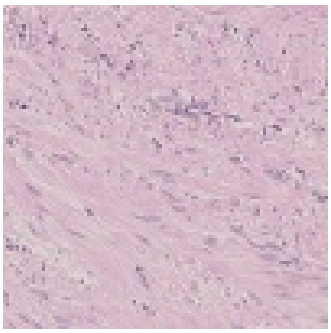
those that were classified with over 70% certainty. It shows that only one tile with muscularis propria was detected, see figure 6.10d, while almost all tiles with fatty tissue were classified correctly. Figure 6.10b shows few classified TC tiles with over 70% confidence score, while classified IM shows more tiles than expected. Overall, prediction of FT is good, while TC and MP have few but correct predicted tiles. Contrary to, classification of IM that shows many but also wrongly classified tiles.



(a) Tiles classified as invasive margin.

(b) Tiles classified as tumour center.

(c) Tiles classified as fatty tissue.



(d) Tiles classified as musculatis propria.

Figure 6.10: Tiles classified with more than 70% confidence score.

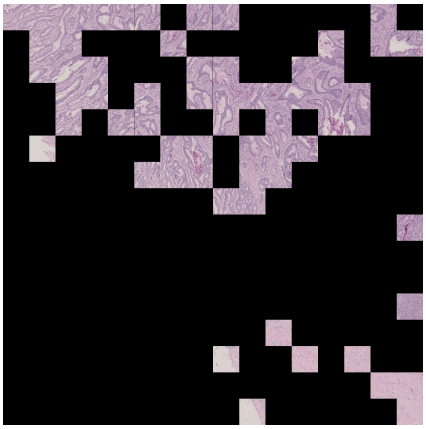
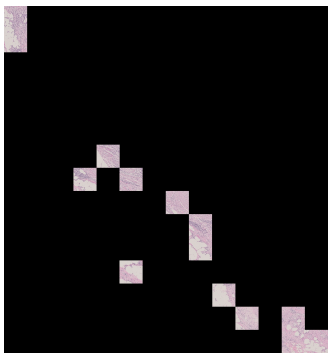


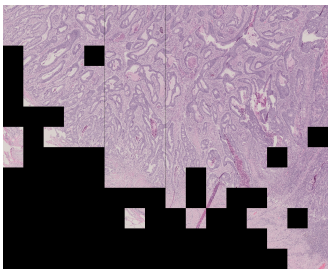
Figure 6.11: Tiles classified with less than 70% confidence score.

Figure 6.11 shows the tiles that did not have a high enough confidence score to be seen as a reliable classification. It shows that muscularis propria and tumour center are the classes that had the worst certainty in their predictions. Combining figure 6.11 and the four images in figure 6.10 will correspond to the original image in figure 6.9.

The same experiment was performed with the model that was trained with several levels from each tile, see figure 6.12. At first glance, one sees a difference in the classification of TC. Here, almost all tiles with TC are predicted correctly, however the majority of IM is misclassified as TC. The classification of FT is performing equally well as previous and MP has slightly more correct classifications. Figure 6.13 shows that the model provides the worst certainty in classifying MP and IM. In a similar manner, all four images from figure 6.12 and the image in figure 6.13 will correspond to the same original image in figure 6.9.

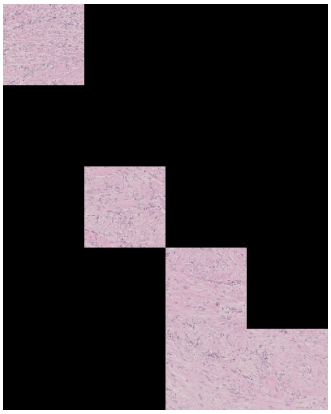


(a) Tiles classified as invasive margin.



(b) Tiles classified as tumour center.

(c) Tiles classified as fatty tissue.

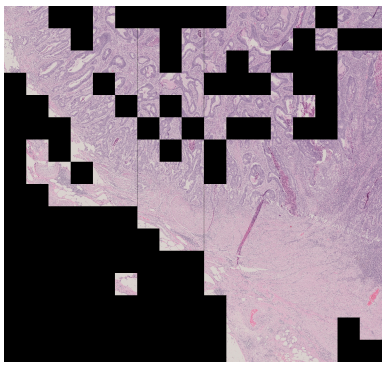


(d) Tiles classified as muscularis propria.

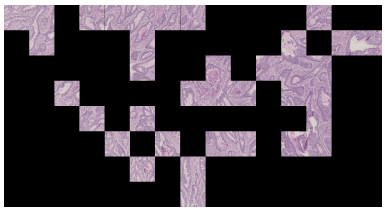
Figure 6.12: Multi level tiles classified with more than 70% certainty.

Figure 6.13: Multi level tiles classified with less than 70% certainty.

Lowering the threshold from 70% to 30% can lead to one of two things; better results or an increase in misclassifications. Figure 6.14 shows the result from the same model as for 6.10, i.e. the model that has been trained on one level. But this time the threshold for reliable classifications has been adjusted down to 30% certainty. It turns out that each classification had over 30% certainty and thus all are seen as credible results. The fatty tissue remains stable with high precision, and MP and TC have received a few more correct classifications. IM, on the other hand, has received an increased misclassification. Thus, even if the classification of TC, MP and FT is better, one cannot set the threshold too low as classification of IM is not reliable.

