Diploma Thesis

Creation of Annotated Datasets for the Development of AI-Supported Colon Polyp-Classification Algorithms

submitted by

Joseph Bela Bukovics

attaining the academic degree of

Doktor der gesamten Heilkunde

(Dr. med. univ.)

at the

Medical University of Graz

conducted at the

Diagnostic- & Research-Institute of Pathology

under the supervision of

Univ. FÄ^{in} Priv.-Doz.^{in} Dr.^{in} med.univ.et scient.med. Iva Brcic

Dipl. Ing. Markus Plass, BSc

Graz, 31.05.2023

Affidavit (Eidesstattliche Erklärung)

I hereby confirm that the present diploma thesis is the result of my own independent scholarly work. I also confirm that in all cases, where material from the work of others (in books, articles, essays, dissertations and on the internet) is acknowledged, quotations and paraphrases are clearly indicated. No material other than that cited in the reference list has been used. I have read and understood the Medical University's regulations and procedures concerning plagiarism.

Graz am 31.05.2023 *Joseph Bukovics eh.*

Acknowledgements

I am grateful for the unwavering support of my two supervisors, Priv.-Doz.in Dr.in med.univ.et scient.med. Iva Brcic and Dipl. Ing. Markus Plass, as I worked on my diploma thesis. I would also like to extend my heartfelt appreciation to my family and friends for their continuous encouragement. Laura, I am truly grateful for your presence in my life.

In loving memory of my grandfather, Rudolf Egger, whose unwavering passion for carpentry inspires me deeply. Just as a microscope unveils the microscopic world, exploring wood up close reveals its hidden wonders, igniting a sense of awe and fascination for the wonders of nature.

Table of Content

Abbreviations

H&E	Hematoxylin and eosin
CA	Conventional adenoma
TA	Tubular adenoma
VA	Villous adenoma
TVA	Tubulovillous adenoma
CRC	Colorectal cancer
CIN	Chromosomal instability
MSI	Microsatellite instability
CIMP	CpG island hypermethylation phenotype
LGD	Low-grade dysplasia
HGD	High-grade dysplasia
HP	Hyperplastic polyp
GCHP	Goblet cell rich hyperplastic polyp
MVHP	Microvesicular hyperplastic polyp
SSL	Sessile serrated lesion
SSL-D	Sessile serrated lesion with dysplastic alterations
TSA	Traditional serrated adenoma
TP	Telepathology
WSI	Whole slide image
CPATH	Computational pathology
AI	Artificial intelligence
ML	Machine learning
DL	Deep learning
CNN	Convolutional neural networks

List of Figures

List of Tables

Abstract

Introduction: The colorectal carcinoma ranks third globally in cancer statistics, preceded by bronchial and mammary carcinomas. Some colon polyps, particularly adenomas and sessile serrated lesions, are considered precursors to colorectal carcinoma. The removal of colon polyps during colonoscopy can significantly reduce the risk of developing colorectal carcinoma. Therefore, the timely detection and their removal play a crucial role in the prevention of colorectal carcinoma. The objective of this thesis is to create an annotated dataset that will serve as the foundation for developing an artificial intelligence-based algorithm.

Methods: The Biobank Graz provided the required slides for this project. Their storage comprises of over 11 million histologic slides with patient data. For this study, a cohort of patients with colon polyps from the years 1984 to 2014 was chosen. Histologic slides of these colon polyps were retrieved and digitized using high-resolution whole slide imaging scanners, followed by anonymization of the data. Further annotations were made using the open-source software QuPath. The dataset comprises of H&E-stained WSIs depicting both colon polyps and normal colonic mucosa. The crop images were exported as rectangular images, with the annotation-polygon indicating the center of each image. All images have a resolution of 1024x1024 pixels.

Results: The dataset comprises 533 whole slide images that were stained with hematoxylin and eosin (H&E). Among them, 33 slides were excluded from the project. In total, 17,937 image samples were collected, with 10,088 representing physiological glands and 7,848 representing dysplastic glands.

Conclusion: Our colon gland dataset is of significant size and quality compared to publicly available datasets that have been used for studies and scientific contests. However, it is important to note that direct comparison of these datasets may be limited due to differences in their intended purposes. Automated analysis will be an important part in digitized histopathology, but diverse tissue structures and subjective evaluations could cause difficulties. Robust computational methods are needed for diagnostic reproducibility. The implementation of medical application-focused deep learning models in digital pathology has the potential to reduce the time and workload of

clinicians and pathologists, minimize potential errors, and improve the accuracy of colorectal cancer screening.

Zusammenfassung

Einleitung: Das kolorektale Karzinom steht weltweit an dritter Stelle in der Krebsstatistik, nach dem Bronchial- und Mammakarzinom. Einige Kolonpolypen, insbesondere Adenome und sessile serratierte Läsionen, gelten als Vorstufen des kolorektalen Karzinoms. Die Entfernung von Kolonpolypen während einer Koloskopie kann das Risiko für die Entwicklung von kolorektalem Karzinom erheblich reduzieren. Daher spielt die frühzeitige Erkennung und deren Entfernung eine bedeutende Rolle bei der Prävention von kolorektalem Karzinom. Das Ziel dieser Diplomarbeit ist die Erstellung eines annotierten Datensatzes, welcher für die Erstellung eines Algorithmus, basierend auf künstlicher Intelligenz, dienen soll.

Methodik: Die Biobank Graz stellte die erforderlichen Präparate für dieses Projekt zur Verfügung. Ihr Bestand umfasst über 11 Millionen histologische Präparate mit Patient*innendaten. Für diese Studie wurde sorgfältig eine Kohorte von Patient*innen mit Kolonpolypen aus den Jahren 1984 bis 2014 ausgewählt. Die histologischen Präparate dieser Kolonpolypen wurden abgerufen und mittels hochauflösenden Whole-Slide-Imaging-Scanner digitalisiert, gefolgt von der Anonymisierung der Daten. Die darauffolgenden Annotationen wurden mit der Open-Source-Software QuPath durchgeführt. Der Datensatz umfasst H&E-gefärbte histologische Schnitte, die sowohl Kolonpolypen als auch normale Kolonschleimhaut beinhalten. Die Ausschnittbilder wurden als rechteckige Bilder exportiert, wobei das Annotation-Polygon das Zentrum jedes Bildes markiert. Alle Bilder haben eine Auflösung von 1024x1024 Pixeln.

Ergebnisse: Das Datenset umfasst 533 histologische Schnittbilder, die mit Hämatoxylin und Eosin (H&E) gefärbt wurden. Davon wurden 33 Schnitte aus dem Projekt ausgeschlossen. Insgesamt wurden 17.937 Bildbereiche gesammelt, wobei 10.088 physiologische Drüsen und 7.848 dysplastische Drüsen repräsentieren.

Conclusio: Unser Datensatz ist im Vergleich zu ähnlichen Datensätzen, die für Studien und wissenschaftliche Wettbewerbe verwendet wurden, von signifikanter Größe und Qualität. Es ist jedoch wichtig zu beachten, dass eine direkte Vergleichbarkeit dieser Datensätze aufgrund von Unterschieden in ihren vorgesehenen Zwecken eingeschränkt sein kann. Automatisierte Analysen werden in

Zukunft in der Histopathologie immer wichtiger, aber unterschiedliche Gewebestrukturen und subjektive Bewertungen können weiterhin eine Herausforderung darstellen. Robuste Algorithmen sind für eine diagnostische Reproduzierbarkeit erforderlich. Die Implementierung von Deep-Learning-Modellen in der digitalen Pathologie hat das Potenzial, die Zeit- und Arbeitsbelastung von Ärzt*innen und Patholog*innen zu reduzieren, potenzielle Fehler zu minimieren und die Genauigkeit der Darmkrebsvorsorge zu verbessern.

1 Introduction

1.1 Scientific Background

Colorectal carcinoma (CRC) is statistically the third most common cancer worldwide. (1) Malignant neoplasms of the colon arise from precursor lesions. Therefore, an adequate diagnosis and management of colonic polyps is of utmost importance to hinder the malignant progression. (2)

1.1.1 Colon: Anatomy and histology

The large intestine begins at the ileocecal valve and consist of three main sections: right colon, left colon and rectum. (3) The right colon includes the cecum, ascending and transverse colon, and the left colon consists of the descending and sigmoid colon (see Figure 1). (4)

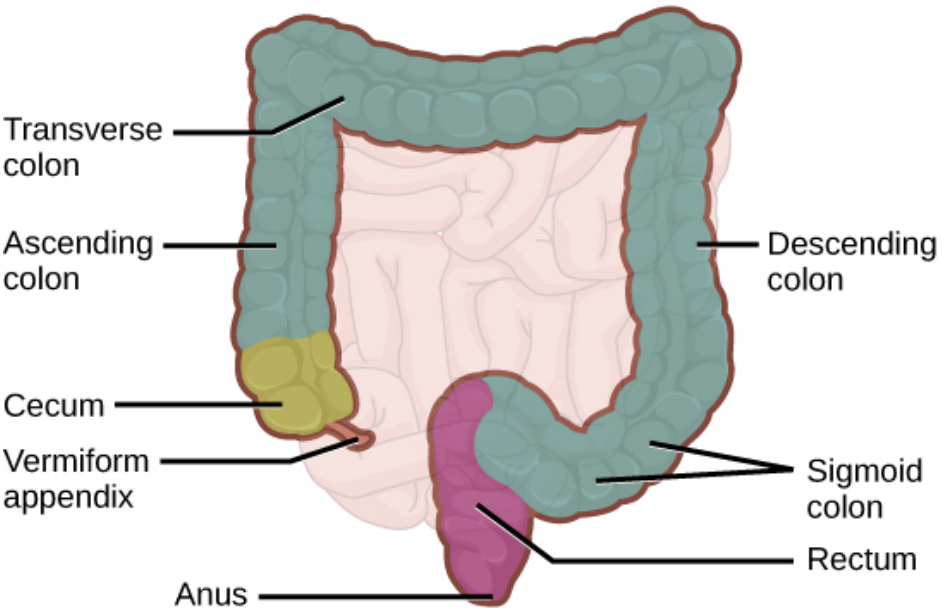

Figure 1: Illustration of the colon. Taken from Anatomy and Physiology, OpenStax *(Licensed: CC-BY) (5)*

Microscopically, the colonic wall is composed of the mucosa, submucosa, muscularis propria, sub-serosa and serosa. (6) The mucosa is comprised of a single-layered epithelium, lamina propria and muscularis mucosae. In the epithelium, the cubic to prismatic absorptive cells are admixed with goblet cells and are lying on a basal membrane and a layer of connective tissue, the lamina propria. Occasionally, scattered lymphocytes can be found between the epithelial cells. A thin muscular layer, the muscularis mucosae, is found underneath. (3)