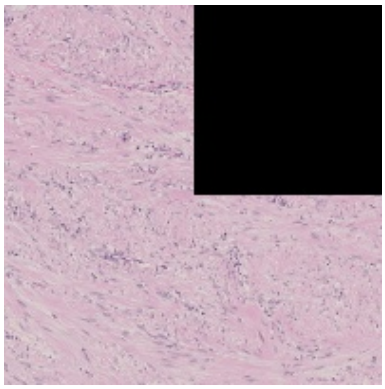


(a) Tiles classified as invasive margin.



(b) Tiles classified as tumour center.

(c) Tiles classified as fatty tissue.



(d) Tiles classified as muscularis propria.

Figure 6.14: Tiles classified with more than 30% certainty.

Chapter 7

Discussion

Profound knowledge and experience have been acquired after completing the task described in section 1.1. It has been created a pipeline that extracts information from a whole slide image (WSI) and utilize this to construct a dataset. Furthermore, a successful deep learning algorithm has been designed to process data from different tissue classes and is trained to recognize the different classes. This forms a solid basis for further work. Although the foundation for the pipeline is established, the result indicates that things can be improved. This chapter will discuss the experiences gained during the project, in addition to what can be improved.

7.1 Experiments

The results from the metrics and the experiments provides a good basis for evaluating the models. What clearly emerges is the lack of data and imprecise annotations. Since the annotations were not carried out by a person of expertise, such as a pathologist, this led to an unstable dataset. For instance, it was discovered later in the project that some of the tumour centers (TC) also contained premalignant epithelium in the annotations.

The biggest surprise was that multi classification did not give a significantly better prediction for IM than binary classification. The hypothesis was that if more tissues were introduced in the classification, the dataset would become more generalized and thus robust. This turned out not to be correct. On the contrary, the multi classification model got a worse f1-score for TC and thus a worse overall accuracy.

It was also experienced that tiles with multi levels did not give noticeably better results. One explanation is that if you double the number of tiles in each class, the imbalance will be even greater. The purpose was to create an extended variation in the dataset, so that the algorithm can learn the structures of the tissue types at different resolutions. For this to yield results, the dataset must take more account of the classes that are underrepresented.

The class that provokes the most issues is invasive margin (IM). This originates from IM being the border between normal tissue and tumour center, and it may be difficult to obtain data with a combination of these classes. Most often, especially with inaccurate annotations, there will be tiles that only contain one of the classes TC, MP or FT in the IM class. In addition, variation of tile sizes will also result in IM containing pure TC, MP or FT classes, referring to figure 5.11.

Unbalanced dataset is also a problem. Invasive margin is only a small part of the whole slide image compared to the other classes. Annotations for TC and FT will have much larger ROIs and therefore be overrepresented. Muscle tissue (MP) is also underrepresented in relation to tumour center and the fatty tissue. Moreover, if the dataset is processed with several levels of tiles, this will also lead to an even greater difference between the classes and hence a more imbalanced dataset.

Throughout the project, the concern about overfitting has been a topic of consideration, as this is a known issue when training deep learning models. However, in this thesis underfitting was rather the problem due to the difference in each WSI, where both the structure and the staining of the slide could

vary. If the training and test set were not separated, the model would have been enormously overfitted. This can be seen from the high values to the validation accuracy.

7.2 Ethical dilemma

A dilemma that is central to the development of digital pathology is how certain can one be that a deep learning model will detect correctly. If false detection occurs, will it be the responsibility of the pathologist, the machine learning model or the developer of the model? Problems such as misdetection can cause a patient to undergo unnecessary surgery. Therefore, the models have to perform with high precision to be a reliable substitute for pathologists. Although there is not yet an answer to these ethical dilemmas, there is still the possibility of using artificial intelligence to facilitate work. But as of now, there is a need for a pathologist to verify the predictions.

7.3 Future work

For future work, it is first recommended to get a person of expertise to annotate the various tissues to be classified. In addition, more data should be processed to create a larger and more robust dataset. If these requirements are met, it is very likely that the model will be remarkably better. Then all uncertainties surrounding the various classes are removed and one can rather focus on specifying the machine learning algorithm to optimize the classification.

As of now, the basis for a model that classifies between different tissues in patients with colorectal cancer has been completed. With the aforementioned improvements, a deep learning model can help ease the workflow of pathologists. An example could be automated cancer detection in whole slide images (WSI), where if cancer is detected the WSI is flagged. If the classification of the invasive border is improved, the model can also be used for automated stage structuring of cancer, as information is obtained on how far the cancer has spread. This will give the pathologists the opportunity to focus on work that cannot be quantized. A pathologist working in parallel with a machine will also lead to increased workforce and thus shorter response time.

Chapter 8

Conclusion

In the course of this master's thesis, several challenges were encountered that exceeded the anticipated time frame. Thus, time was mostly spent on pre-processing the whole slide images rather than creating a solid dataset and classification model. On the contrary, the effort behind establishing a high-quality dataset was recognized by all the time spent on the pre-processing.

Although the results obtained in this thesis do not highlight exceptional accomplishments, it is a promising start for further work. The completed pipeline serves as a foundation for processing whole slide images into a dataset, and then utilize this to train a model with VGG16 as a base. Improvements such as accurate annotations and a larger dataset will provide great potential for using both the binary and multi tissue classifications as an automated diagnostic tool for patients with colorectal cancer.

I've learned a tremendous amount about the precision and standard of a robust dataset, as well as how to make the model training more efficient. After working on this topic, it can be concluded that artificial intelligence serves a purpose in the field of pathology. Automation of pathological tasks will lead to faster diagnosis, shorter response time and, most importantly, ease the work for pathologists.

References

[1] Mahul B. Amin, Stephen B. Edge, et al. *AJCC Cancer Staging Manual*. URL: https://link.springer.com/book/9783319406176. (accessed: 12.06.2023).

[2] Abid Ali Awan. *A Complete Guide to Data Augmentation*. URL: https://www.datacamp.com/tutorial/complete-guide-data-augmentation. (accessed: 22.05.2023).

[3] Pragati Baheti. *Train Test Validation Split: How To and Best Practices [2023]*. URL: https://www.v7labs.com/blog/train-validation-test-set. (accessed: 29.05.2023).

[4] Randall Balestriero and Richard G. Baraniuk. "Batch Normalization Explained". In: (2022). DOI: arXiv:2209.14778. URL: https://arxiv.org/abs/2209.14778.

[5] Aniruddha Bhandari. *Image Augmentation on the fly using Keras ImageDataGenerator!* URL: https://www.analyticsvidhya.com/blog/2020/08/image-augmentation-on-the-fly-using-keras-imagedatagenerator/. (accessed: 20.05.2023).

[6] Dr. Sreenivas Bhattiprolu. $128_{M}alaria_{c}ell_{c}lassification_{C}NN_{w}ith_{d}ata_{a}ug$. URL: https://github.com/bnsreenu/python_for_microscopists/blob/master/128_Malaria_cell_classification_CNN_with_data_aug.py. (accessed: 14.05.2023).

[7] Dr. Sreenivas Bhattiprolu. $266_{o}penslide_{f}or_{w}hole_{s}lide_{i}mages$. URL: https://github.com/bnsreenu/python_for_microscopists/blob/master/266_openslide_for_whole_slide_images/openslide_library_for_whole_slide_images.py. (accessed: 10.03.2023).

[8] Johan Bjorck et al. "Understanding Batch Normalization". In: (2018). DOI: arXiv:1806.02375. URL: https://arxiv.org/abs/1806.02375.

[9] Seldon Blog. *Supervised vs Unsupervised Learning Explained*. URL: https://www.seldon.io/supervised-vs-unsupervised-learning-explained. (accessed: 24.05.2023).

[10] Gaudenz Boesch. *VGG Very Deep Convolutional Networks (VGGNet) – What you need to know*. URL: https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/. (accessed: 11.05.2023).

[11] Akash Borgalli. *Confusion Matrix for N X N Matrix*. URL: https://akash-borgalli.medium.com/confusion-matrix-for-n-x-n-matrix-488e8ff18321. (accessed: 29.05.2023).

[12] Freddie Bray et al. *Tykktarms- og endetarmskreft i Norge - epidemiologi*. URL: https://tidsskriftet.no/2007/10/tema-kolorektal-kreft/tykktarms-og-endetarmskreft-i-norge-epidemiologi (visited on 05/23/2023).

[13] Michael Bretthauer et al. *Effect of Colonoscopy Screening on Risks of Colorectal Cancer and Related Death*. URL: https://www.nejm.org/doi/full/10.1056/NEJMoa2208375. (accessed: 12.06.2023).

[14] Jason Brownlee. *A Gentle Introduction to Dropout for Regularizing Deep Neural Networks*. URL: https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/. (accessed: 20.05.2023).

[15] Jason Brownlee. *How to Normalize, Center, and Standardize Image Pixels in Keras*. URL: https://machinelearningmastery.com/how-to-normalize-center-and-standardize-images-with-the-imagedatagenerator-in-keras/. (accessed: 20.05.2023).

[16] Jason Brownlee. *Use Early Stopping to Halt the Training of Neural Networks At the Right Time*. URL: https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/. (accessed: 19.05.2023).

[17] Christian Chow. *Preventing Training Data Contamination*. URL: https://chrischow.github.io/dataandstuff/2018-09-01-preventing-contamination/. (accessed: 11.05.2023).

[18] PhD. Cindy Seiwert and Goodwin University. *Human Biology*. URL: https://www.labxchange.org/library/pathway/lx-pathway:ba6a713c-0256-46e2-819e-8fe90b0e3660/items/lx-pb:ba6a713c-0256-46e2-819e-8fe90b0e3660:html:c4a58580. (accessed: 15.05.2023).

[19] National institute of neurological disorders and stroke. *Brain Basics: The Life and Death of a Neuron*. URL: https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-life-and-death-neuron. (accessed: 27.04.2023).

[20] Linda Hatleskog Dordi Lea. *Fremtidens patologi er digital*. URL: https://tidsskriftet.no/2022/06/kronikk/fremtidens-patologi-er-digital. (accessed: 23.05.2023).