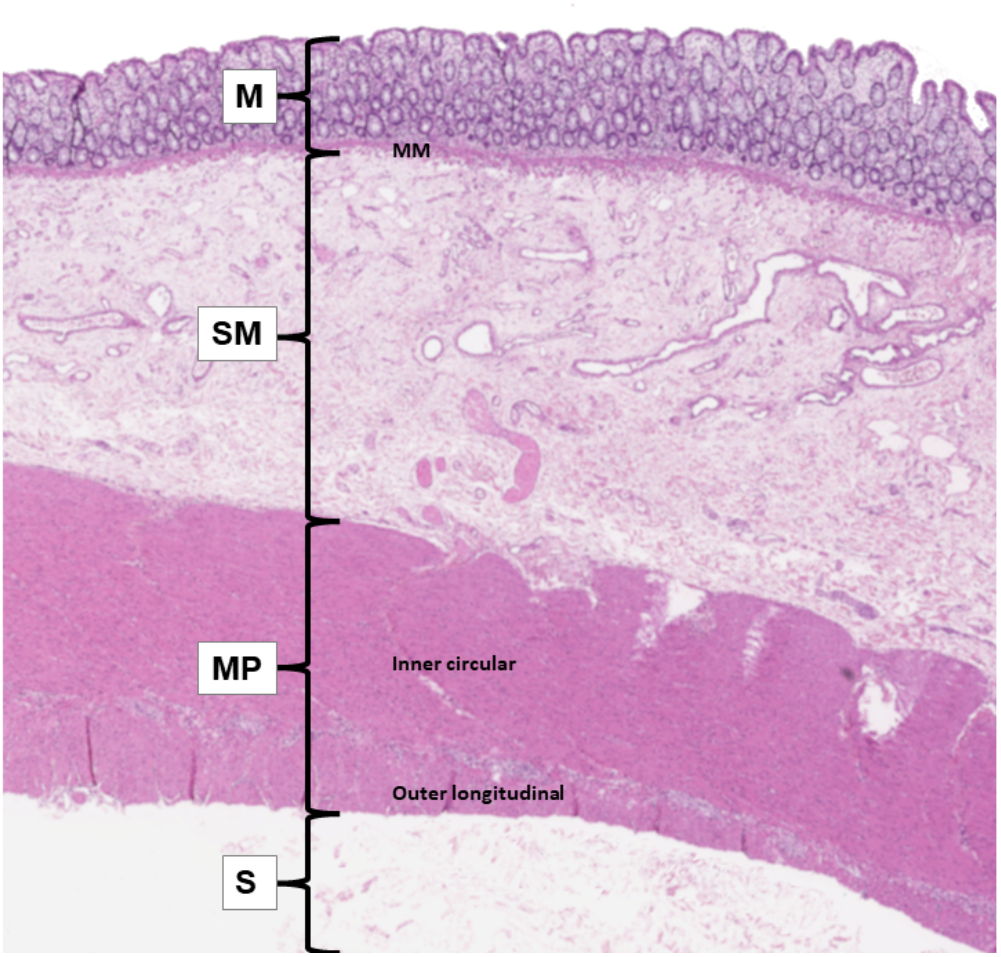

Figure 2: Layers of the colon. *Consisting of the mucosa (M), submucosa (SM), muscularis propria (MP), sub-serosa and serosa (S)*

A uniform architecture of U-shaped and evenly arranged crypts is characteristic for the colon. (4) Other specific cells are found in these crypts, such as: undifferentiated stem cells, endocrine cells, and Paneth cells. At the crypt base, regeneration of all cell types takes place through meiotic division of a common stem cell. The resulting cells migrate up to the apical part while the luminal quarter of the crypt form the differentiation zone. (3)

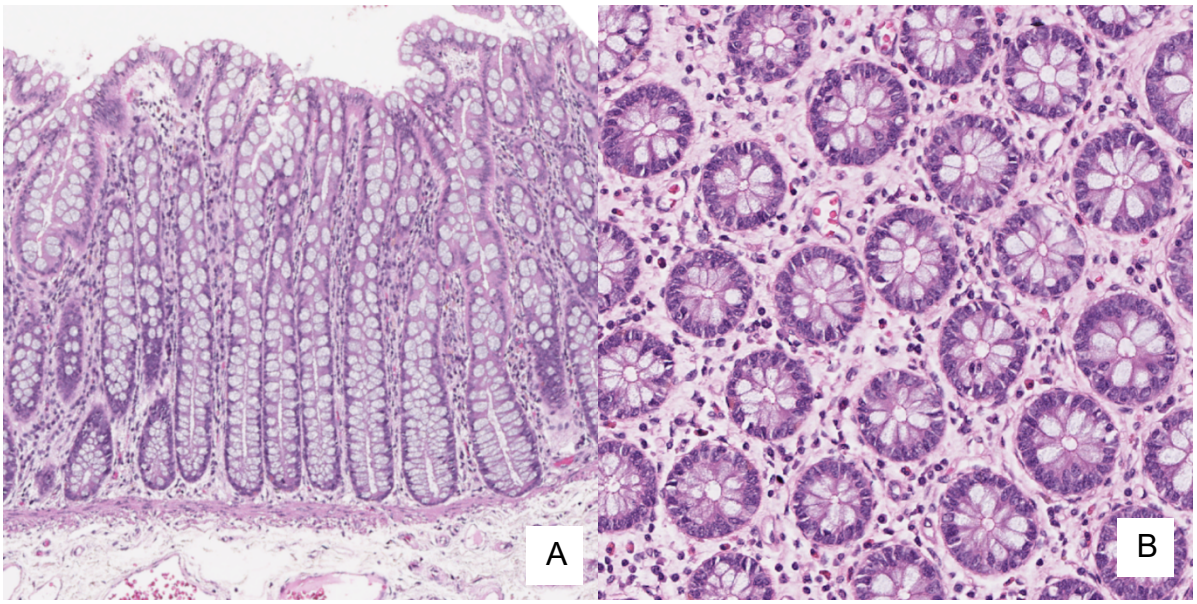

Figure 3: Normal colon mucosa. ***A*** *Longitudinal view.* ***B*** *Cross-section of colonic crypts.*

1.1.2 Classification of benign epithelial lesions of the colon and rectum

Colon polyps can be divided into a neoplastic and non-neoplastic group. The neoplastic group includes most commonly conventional adenomas (CA), serrated lesions and traditional serrated adenomas (TSA). They present precursor lesions to CRC. (7) Due to this fact, early detection and removal is important to prevent further progression into a malignant tumor.

The most relevant screening method for early diagnosis is colonoscopy, which is recommended for patients aged ≥50 years. (8) From a morphologic standpoint, these tumors can present as a polypoid or non-polypoid/flat lesions. (9)
In the last decades, three major signaling pathways responsible for the development of CRC have been found: the chromosomal instability (CIN) pathway, the microsatellite instability (MSI) pathway and the CpG island hypermethylation phenotype (CIMP) pathway (10) (3). They are characterized by a certain precursor lesion and specific mechanism of carcinogenesis and will be described in more detail in the following chapters. Approximately 5% of CRC arise in patients with inherited syndromes such as MUTYH-associated polyposis (MAP), Lynch syndrome (LS), and familial adenomatous polyposis (FAP). (10)

Non-neoplastic polyps include hyperplastic polyps, inflammatory polyps and

hamartomatous polyps, which in most cases do not have a malignant potential. Hyperplastic polyps are the most common colonic polyps. (11) Inflammatory polyps are comprised of epithelial and stromal components with a great number of inflammatory cells. This group includes prolapse type inflammatory polyps and pseudopolyps that arise in response to chronic inflammatory processes like Crohn disease or ulcerative colitis. (3)

Hamartomatous polyps consist of local tissues that grow in an unorganized mass. Most common lesions in this group are juvenile polyps with an increased number of dilated cystic glands and Peutz-Jeghers polyps characterized by smooth muscle proliferation with a distinctive arborizing pattern. A sporadic form of juvenile polyps is seen in 2% of children under 10 years of age. A hereditary form is associated with an autosomal dominant mutation that is causing the juvenile polyposis syndrome (JPS), where hamartomatous polyps appear in the gastrointestinal tract. Peutz-Jegher polyps usually develop due to *STK11* mutations resulting in Peutz-Jeghers-Syndrome (12,13)

1.1.3 Conventional colorectal adenomas

Conventional colorectal adenomas are benign epithelial neoplasms, which can progress to CRC. Epithelial dysplasia of low- and/or high-grade is typical for this entity. (3) After hyperplastic polyps, adenomas are the most common colonic polyps. (14)

1.1.3.1 Epidemiology, etiology, and risk factors

The epidemiology of CAs is similar to the epidemiologic data of colorectal adenocarcinoma. (15) The prevalence and incidence vary depending on the used detection method. (3) Colonoscopy screening studies show that in western industrialized countries the mean prevalence of colorectal adenomas is 25%. If the examined area is limited to the sigmoid only, the prevalence is lower (10%). (16) Age is an important risk factor for the adenoma formation. Autopsy studies have shown that only 1 to 4% of cases had an adenoma in the age of the twenty- to thirty-years. (17) By age 70, in up to 50% of all autopsies polyps could be detected. (18) Men are more commonly affected and 39% of all patients have more than one. (3) Risk factor important for the development of CAs is a diet, especially one that is high in calories and fat, but low in fiber and vitamins(3) On the contrary, physical activity

seems to be having a protective effect. According to Sanchez et al., the prevalence is significantly lower in the group that exercises one hour per week. (19)

1.1.3.2 Pathogenesis

In the late 1980s, Fearon et al. described new molecular findings that lead to the modern understanding of the colorectal carcinogenesis. (20)
For CAs, the development happens in sequential genetic alterations of oncogenes or tumor suppressor genes. The most important and earliest genetic aberration involves a mutation in the WNT signaling pathway, in detail an *APC* gene mutation. (21) Other driver gene alterations that lead to malignant transformation and propensity to growth are *KRAS*, *SMAD4*, *PIK3CA* and *TP53* mutations (see Figure 4). (15)

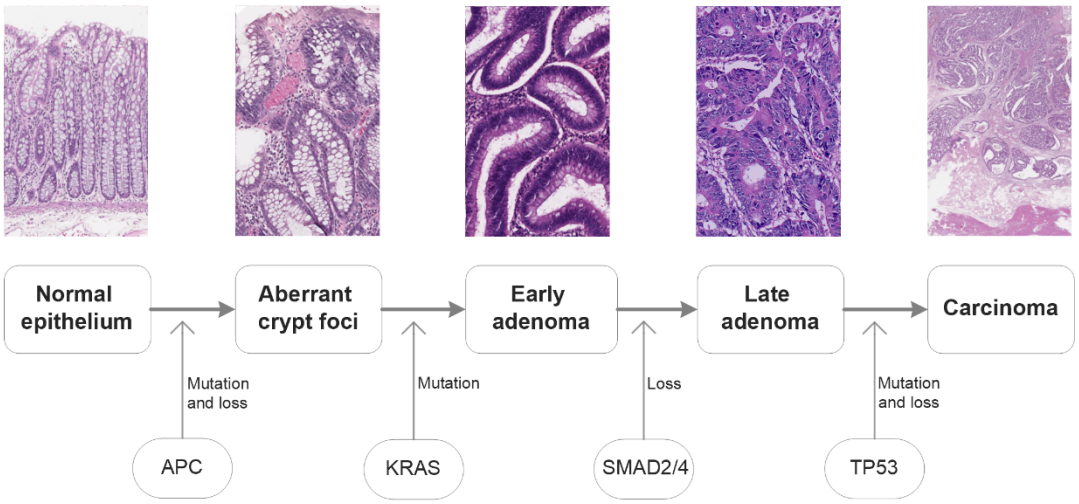

Figure 4: Basic illustration of the adenoma carcinoma sequence. *Own figure based on Remmele (2013; p. 646) (3)*

The central role of the APC protein in this carcinogenesis pathway is based on its involvement in key cellular regulatory processes: cell proliferation and differentiation, cell cycle regulation and apoptosis via the β-catenin signaling pathway and chromosomal segregation. (3)
In detail, when the β-catenin signaling pathway is not activated by WNT signaling, enzymatic phosphorylation, in which APC is also involved, results in inactivation of cytoplasmic β-catenin. However, physiological activation of WNT or inactivating

mutations of *APC* or activating mutations of *β-catenin* result in an accumulation of cytoplasmic β-catenin. (22) This protein travels to the nucleus and acts as an transcription factor in combination with a T-cell factor (TCF), which leads to an activation of specific proto-oncogenes like *c-myc* and *Cyclin D*. (3)
In addition, during mitosis APC is involved in the chromosomal stability. Due to genetic alterations of the *APC* gene chromosomal instabilities can lead to aneuploidy, deletions and translocations. (23) As a consequence loss of function of tumor suppressor genes (e.g. *TP53* or *SMAD2/4*) or gain of function of oncogenes (e.g. *KRAS* or *PIK3CA*) promote colorectal carcinogenesis. (3) As seen in Figure 4, for the early progression of aberrant crypt foci to small and eventually large adenomas, *KRAS* mutations play a central role. Dysregulations in the TGF-β growth inhibitory pathway due to deletion of *SMAD4*, activating mutations of *PIK3CA* or functional loss of *TP53* are important for the later progression of the adenomas and further malignant transformation to invasive carcinoma. (24)
Tumors that arise via the adenoma-carcinoma sequence are more likely to be localized in the left colon, especially in the rectosigmoid. (3)

1.1.3.3 Morphological appearance

CAs can be pedunculated, sessile, slightly elevated, flat or depressed. (15) Histologically, they are divided in tubular, villous or tubulovillous subtypes. The classification is based on the percentage of the villous architecture. (3) The epithelial proliferation zone of adenomas migrates from the crypt base to the surface, which leads to the characteristic top-down morphology. This architecture helps to distinguish dysplasia from reactive changes, due to a basal proliferation zone. (4)
Intraepithelial neoplasia can be graded in low-grade (LGD) and high-grade dysplasia (HGD). The epithelium of LGD adenomas show enlarged, hyperchromatic, and relatively uniform nuclei, which can be oval, or spindle shaped, but a nuclear polarity is retained. Furthermore, the relation of nuclear to cell volume is slightly shifted in favor of the nuclei and the number of goblet cells is reduced.
On the other hand, HGD is characterized by increased nuclear pleomorphism with prominent nucleoli, complete loss of nuclear polarity, often atypical mitotic figures, and architectural complexity including, irregularity of glands, cribriform architecture, and intraluminal necrosis. (3,4,15)

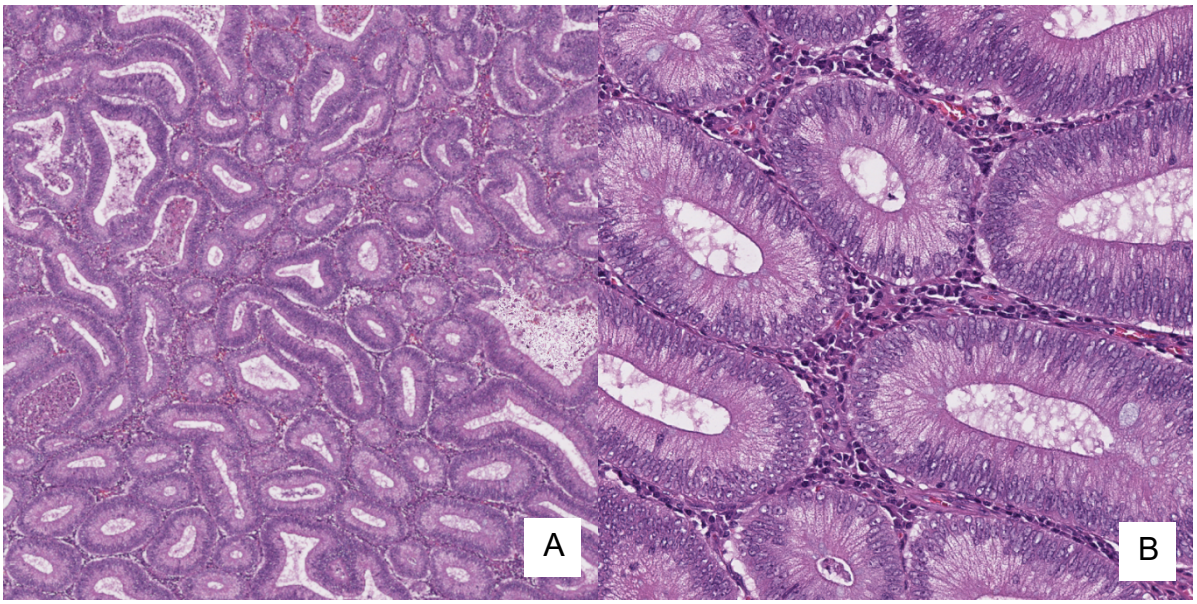

Figure 5: Tubular adenoma with low-grade dysplasia. *(**A** 10x, **B** 40x)*

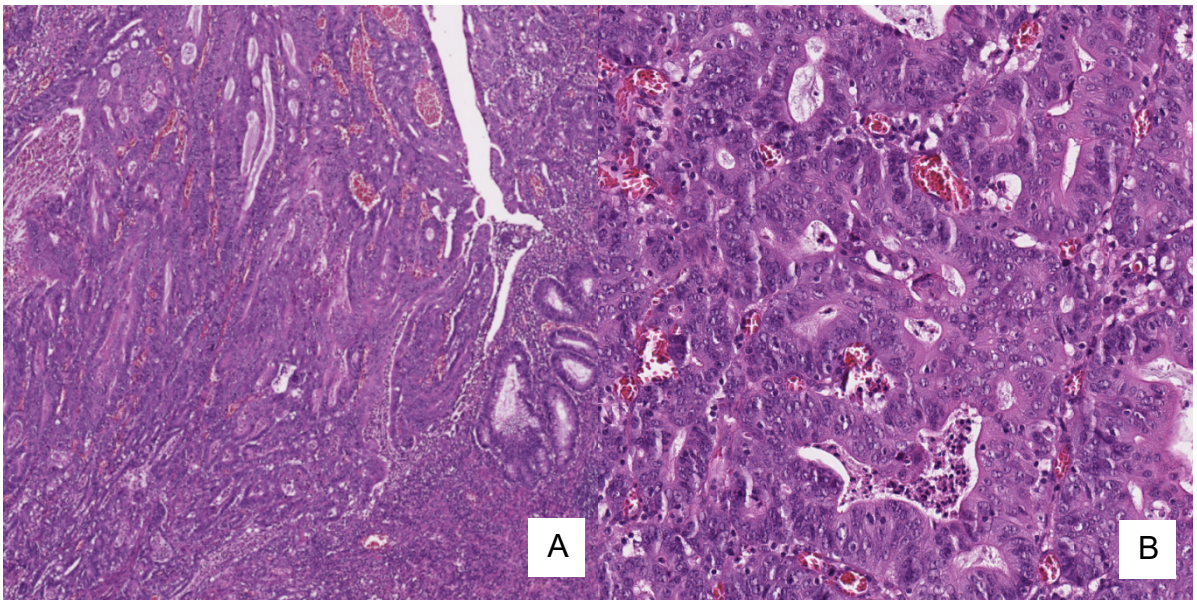

Figure 6: Tubular adenoma with high-grade dysplasia. *(**A** 10x, **B** 40x)*

1.1.3.3.2 Tubular adenoma

Tubular adenomas (TA) are the most common colonic adenomas. Approximately 70-80% of CAs are of tubular architecture. (25) Endoscopic findings can be presented with exophytic, flat or depressed growth. Histological characteristics are branched glandular structures embedded in the lamina propria. To be classified as a TA, the villous fraction must be under 25%. (3)