[21] Joshua Ebner. *How to Use Sklearn train$_t$est$_s$plitinPython*. URL: https://www.sharpsightlabs.com/blog/scikit-train_test_split/. (accessed: 29.05.2023).

[22] B. Ehteshami Bejnordi et al. *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer*. URL: https://pubmed.ncbi.nlm.nih.gov/29234806/#full-view-affiliation-1. (accessed: 07.06.2023).

[23] Wendy L Frankel and Ming Jin. *Oh my: challenges in staging colorectal carcinoma*. URL: https://www.nature.com/articles/modpathol2014128. (accessed: 16.04.2023).

[24] Machine Learning Glossary. *Harmonic Precision-Recall Mean (F1 Score)*. URL: https://machinelearning.wtf/terms/harmonic-precision-recall-mean-f1-score/. (accessed: 30.05.2023).

[25] Gorlapraveen. *VGG16*. URL: https://commons.wikimedia.org/wiki/File:VGG16.png. (accessed: 27.05.2023).

[26] MK Gurucharan. *Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network*. URL: https://www.upgrad.com/blog/basic-cnn-architecture/. (accessed: 19.05.2023).

[27] Grete Hansen. *Kunstig intelligens kan føre til et paradigmeskifte innen cytologi*. URL: https://www.bioingenioren.no/aktuelt/2023/kunstig-intelligens-kan-fore-til-et-paradigmeskifte-innen-cytologi/. (accessed: 07.06.2023).

[28] Muneeb ul Hassan. *VGG16 – Convolutional Network for Classification and Detection*. URL: https://neurohive.io/en/popular-networks/vgg16/. (accessed: 30.05.2023).

[29] Helsedirektoratet. *Benigne polypper*. URL: https://www.helsedirektoratet.no/retningslinjer/kreft-i-tykktarm-og-endetarm-handlingsprogram/polypper-i-tykk-og-endetarm-og-dysplasi-ved-ulceros-kolitt/oppfolging. (accessed: 10.06.2023).

[30] Helsedirektoratet. *https://www.helsedirektoratet.no/retningslinjer/kreft-i-tykktarm-og-endetarm-handlingsprogram*. URL: https://www.helsedirektoratet.no/retningslinjer/kreft-i-tykktarm-og-endetarm-handlingsprogram/kirurgisk-behandling-av-endetarmskreft-uten-fjernmetastaser. (accessed: 06.06.2023).

[31] Johann Huber. *Batch normalization in 3 levels of understanding*. URL: https://towardsdatascience.com/batch-normalization-in-3-levels-of-understanding-14c2da90a338#b93c. (accessed: 15.05.2023).

[32] Petr Kalinichenko Igor Jokic. *Histologi*. Dalefag, 2014.

[33] IndianTechWarrior. *Fully Connected Layers in Convolutional Neural Networks*. URL: https://indiantechwarrior.com/fully-connected-layers-in-convolutional-neural-networks/. (accessed: 28.05.2023).

[34] National Cancer Institute. *Cancer Staging*. URL: https://www.cancer.gov/about-cancer/diagnosis-staging/staging. (accessed: 19.05.2023).

[35] Intel. *What Is a GPU?* URL: https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html. (accessed: 19.05.2023).

[36] Creative Commons Attribution 4.0 International. *FIGURE 2*. URL: https://creativecommons.org/licenses/by/4.0/. (accessed: 28.05.2023).

[37] Moinfar F Jahn SW Plass M. *Digital Pathology: Advantages, Limitations and Emerging Perspectives*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7698715/. (accessed: 23.05.2023).

[38] Shubham.jain Jain. *An Overview of Regularization Techniques in Deep Learning (with Python code).* URL: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/. (accessed: 19.05.2023).

[39] Vandit Jain. *Everything you need to know about "Activation Functions" in Deep learning models.* URL: https://towardsdatascience.com/everything-you-need-to-know-about-activation-functions-in-deep-learning-models-84ba9f82c253. (accessed: 15.05.2023).

[40] Jeremy Jordan. *Evaluating a machine learning model.* URL: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/. (accessed: 30.05.2023).

[41] JoVE. *Harmonic Mean.* URL: https://www.jove.com/science-education/12590/harmonic-mean. (accessed: 29.05.2023).

[42] Andrew Zisserman Karen Simonyan. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* URL: https://arxiv.org/abs/1409.1556. (accessed: 05.05.2023).

[43] Maria Khalusova. *MACHINE LEARNING MODEL EVALUATION METRICS PART 2: MULTI-CLASS CLASSIFICATION.* URL: https://www.mariakhalusova.com/posts/2019-04-17-ml-model-evaluation-metrics-p2/. (accessed: 12.06.2023).

[44] Renu Khandelwal. *Convolutional Neural Network: Feature Map and Filter Visualization.* URL: https://towardsdatascience.com/convolutional-neural-network-feature-map-and-filter-visualization-f75012a5a49c. (accessed: 29.05.2023).

[45] Aman Kharwal. *Classification Report in Machine Learning.* URL: https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/. (accessed: 29.05.2023).