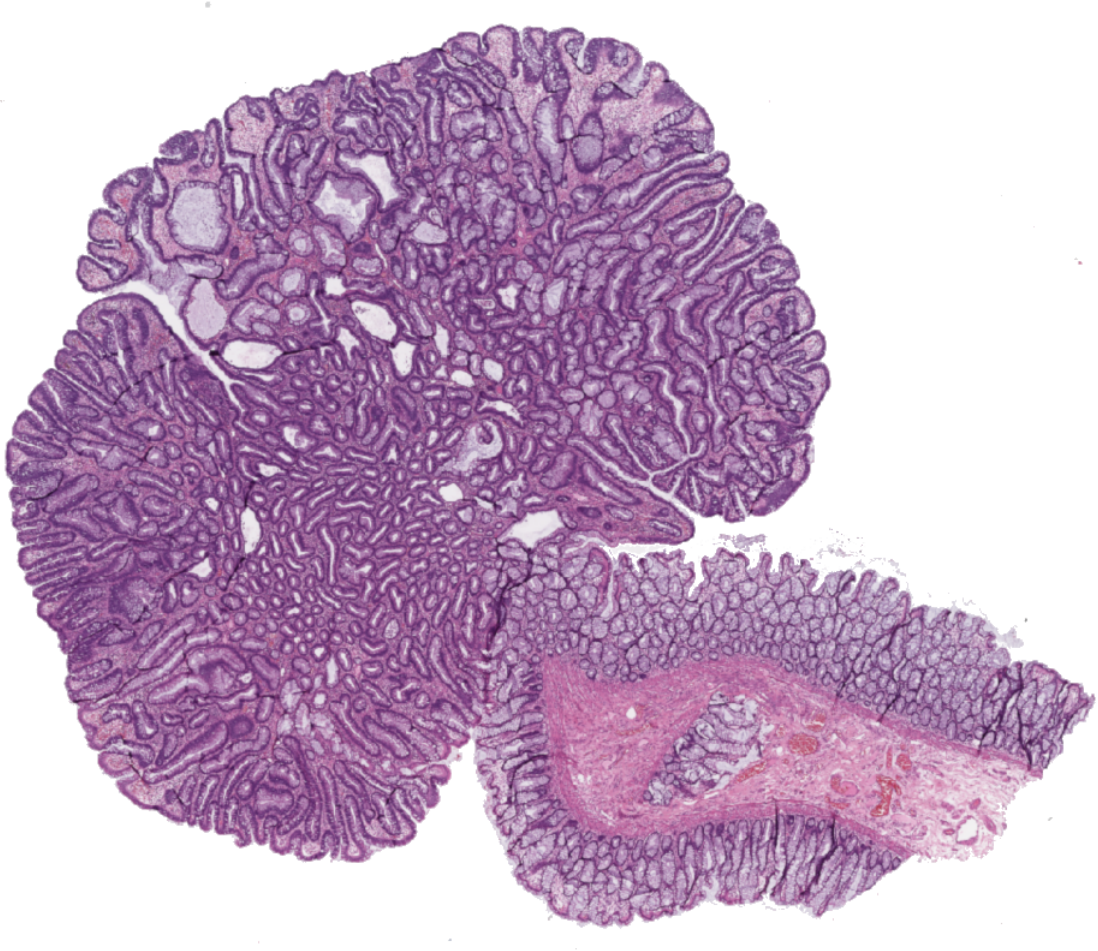

Figure 7: Tubular adenoma overview *(4x)*

1.1.3.3.4 Villous adenoma

Villous adenomas (VA) are more uncommon and comprise around 5-15% of all colonic polyps. (26,27) They can appear as sessile exophytic lesions and the surface resembles small intestinal villi. Under the microscope the villi are narrow, finger-like protrusions of the lamina propria which are covered by dysplastic epithelium. These elongated glands can reach from the surface to the center of the polyp. (3) For the diagnosis of VA, the villous component has to be above 75%. (15)

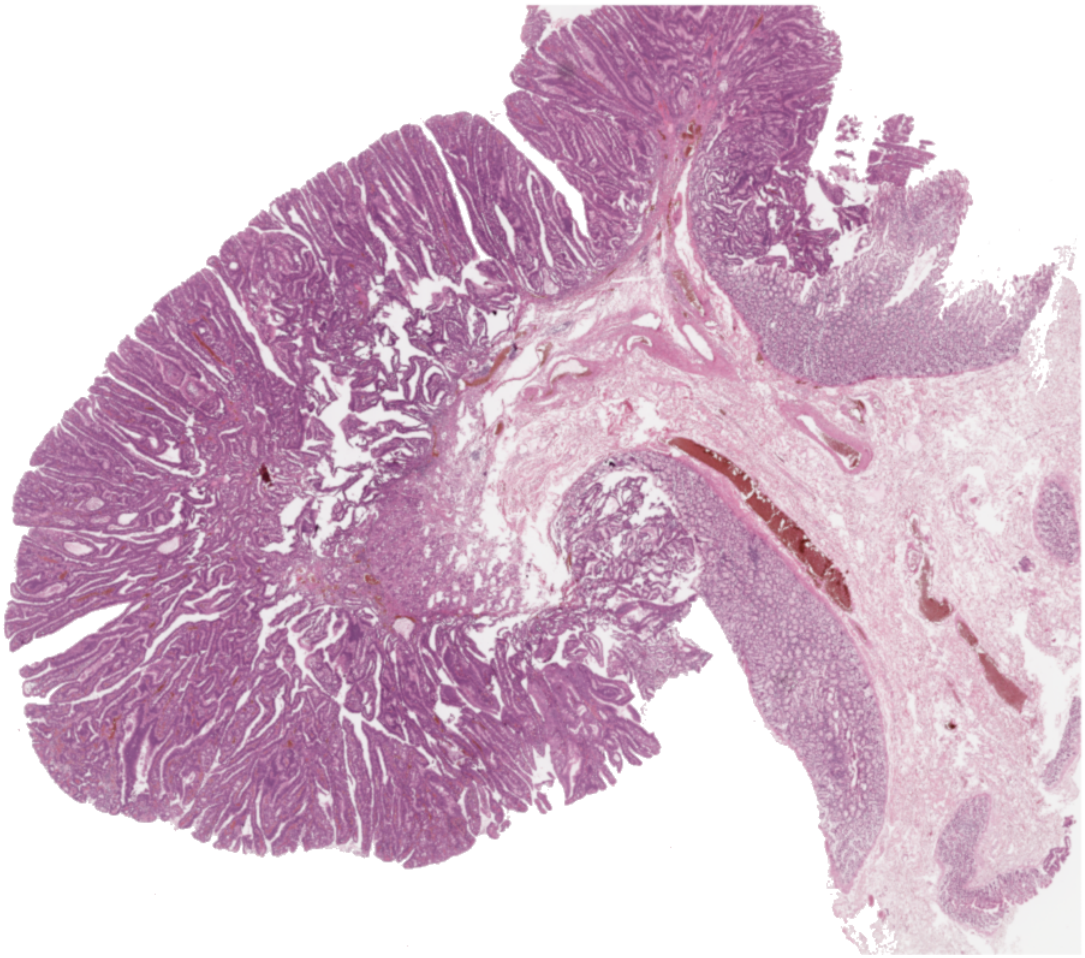

Figure 8: Villous adenoma overview *(2x)*

1.1.3.3.6 Tubulovillous adenoma

Tubulovillous adenoma (TVA) have a similar frequency as VA with 5-15%. In these polyps, the villous architecture must be between 25-75%. (26,27)

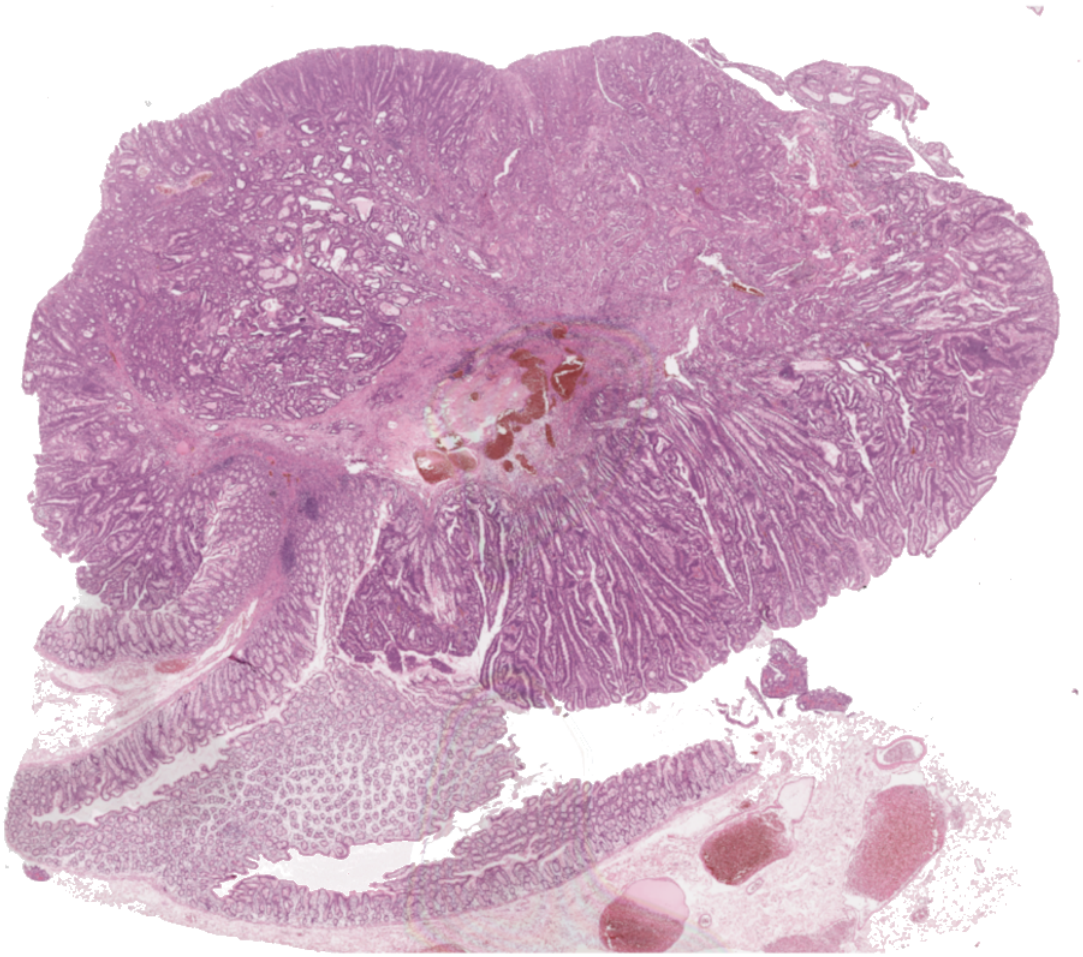

Figure 9: Tubulovillous adenoma overview *(2x)*

All these precursor lesions are confined to the mucosa. An invasion beyond the muscularis mucosae is indicative for progression to CRC. (4)

1.1.5 Serrated pathway precursor lesions

Besides the colorectal carcinogenesis via the adenoma carcinoma sequence, the serrated pathway gained considerable interest in the past decades. (28) Diagnosis and classification of serrated polyps can be challenging due to evolving nomenclature, changing histologic criteria and unclear prognostic factors. Serrated polyps summarize an heterogenous group of polyps varying in morphology, genesis, and possible malignancy. (29) The three distinguishable subtypes are hyperplastic polyps (HP), sessile serrated lesions (SSL), and traditional serrated adenoma (TSA). (3) HPs can further be subdivided into goblet cell rich (GCHP) and microvesicular (MVHP) variants. SSLs present most frequently without dysplasia, however cases with dysplasia and progression into an invasive carcinoma occur as well.