[46] Olbjørn Klepp and Eva Hofsli. *Endetarmskreft.* URL: https://sml.snl.no/endetarmskreft. (accessed: 22.05.2023).

[47] Shivam Kohli. *Understanding a Classification Report For Your Machine Learning Model.* URL: https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397. (accessed: 29.05.2023).

[48] Muriel Kosaka. *Fine-Tuning Pre-trained Model VGG-16.* URL: https://towardsdatascience.com/fine-tuning-pre-trained-model-vgg-16-1277268c537f. (accessed: 01.06.2023).

[49] Simeon Kostadinov. *Understanding Backpropagation Algorithm.* URL: https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd. (accessed: 09.05.2023).

[50] Krefforeningen. *Hva er kreft?* URL: https://kreftforeningen.no/om-kreft/hva-er-kreft/. (accessed: 24.05.2023).

[51] Kreftlex. *Utredning ved kreft i tykk- og endetarm.* URL: https://www.kreftlex.no/Tykk-og-endetarmskreft/ProsedyreFolder/UTREDNING/New-ksProcedureChapter. (accessed: 19.05.2023).

[52] Kreftregisteret. *Kreft i Norge.* URL: https://www.kreftregisteret.no/Temasider/om-kreft/. (accessed: 24.05.2023).

[53] Kreftregisteret. *Nasjonalt kvalitetsregister for tykk- og endetarmskreft.* URL: https://www.kreftregisteret.no/Registrene/Kvalitetsregistrene/Tykk-ogendetarmskreftregisteret/. (accessed: 11.05.2023).

[54] Muthu krishnan. *Understanding the Classification report through sklearn.* URL: https://muthu.co/understanding-the-classification-report-in-sklearn/. (accessed: 29.05.2023).

[55] Khuyen Le. *An overview of VGG16 and NiN models.* URL: https://medium.com/mlearning-ai/an-overview-of-vgg16-and-nin-models-96e4bf398484. (accessed: 05.05.2023).

[56] Dordi Lea. *Use of quantitative pathology to improve grading and predict prognosis in tumours of the gastrointestinal tract.* University of Bergen, 2022.

[57] Alexander LeNail. *NN SVG.* URL: http://alexlenail.me/NN-SVG/. (accessed: 05.06.2023).

[58] Sampurna Mandal et al. "Chapter Four - Single shot detection for detecting real-time flying objects for unmanned aerial vehicle". In: *Artificial Intelligence for Future Generation Robotics.* Ed. by Rabindra Nath Shaw et al. Elsevier, 2021, pp. 37–53. ISBN: 978-0-323-85498-6. DOI: https://doi.org/10.1016/B978-0-323-85498-6.00005-8. URL: https://www.sciencedirect.com/science/article/pii/B9780323854986000058.

[59] Sander A. Søndeland Mari A. Hognestad. *Cardiac rhythm interpretation.* URL: https://github.com/SanderSondeland/ELE690_project1/blob/main/ELE690_project.pdf. (accessed: 10.03.2023).

[60] Mayank Mishra. *Convolutional Neural Networks, Explained*. URL: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939. (accessed: 29.05.2023).

[61] Sarah Moore. *History of Digital Pathology*. URL: https://www.news-medical.net/life-sciences/History-of-Digital-Pathology.aspx. (accessed: 23.05.2023).

[62] Richard Nagyfi. *The differences between Artificial and Biological Neural Networks*. URL: https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7. (accessed: 22.05.2023).

[63] Vibhor Nigam. *Understanding Neural Networks. From neuron to RNN, CNN, and Deep Learning*. URL: https://medium.com/analytics-vidhya/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90. (accessed: 20.05.2023).

[64] Cancer Registry of Norway. *Cancer in Norway*. URL: https://www.kreftregisteret.no/globalassets/cancer-in-norway/2022/cin_report-2022.pdf. (accessed: 24.05.2023).

[65] Openslide. *Hamamatsu format*. URL: https://openslide.org/formats/hamamatsu/. (accessed: 05.03.2023).

[66] World Health Organization. *Cancer*. URL: https://www.who.int/health-topics/cancer#tab=tab_1. (accessed: 15.05.2023).

[67] Oritnk. *Binary confusion matrix*. URL: https://commons.wikimedia.org/wiki/File:Binary_confusion_matrix.png. (accessed: 27.05.2023).

[68] Liron Pantanowitz et al. *Review of the current state of whole slide imaging in pathology*. URL: https://www.sciencedirect.com/science/article/pii/S2153353922002127. (accessed: 15.05.2023).

[69] Christophe Pere. *What are Loss Functions?* URL: https://towardsdatascience.com/what-is-loss-function-1e2605aeb904. (accessed: 09.05.2023).

[70] Hamamatsu Photonic. *Software for NanoZoomer*. URL: https://www.hamamatsu.com/content/dam/hamamatsu-photonics/sites/documents/99_SALES_LIBRARY/sys/SBIS0066E_NDPVIEW2.pdf. (accessed: 10.04.2023).

[71] Hamamatsu Photonic. *Whole Slide Imaging*. URL: https://nanozoomer.hamamatsu.com/jp/en.html. (accessed: 10.04.2023).

[72] DR.MED. Per Holck PROFESSOR EMERITUS. *Tykktarmen*. URL: https://sml.snl.no/tykktarmen. (accessed: 19.05.2023).

[73] Institute for Quality and Efficiency in Health Care (IQWiG). *Colorectal cancer: Overview*. URL: https://www.ncbi.nlm.nih.gov/books/NBK279198/. (accessed: 11.04.2023).

[74] Kreft registeret. *Nasjonalt kvalitetsregister for tykk- og endetarmskreft*. URL: https://www.kreftregisteret.no/Registrene/Kvalitetsregistrene/Tykk-ogendetarmskreftregisteret/. (accessed: 19.05.2023).

[75] Kreft registeret. *Tykk- og endetarmskreft*. URL: https://www.kreftregisteret.no/Temasider/kreftformer/Tykk--og-endetarmskreft/. (accessed: 19.05.2023).

[76] Signe Riemer-Sørensen. *Hva er kunstig intelligens?* URL: https://www.sintef.no/fagomrader/kunstig-intelligens/hva-er-kunstig-intelligens/. (accessed: 11.04.2023).

[77] Moritz Ringler. *How to get a classifier's confidence score for a prediction in sklearn?* URL: https://stackoverflow.com/questions/31129592/how-to-get-a-classifiers-confidence-score-for-a-prediction-in-sklearn. (accessed: 09.06.2023).

[78] Adrian Rosebrock. *ImageNet: VGGNet, ResNet, Inception, and Xception with Keras*. URL: https://pyimagesearch.com/2017/03/20/imagenet-vggnet-resnet-inception-xception-keras/. (accessed: 11.05.2023).

[79] Ram Sagar. *What Does Freezing A Layer Mean And How Does It Help In Fine Tuning Neural Networks*. URL: https://analyticsindiamag.com/what-does-freezing-a-layer-mean-and-how-does-it-help-in-fine-tuning-neural-networks/. (accessed: 01.06.2023).

[80] Geetika saini. *Artificial neural network*. URL: https://commons.wikimedia.org/wiki/File:Artificial_neural_network.png. (accessed: 27.05.2023).

[81] Shibani Santurkar et al. "How Does Batch Normalization Help Optimization?" In: (2018). DOI: arXiv:1805.11604. URL: https://arxiv.org/abs/1805.11604.

[82] scikit-learn. *sklearn.model$_s$election.train$_t$est$_s$plit*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. (accessed: 10.05.2023).

[83] Yashowardhan Shinde. *Custom Data Augmentation using Keras ImageDataGenerator*. URL: https://medium.com/geekculture/custom-data-augmentation-using-keras-imagedatagenerator-7cfd58e54171. (accessed: 20.05.2023).

[84] American Cancer Society. *Colorectal Cancer Screening Tests*. URL: https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/screening-tests-used.html. (accessed: 11.06.2023).

[85] American Cancer Society. *Colorectal Cancer Stages*. URL: https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/staged.html. (accessed: 10.06.2023).

[86] Mayo Clinic Staff. *Colon cancer*. URL: https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669. (accessed: 11.05.2023).

[87] Anthony M. Taylor and Bruno Bordoni. *Histology, Blood Vascular System*. URL: https://www.ncbi.nlm.nih.gov/books/NBK553217/. (accessed: 10.06.2023).

[88] Goran tek-en. *Layers of the Alimentary Canal. The wall of the alimentary canal has four basic tissue layers: the mucosa, submucosa, muscularis, and serosa.* URL: https://commons.wikimedia.org/wiki/File:Layers_of_the_GI_Tract_english.svg. (accessed: 01.06.2023).

[89] TensorFlow. *tf.keras.preprocessing.image.ImageDataGenerator*. URL: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator. (accessed: 20.05.2023).

[90] Rohit Thakur. *Beginner's Guide to VGG16 Implementation in Keras*. URL: https://builtin.com/machine-learning/vgg16. (accessed: 05.05.2023).

[91] Axel Tidemann and Anne Cathrine Elster. *Maskinlæring*. URL: https://snl.no/maskinl%C3%A6ring. (accessed: 11.04.2023).

[92] Cancer Research UK. *How cancer can spread*. URL: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-can-spread. (accessed: 10.06.2023).

[93] Rune Wetteland. *A Multiscale Approach for Whole-Slide Image Segmentation of five Tissue Classes in Urothelial Carcinoma Slides*. URL: https://github.com/Biomedical-Data-Analysis-Laboratory/multiscale-tissue-segmentation-for-urothelial-carcinoma. (accessed: 10.03.2023).

[94] Rune Wetteland. *Automated Grading of Bladder Cancer using Deep Learning*. URL: https://hdl.handle.net/11250/2977487. (accessed: 06.02.2023).

[95] Harsh Yadav. *Dropout in Neural Networks*. URL: https://towardsdatascience.com/dropout-in-neural-networks-47a162d621d9. (accessed: 20.05.2023).

[96] Yadnesh. *Classification and its Performance Metrics in Machine Learning*. URL: https://medium.com/analytics-vidhya/classification-and-its-performance-metrics-in-machine-learning-f0ad57866ec9. (accessed: 29.05.2023).