1.1.5.1 Epidemiology, etiology, and risk factors

Data on epidemiology and etiology are only possible to a limited extent, due to the changing terminology and taxonomy and varying detection and resection practices of serrated polyps. As already mentioned, the most common entity are distally located hyperplastic polyps, which are sometimes not biopsied. (4)
Serrated polyps seem to occur all over the globe, independent from location and ethnicity. (30) Regarding the prevalence, analysis of endoscopic and autopsy data can provide an estimate of the prevalence. Carr et al. showed in their case series study, that in an unselected group of adult patients undergoing colonoscopy screening, 35% of all resected polyps were serrated polyps. (31) Moreover, an overlap of the prevalence of SP and CA is observed, around 50% of patients with one or more SSLs have a CA at the same time. (32)
A true prevalence is difficult to quote as autopsy studies published varied in length and the distinction between different serrated polyp types was not clear. A low prevalence with under 10% is observed in different parts of Asia. In contrast, studies in the US and Europe reported a prevalence of up to 30%. (18,29,33)
Several modifiable and non-modifiable risk factors are found. Bailie et al. presented in their systematic review and meta-analysis a strong connection between smoking and risk for serrated polyps, especially SSLs. (11) Furthermore, an association with alcohol intake is probable, as several studies indicate an increased risk for SSLs. (34–36) In addition, other possible risk factors with less evidence are obesity, diabetes, supplementary folate and a high socioeconomic status. (30) Intake of non-steroidal anti-inflammatory like aspirin result in a significant lower risk of SSLs. (36)

Age and male sex are strong risk factors for CAs, while for serrated polyps these two seem to be less relevant. The risk for developing a SP in patients over the age of 50 is not strongly increasing, and men and women are both affected. As mentioned above, a consistent risk factor for a higher prevalence is European or North American origin. (33) Finally, family history of CRC or previous diagnosis of serrated polyps are significant risk factors. (34)

1.1.5.2 Pathogenesis

Similar to the adenoma carcinoma sequence, the serrated pathway summarizes a sequence of genetic and epigenetic changes that consequently lead to a further progression of serrated polyps. Initially alterations regarding the regulation of mitogen-activated protein kinase pathway, predominantly *BRAF* (in MVHP, SSL, and TSA) and in some instances *KRAS* (in GCHP and TSA), start the development of serrated polyps. (15) Cellular senescence happens as a result to a disruption of the apoptosis. The mutation of the *BRAF* gene is associated with epigenetic alterations, namely the hypermethylation of CpG islands, labeled CpG island methylator phenotype (CIMP). For this reason, tumor suppressor genes are silenced (such as *TP53* and *p16INK4a*) and malignant transformation is possible. Depending which gene is affected, high (MSI-H; e.g. *MLH1* gene) or low (MSI-L; e.g. *MGMT* gene) microsatellite instability develops. (3,30) Further progression of serrated polyps is induced through the activation of the WNT signaling pathway. Rather than the *APC* gene mutation in the adenoma carcinoma sequence, the WNT pathway is activated due to a mutation in the RNF43-ZNRF3 complex. (37) In *KRAS*-mutated TSAs a fusion of genes in the R-spondin family occurs, which leads to R-spondin overexpression, and consequently a down-regulation of RNF43. (38) A schematic representation of the serrated pathway is summarized in Figure 10.

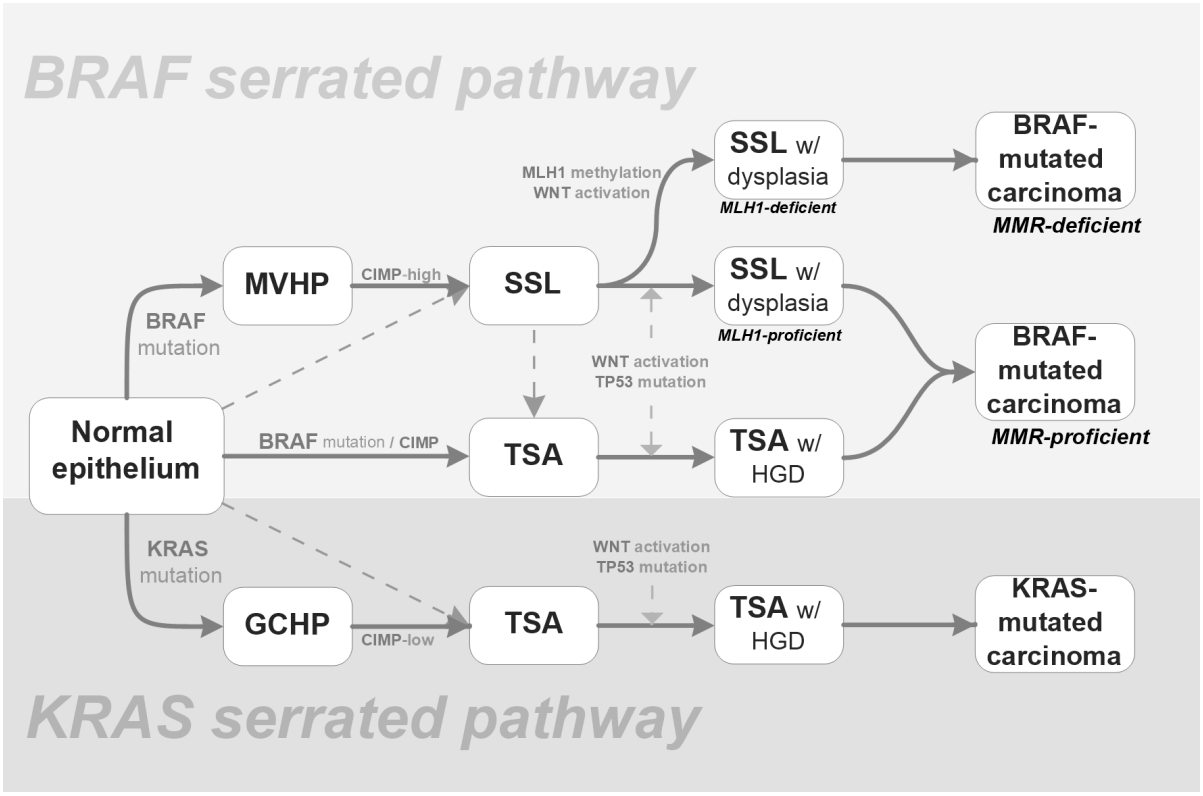

Figure 10: Schematic illustration of the serrated pathway. *Own figure based on WHO Classification of Tumours / 5th Edition / Digestive System Tumours (2019; p.163) (15)*

1.1.5.3 Morphological appearance

Over the past decades the terminology of serrated polyps evolved and got more complex since it's discovery as an important precursor lesion for CRC. Nowadays three entities are well-described and will be discussed in more detail in the following chapters. (28)

1.1.5.3.1 Hyperplastic polyps

HP are the most common serrated polyps, predominantly located in the left colon with a diameter of less than 5 mm. They can be classified into MVHP and GCHP subtypes, whereas the mucin poor HP isn't considered a separate subtype anymore (is considered to represent damaged MVHPs).(4) The most frequent subtype is the MVHP and it occurs after *BRAF* mutations, as shown in Figure 10. (15) GCHP mostly develops after *KRAS* mutations and is mostly found in the rectosigmoid. (3) Histologically, HPs are characterized by elongated, funnel-shaped and evenly spaced crypts. The surface resembles a pattern like a serrated knife, arising due to an early *KRAS*/*BRAF* mutation and reduced apoptosis, which leads to an excess of epithelium cells folding to endoluminal papillary structures (Fig. 11). While at the crypt base a

pseudostratification of the cell nuclei can be observed, the epithelium towards the surface matures normally with small, oval and basally located nuclei. Any nuclear atypia is uncommon. (15,30) MVHP are characterized by microvesicular epithelium cells and plentiful cytoplasm. GCHP have subtler morphologic changes like taller crypts, slight serration, and an increased number of goblet cells. The goblet cell rich subtype is mostly found in the rectosigmoid. (3)