Appendix A

Preprocessing

= README

This the the souce code for the pre-processing part of my master thesis. The structure and algorithm for the pipeline is inspired by Rune Wetteland's paper "A Multiscale Approach for Whole-Slide Image Segmentation of five Tissue Classes in Urothelial Carcinoma Slides". While the content is based on Dr. Sreenivas Bhattiprolu's work on the use of OpenSlide for whole slide images. My supervisor, Trygve Christian Eftestøl, has also played a crucial role in the development of this source code.

== Library versions

- et-xmlfile==1.1.0

- matplotlib==3.5.3

- matplotlib-inline==0.1.6

- numpy==1.23.0

- keras==2.12.0

- openslide-python==1.2.0

- pandas==1.4.4

- Pillow==9.2.0

- pyvips==2.2.1

- tensorflow==2.12.0

== Pre-processing

The scripts is designed to read whole slide images (WSI) in NDPI format, extract annotation points from corresponding NDPA files, and generate tiles based on the annotations. The scripts uses the OpenSlide library to handle NDPI files and the ElementTree library to parse NDPA files. The algorithm is designed so that it can generate tiles from up to four different classes, i.e. colors of annotation

=== Setup

Firstly install the required libraries. Then place your NDPI files in the designated directory (ndpi_dir). Store the corresponding NDP files in the specified directory (ndpa_dir). Set the desired parameters in the code such as 'tile_size', 'physical_pixel_size', 'tile_step', and 'downsample'. Note, pyhsical pixel is already given from the NDP.View2 software. In this project it turned out to be 220 pixels/nm.

=== Usage

The script will iterate through each NDPI file in the ndpi_dir directory and search for a matching NDPA file in the ndpa_dir directory If a match is found, the script will proceed with the extraction process. The script reads the WSI using OpenSlide and retrieves the annotation points from the NDPA file. The annotations are categorized into different tissue types (e.g., red, green, blue, cyan) based on their color

codes. Tiles are generated using the DeepZoomGenerator, and the script iterates over each tissue type to extract tiles within the specified regions of interest (ROIs). The tiles are saved to local directories based on their tissue type. The script also updates a DataFrame (tile_table) with information about each extracted tile, including the slide name, tissue type, column, and row. This table is then used to count each instance of the tissue types to get an overview of the data set.

The process is repeated for each NDPI-NDPA pair found in the directories.

=== Output

Extracted tiles are saved in separate directories based on their tissue type. The tile_table DataFrame is updated with information about each extracted tile.

Please note that the script assumes specific naming conventions for NDPI and NDPA files and that the directories for NDPI and NDPA files are correctly specified. Adjust the code as necessary to fit your file structure and naming conventions. Feel free to modify the script or integrate it into your own workflow to suit your specific requirements.

== WSI tile drawing

In order to see if the extracted tiles is within the ROI, the script '*draw_tiles.py*' is used. This code does NOT take in multiple annotations as it is just used as an iillustration and is not a part of the pre-processing.

The script allows you to extract coordinates from an NDPA file (annotation file) and use those coordinates to draw tiles on a whole slide image (WSI) file. The tiles are then saved as a separate image.

=== Usage

Import the necessary libraries and define the following parameters; filename_ndpi, filename_ndpa and tile_size. Call the functions 'extract_coordinates' and 'draw_tiles' and display the image with the drawn tiles.

NOTE: Change the tissue type (tissue1, tissue2, ..) according what color annotation you are using.

== Image Data Splitting

Two scripts are used to split the dataset, *'split_multi'* an *'split_binary'*. They share the same purpose in that they will splitt the dataset into three different folders; train, test and validation. The difference is that binary is for creating data set to binary classification and multi is for multi classification.

=== Usage

Both scrips include two functions 'split_train_val' and 'split_test'. To use the script, specify the path to the folders where the extracted tiles are stored and the path to the train, test and validation folders. Select the folder type to determine the splitting behaviour. This is to prevent cross-contamination, if the functions is not handeled seperately, tiles from the same biopsy test can occur in both train, test and validation folders.

== How to

A short description of the workflow:

- Run the preprosessing script with the data that is to be used for training and validation. Depending on how many levels you want to extract from, run the script for each level - adjusting only - Depending on if you are running with two or more classes, choose between *'split_multi'* an *'split_binary'*. - Check that foldertype = 1 and run the script. Now you have split the data into train and validation set. - Run preprocessing again, but this time with a new set of data (.ndpi and .ndpa pair) that is to be used for testing. - Go to the splitting script (either *'split_multi'* or *'split_binary'*) and change to another number (0 or 2) before running. Now you have created the test set. - The train, val and test folders can then be used as input for a deep learning model.

Appendix B

Model training

= Train and evaluate model

This the the souce code for the model training part of my master thesis. The code is based on previous work, and inspired by Dr. Sreenivas Bhattiprolu's work on CNN classification with data augmentaton. Additionally, I received assistance from the Stack Overflow community.

== Library versions

- et-xmlfile==1.1.0

- matplotlib==3.5.3

- matplotlib-inline==0.1.6

- numpy==1.23.0

- keras==2.12.0

- openslide-python==1.2.0

- pandas==1.4.4

- Pillow==9.2.0

- pyvips==2.2.1

- tensorflow==2.12.0

== Train Model

The scripts '*train_model_binary.py*' and '*train_model_multi.py*' is designed to train a model based on the pre-trained VGG16 model. The only difference is how many classes they have as input and output. In '*train_model_multi.py*' four classes are considered, while binary will naturally only have two.