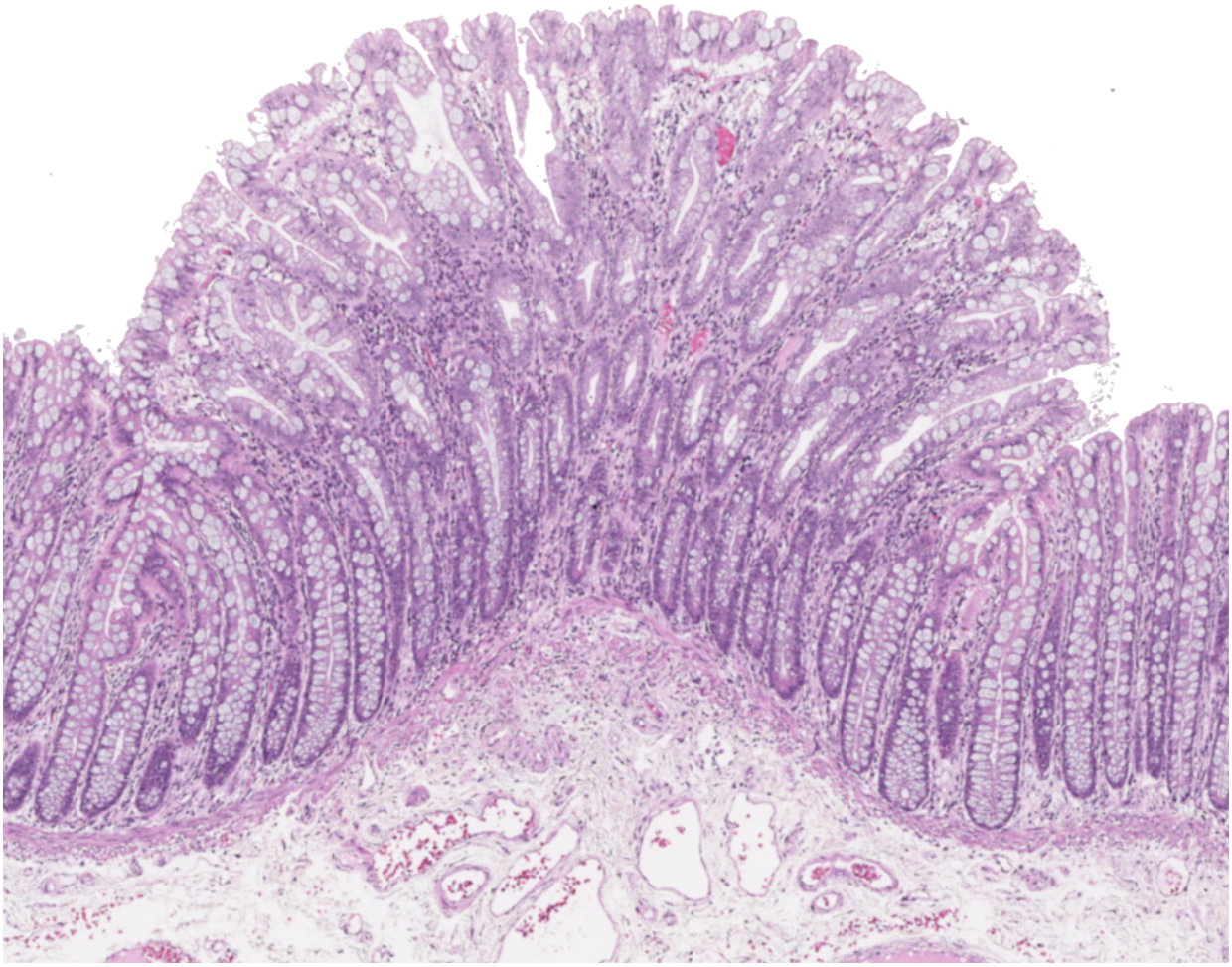

Figure 11: Hyperplastic polyp. *Microvesicular subtype*

1.1.5.3.2 Sessile serrated lesions

The second most common subtype of serrated polyps are SSL, with a proportion of around 11% of all serrated polyps. (4) They are mostly located in the right colon, larger than 5 mm and their endoscopic appearance is best described as sessile, i.e., flat, lightly elevated and they do not protrude into the intestinal lumen. (3) The differentiation to the other serrated subtypes is based on the architecture and cytological characteristics, due to a shifted proliferative zone. Histopathologic diagnostic criteria

are: minimum of one unequivocal distorted crypt, with horizontally branched and dilated crypt base. Other architectural characteristics include continuous serration to the base, thinned out muscularis mucosae and herniation of crypts through the muscularis mucosae (Fig. 12A-C). (15)
Although the prevalence of dysplasia is rather low, about 4-8% of all SSLs have dysplastic alterations (SSL-D). (33) The awareness of the possibility of a malignant transformation is important because SSL-D can progress to CRC. Rather than the dysplasia in CA, the morphologic changes are very heterogenous (Fig. 12D). (30) Three morphologic subtypes are recognized, often occurring in the same SSL. Firstly, serrated dysplasia is the most common one, which exhibits following features: eosinophilic cytoplasm, tightly packed glands, and possible progression to TSA. Cytologic changes like nuclear atypia and numerous mitoses are more pronounced. (39) Intestinal dysplasia looks like the dysplasia in CA but is relatively uncommon. MLH1 staining is unchanged and a progression to carcinoma is questionable. The final pattern observed is characterized by minimal deviation dysplasia, which exhibits subtle alterations in comparison to SSL without dysplasia, but is distinguished by the complete absence of MLH1 staining. (3,15)

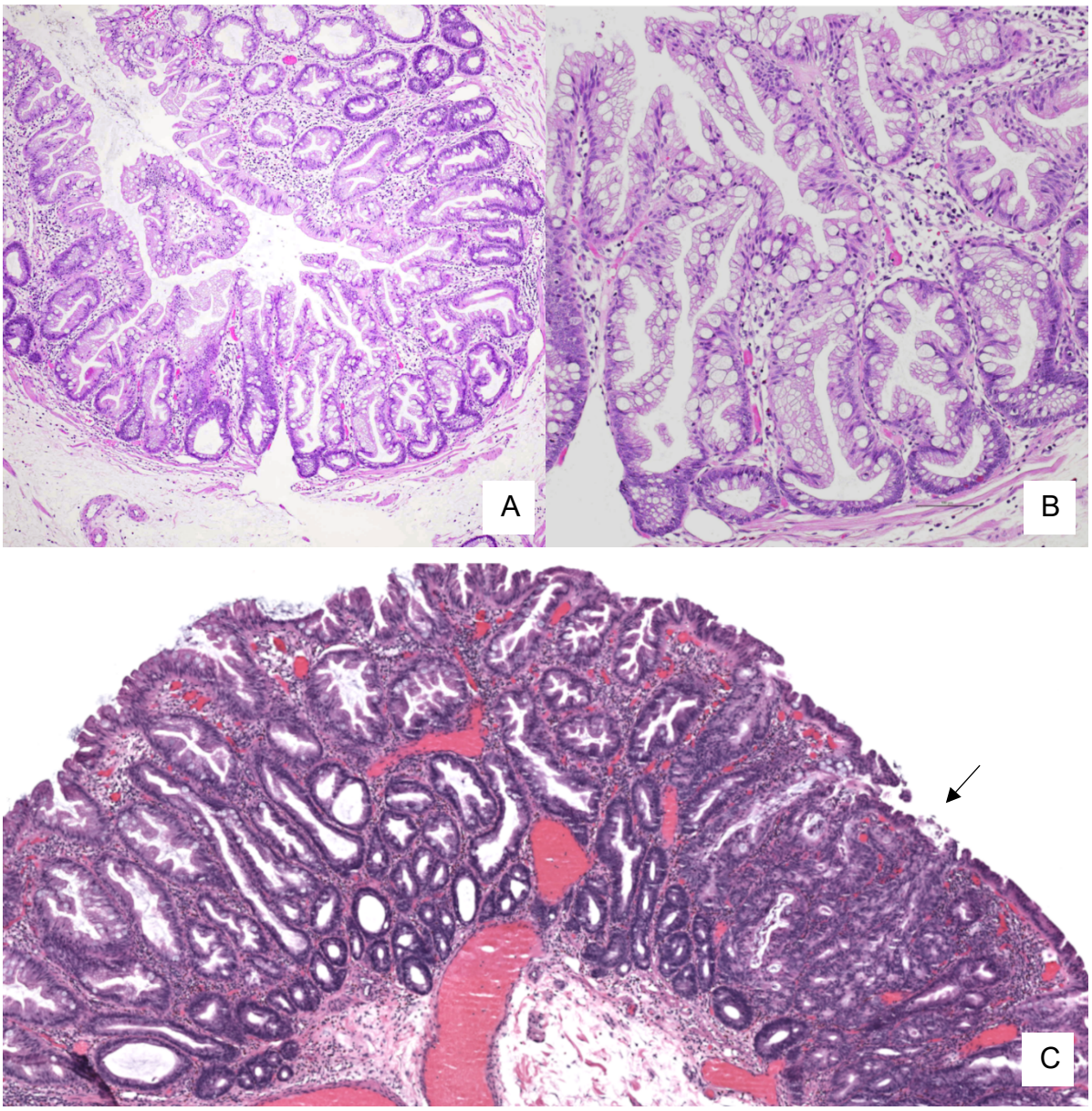

Figure 12: Sessile serrated lesion. *(**A** SSL without dysplasia 4x, **B** SSL without dysplasia 10x, **C** SSL with dysplasia (arrow). 10x)*

Traditional serrated adenoma

The TSA is unfrequently found (about 1% of all serrated polyps). On colonoscopy, they can be polypoid protuberant, mainly located in the distal colorectum, or flat usually in the proximal colon. (15) TSA combines the sawtooth-like architecture of HP with dysplasia characteristics of CA (Fig. 13). In more detail, penicillate nuclei, ectopic crypts and eosinophilic cytoplasm are the most distinctive histologic features. They can develop after KRAS or BRAF mutations. Half of all TSAs arise from other serrated polyps, as seen in Figure 10. (40)

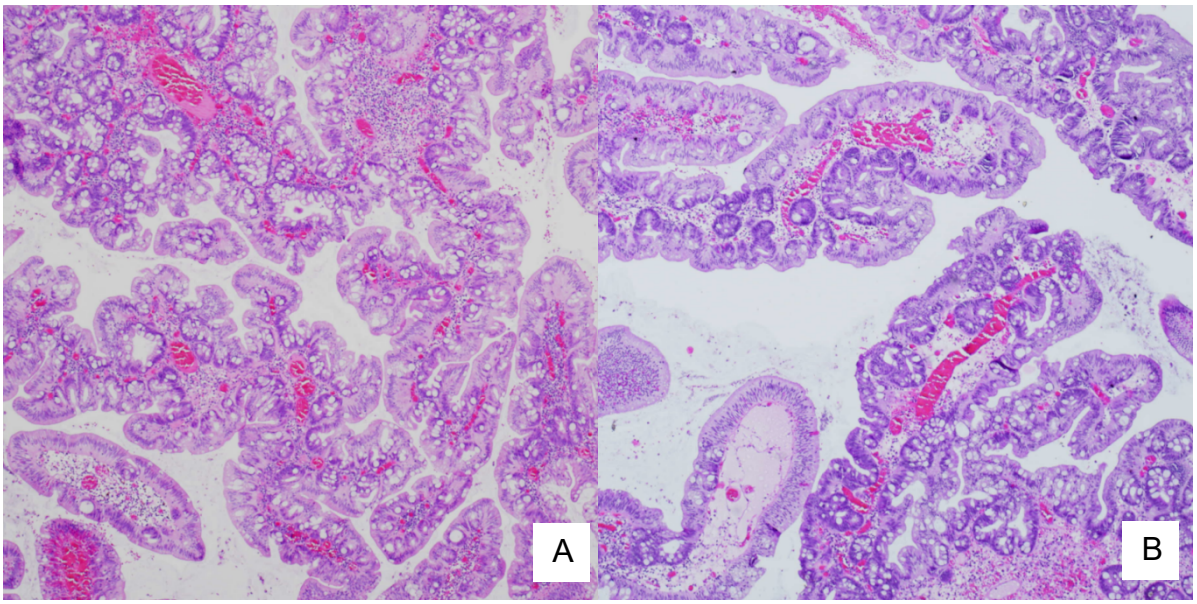

Figure 13: Traditional serrated adenoma. *(**A-B** TSA with low-grade dysplasia 4x)*

1.2 Digital Pathology

Digitization in medicine is an ongoing process with the goal of improved care, making infrastructures more efficient and development of new diagnostic and therapeutic options. This change is driven by technological achievements over the past decades. New priorities are set to meet new needs, focusing on collecting and handling large amounts of data. Each medical specialty has different possibilities to implement digitized environments and workflows, with varying standards and technical demands. (41)