=== Setup

In order to run either of the scrips, you need to provide the path to the training, test and validation folders. Additionally 'tile_size' and 'batch_size' need to be specified. As output, you will get a classifiaction report and confution matrix for each training.

== Test model

The scripts '*test_model_binary.py*' and '*test_model_multi.py*' is created to test the models. Both scripts uses the confidence score to show the probability of the image being detected correctly by the algorithm.

=== Setup

In order to run the scrips, you need to provide path to the test folder, path to the best model and define the class labes. As output, the binary test will print out the image name with the conficence score. While

the multi test, will to the same in addition to store the images in new predicted folders. Then these images can be joint back together to create the image based on what the algorithm is classifying.

Appendix C

Poster presentation

AI-based Tissue Classification in Whole Slide Images: Differentiating Tumor Center and Invasive Margin

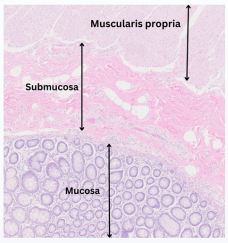
Mari Hognestad
University of Stavanger

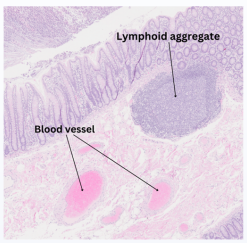
ABSTRACT

MSc project within biomedical image analysis of histological images of colon and rectal cancer in collaboration with Stavanger University Hospital (SUS). The primary objective of the project is to develop a classification system that can accurately differentiate between different tissue types within whole slide images (WSI), specifically focusing on the distinction between tumor center and invasive margin.

BACKGROUND

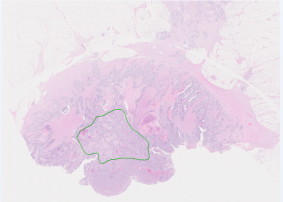
Colorectal cancer is one of the most frequent forms of cancer in Norway. In 2022, 3,385 people got colon cancer and 1,267 people rectal cancer [1]. cases of cancer, the surgical preparation (tumor) will be assessed by a pathologist to determine the degree of severity that governs further treatment. To facilitate the workflow of pathologists, it is possible to develop machine learning models that can recognize tissue types and structures in digitized histological images, referred to as whole slide images (WSI).

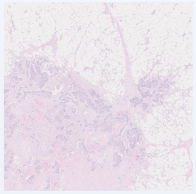


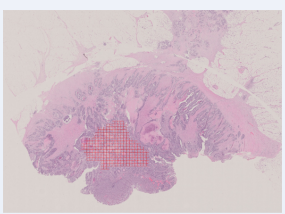


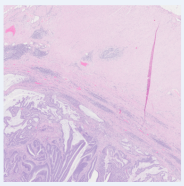
METHODS

The proposed approach for tissue classification involved the utilization of a pre-trained model, VGG16. However, prior to initiating the training process, it was imperative to pre-process the raw data (WSI) and transform it into a suitable dataset format.









RESULT

From the overall accuracy, the model appears to perform well with an accuracy of 74%, meaning that it correctly classified 74% of all the test data. Upon closer examination of each class, this suggests something else. The prognosis for the tumor center is good. The precision is 0.79 indicating that 79% of the output that were classified as tumor center were correct. The recall shows an even better result with 91% of the actual tumor center samples were correctly classified.

However, for invasive margin the classification is exceptionally poor. This shows the recall is only 0.06 suggesting that only 6% were correctly classified. Furthermore, only 14% of the invasive margin classifications were correct.

	precision	recall	f1-score	support
0	0.14	0.06	0.08	67
1	0.79	0.91	0.85	266
accuracy			0.74	333
macro avg	0.47	0.48	0.46	333
weighted avg	0.66	0.74	0.69	333

CONCLUSION

The results obtained from the classification model indicate good performance in detecting the tumor centers, while the classification of the invasive areas shows relatively poorer performance. Several factors may contribute to this outcome.

The poor results may be due to TC being overrepresented, which creates an imbalance in the training. The support tells how much a class is represented, and in the classification report you can see that the model trains with 266 samples from TC versus only 67 samples from IM. This will also lead to a potential bias in the model's performance. Additionally, the presence of tumor cells within the invasive areas can introduce problems, as these regions can be easily misclassified as tumors. It is also worth noting that the annotations were not performed by pathologists, which could introduce some variation in the quality and accuracy of the annotations.

A dilemma that is central to the development of digital pathology is how certain can one be that a deep learning model will detect correctly. If false detection occurs, will it be the responsibility of the pathologist, the machine learning model or the developer of the model? Although there is not yet an answer to these ethical dilemmas, there is still the possibility of using artificial intelligence to facilitate work. But as of now, there is a need for a pathologist to check and verify the predictions.

REFRENCES

[1] https://www.kreftregisteret.no/Generelt/Rapporter/Cancer-in-Norway/

CONTACT

Mari Amdalsrød Hognestad
250627@uis.no