1.2.1 Telepathology

First described in 1986 by Weinstein, the definition of telepathology (TP) in the early days was the practice of pathology at long distances. (42) Over time, the scope of TP got more complex due to technical advancements, which led to the development of whole-slide imaging. Telepathology was the first application of digital pathology in clinical use, with an infrastructure to view images on a monitor at distance rather than through a light microscope on site. (43) The architecture of a telepathology system is straightforward: a digital imaging setup to capture images, a data connection for transfer, and a monitor for evaluation. Further use can be static, in which the images are saved on a shared server or sent directly via email, or dynamic, where the slides can be examined live with a motorized microscope and a constant video transmission, or hybrid, with a merge of both static and dynamic components. (44) To operate a telepathology solution, a competent host on site is vital to operate and support the imaging of slides. While the downside of static telepathology is the limitation of exploring the histologic slide, dynamic settings need a highly trained host, to ensure an uninterrupted use for the pathologist, which can examine the slide independently via a remotely controlled robotic telepathology system or a locally controlled by the host. (45)

In some cases, TP is employed in surgical pathology, such as frozen section consultation, or niche areas, like hematopathology, transplant pathology, and ultrastructural telepathology. (45,46)

1.2.2 Whole-slide imaging (WSI)

The everyday tool for pathologists, the conventional light microscope, played an essential role in the past decades and is still important for a proper diagnostic workflow with histopathology as a diagnostic gold standard. The idea of a virtual microscope

and digitilizes glass slides transformed the field of digital pathology but was held back by the missing technological advancements and regulatory barriers. (45) Eventually, new innovations, especially relating to computer processing power, data transfer, software, and cloud usability, paved the way for the implementation we are seeing today with the creation, processing, analysis and exchange of data and images on a larger scale. (47)

In general, WSIs are scanned histopathologic slides that are stored in pyramidal (multi-scale) format (see Figure 14). The resolution of the WSI specified by the optical resolution, i.e., the magnification level of the lens, and spatial resolution, which is defined by the minimum distance a scanner captures two distinguishable objects measured by µm/pixel. As of today, WSI scanners can reach a spatial resolution of 0.11 µm/pixel to 0.25 µm/pixel at a magnification of 40x. Applied on tissue samples in surgical pathology, which are on average 15mm x 20mm, the scan can be the size of 200 000 x 200 000 pixels, resulting in an image file several gigabytes of size. Due to the pyramidal format, the pathologist is capable of examining the tissue in different magnification levels similar to a conventional light microscope. (48,49)

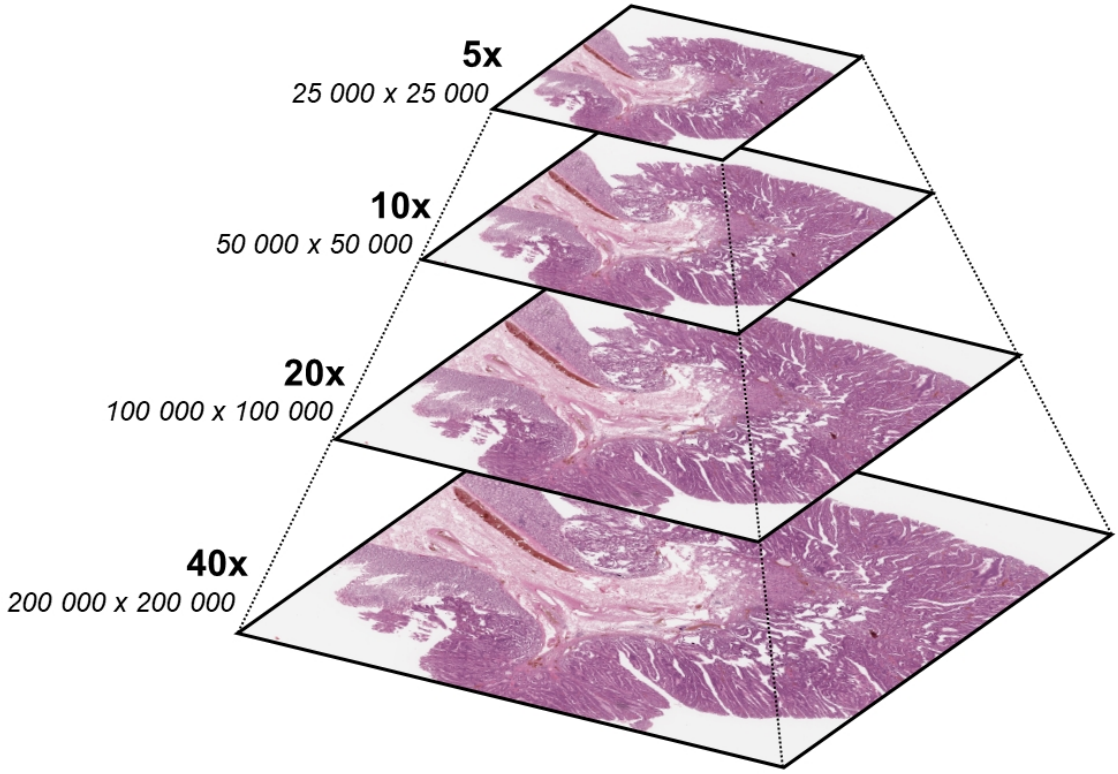

Figure 14: An example of a WSI stored in pyramidal format, *which includes multiple magnification levels and its respective spatial resolution. The image with the highest resolution is the base of the pyramid. Own figure based on Marini et al. (2021) (48)*

1.2.2.1 Technical requirements

WSI is a multistep process to visualize histologic slides. A synergetic solution between hardware and software results in a functioning process of image acquisition, storage, processing, and display of the scanned areas. (50)

WSI scanners handle the first step of this process. The first implementations 20 years ago were not feasible for the goal of high-throughput scanning, which is necessary for clinical use. The first system was priced at $300,000 and could scan one slide every 24 hours. (45) Nowadays, the hardware is more time and cost efficient and can produce high-resolution images in 30 seconds to a few minutes and the WSI scanner with the largest capacity currently (i.e., 3DHISTECH P1000) can handle up to 1000 slides at a time. (51)

Modern scanners are essentially monocular or trinocular microscopes under robotic controls which modify several factors like illumination intensity and focusing facilities. (50) As described in the paper published by Iyengar the main components of a WSI scanner are ”a) microscope with lens objectives, b) light source (bright field/fluorescence), c) robotics to load and to move the slides around, d) digital camera(s) that capture the image, e) computer, and f) software to manipulate, manage, and view the digital slides”. (51) Contrary to TP the WSI hardware is capable of scanning sequential areas and merging them into a virtual slide, which is beneficial for the quality of the image and produces an accurate replica of the slide (see Figure 15). It is possible to alter the used scanning resolution, depending on the used objective and the quality of the photo sensor. As mentioned above, the resolution and file size are directly proportional and depending on the used commercial WSI scanner, the

resulting file can be several gigabytes of size. Therefore, data storage and fast accessibility are challenges to overcome for routine use. (51,52)

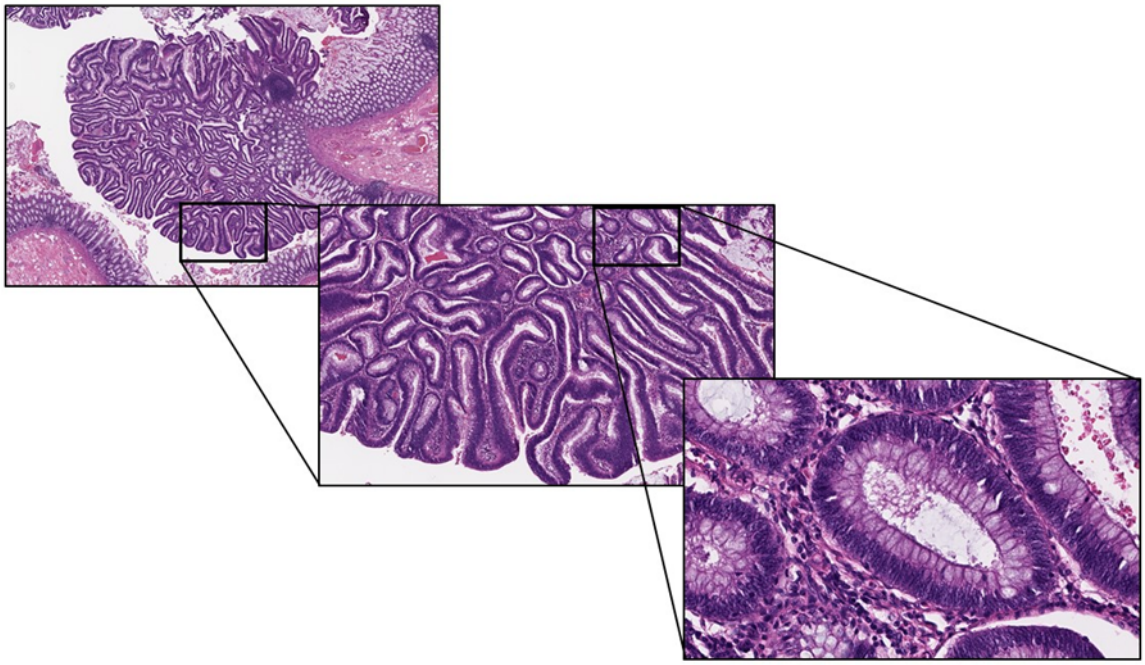

Figure 15: Example of zooming capability, *by merging high-resolution sequential scans.*

WSI scanners must utilize not only the x- and y-axes, but also the focus with the z-axis, to create a similar quality like the direct view through the microscope. Different approaches to z-focusing are possible, like focusing every tile, just a few selected tiles or even a multilayered approach, where multiple focus layers are used to combine all in-focus regions of a slide. (50,53)

The provided software packages are aligned accordingly, depending on the use of the WSI system, regarding viewing, management, analysis, and sharing of the virtual slide. Even artificial intelligence applications and algorithms for scoring and image analysis can be implemented in the software, which will be described in more detail in the following chapter. Besides more complex tasks, the image viewer should provide a similar experience to the handling of a real microscope, for example, to navigate slides and zoom-in on regions of interest. Other tools include direct slide annotation (for taking measurements, marking parts of the slide, and adding notes), synchronized viewing and sharing capabilities. (51)

The applications for WSI are diverse. An early introduction was seen in research and thereafter for teaching purposes. To implement digital pathology for routine diagnostics, several obstacles must be cleared, like building a cost-efficient

infrastructure, that can be integrated into the daily diagnostic workflow. Furthermore, regulations must be met to be able to use it in a clinical setting. (44) In addition, WSI is an enhancement for expert consult, because glass slides don't need to be transported to the location of the specialist. The cases can be digitally transferred in whole with attached clinical data with direct communication between the expert and the submitter. (51) In addition to patient-oriented usages, teaching and research benefited immensely by WSI. Virtual slides can be shared for educational purposes, and they don't degrade over time when archived. A variety of applications can be used for textbooks and journals. (50)
WSI is the steppingstone for a new revolution in the field of pathology. Virtual slides can be used for the development of diagnostic algorithms utilizing the latest advancements regarding artificial intelligence.

1.2.3 Computational pathology

Computational pathology (CPATH) is the incorporation of computer sciences in the field of pathology. Various forms of data, for example image data retrieved from WSI and the corresponding meta-data of the patient, such as multi-omics data, can be used for artificial intelligence (AI) and machine learning (ML) techniques. (54)
Like the main objectives of digitization in medicine, CPATH aims to optimize patient care at a reduced cost and improve the diagnostic accuracy, specifically regarding pathology, to minimize the intra- and interobserver variability, and to move in the direction of personalized medicine. (55)

1.2.3.1 Artificial intelligence and Deep learning

In the 1950s Dr. Alan Turing established the concept of AI. Later in 1955, Dr. John McCarthy tried to accomplish the development of a humanlike intelligence, that could solve simple puzzles and games, but progress was limited by the missing computational power. (56)
With the innovation of expert systems, which concluded knowledge of experts in a concept of logical rules for problem solving, AI experienced its second boom in the 1980s. (57)
Due to technological advancements in the areas of data gathering and computing technology, the realization of AI methods got accelerated after the year 2000. Nowadays, AI is a branch of computer sciences and is used to develop algorithms with multiple use cases, with a potential implementation in the field of pathology.

Depending on the degree of intelligence, AI can be classified as weak AI or strong AI. Weak AI is trained to execute specific tasks based an established statistic model, while strong AI is more like a general intelligence, which is capable of independent creation of a system, by performing ML methods on available data. (54)

1.2.3.1.1 Machine learning

In general, ML is a branch of AI that refers to systems where computers can process representative data and learn from it repeatedly. This enables further interpretation and enables the system to act accordingly, all without human guidance. (54) Various approaches to ML show potential and can be used to assist the pathologist in the diagnostic process by utilizing histomorphologic patterns found within cells or architectural structures. In detail, the system creates predictive models to determine patterns and, in addition, statistical methods like classification or regression can be executed. ML algorithms can be divided into supervised and unsupervised learning methods. Supervised learning is defined by using a labeled dataset as a ground truth and training algorithms that pair associated features and further classify the data. After training, the algorithm need to be validated with an independent dataset to evaluate the overall performance. (56) Specifically, in pathology, the data is often retrieved by annotations on WSI, that label various histologic structures like cellular components (e.g., nuclei of different cell types), regions with or without cancer cells, or whole areas of tumors. (54)
Different challenges must be overcome, such as needing a large quantity of annotations to achieve a viable accuracy, reducing the subjectivity of annotations, having access to preprocessed data without many disruptive factors (e.g., quality of the WSI and areas of staining) and in general the complexity of the algorithm. (56)

On the contrary, unsupervised learning does not utilize labels. Algorithms try to identify clusters on unlabeled datasets based on similar properties or hidden patterns and don't have the goal of predefined outcomes. This makes the unsupervised approach more exploratory and usually does not need human intervention. Often the goal is to reduce the data into fewer dimensions or features with techniques like principal component analysis or t-distributed Stochastic Neighbor Embedding (t-SNE). (58)

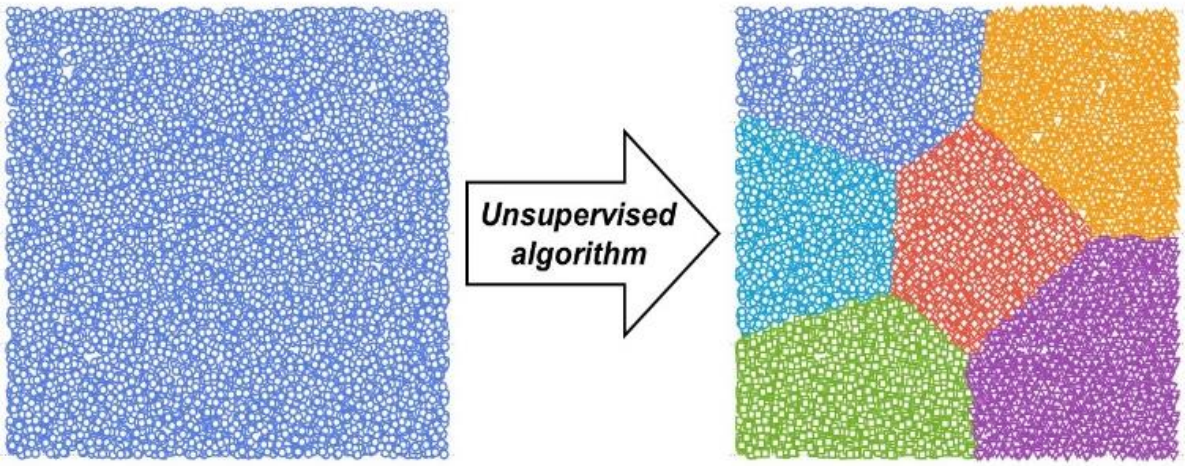

Figure 16: Schematic illustration of an unsupervised dimension reduction. *Own figure based on Sidey-Gibbons (2019; p. 5) (58)*

1.2.3.1.2 Deep learning

Deep learning (DL) is a specialized branch of machine learning (ML) that employs intricate neural network architectures consisting of an input layer, multiple hidden layers, and an output layer, comprising numerous artificial neurons (see Figure 17). These are linked with each other, and the strength of the connection is determined by different clustering methods, such as K-means clustering or other statistical methods. In contrast to ML, the algorithms are more adaptive and can process more complex data without any bias regarding labels and learn features totally independent. This makes ML algorithms more transparent and easier to understand, but they sometimes struggle to achieve a similar accuracy to deep neural networks. (54,56) However, DL algorithms are difficult to comprehend and often described as a “black box”, which makes them in some instances unpredictable and not easily suitable for clinical use. (7)

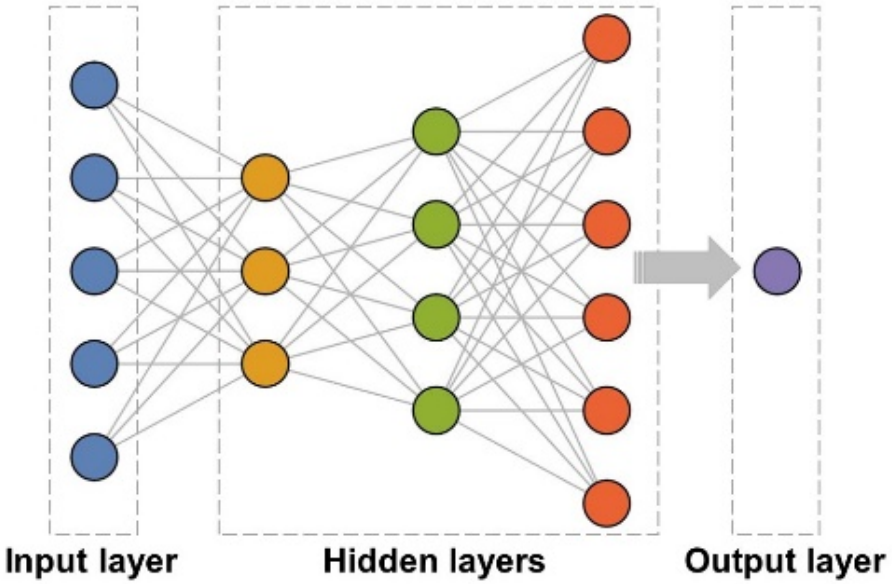

Figure 17: Basic structure of a deep neural network.

In the context of pathology and WSI, convolutional neural networks (CNN) are used, as they are designed specifically for image analysis, which makes them more suited for classification tasks and image recognition. The input image is modeled to a spatial structure using convolutional operations, by creating convolutional layers and reducing the dimensions by generating a pooling layer, with different pooling methods available (see Figure 18). (54)

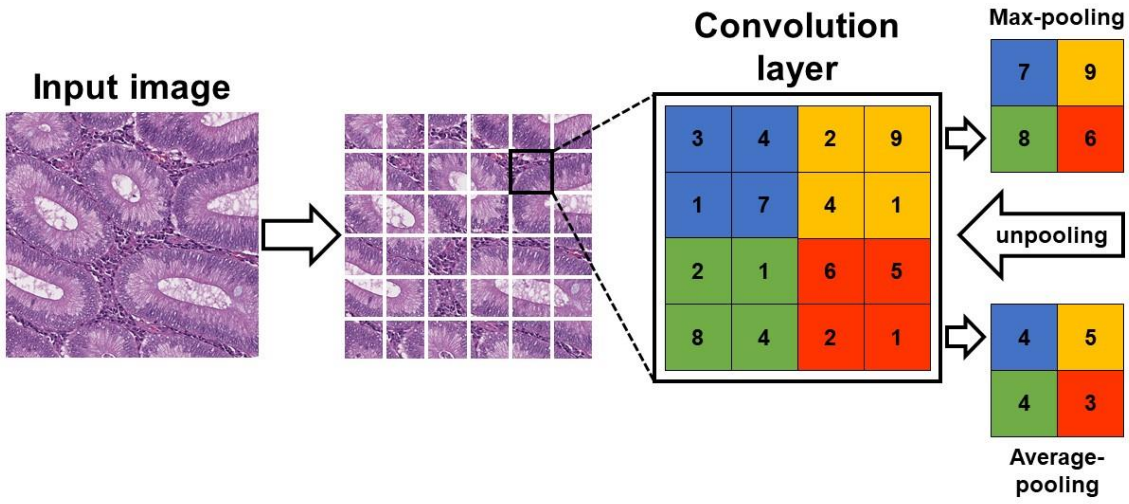

Figure 18: Structure of convolutional neural networks. *Own figure based on Cui (2021; p. 414) (54)*

Developers can design their CNN by varying the number of used neurons and layers, with the complexity of the network often correlating with the difficulty of the task. Moreover, complex systems need a large amount of training data to prevent prediction errors and achieve their potential. (56)

1.2.3.1.3 Applications of AI in pathology

Multiple implementations of AI have the potential to be significant in routine diagnostics after several hurdles are overcome. As mentioned above, a major challenge is the introduction of a digitized workflow and a sophisticated digital infrastructure. Furthermore, well-annotated datasets are crucial for the development of algorithms, which rely on a predefined ground truth for a standardized annotating process. (54) Examples of AI applications in pathology can be in detecting and subtyping histologic cancer regions, patient prognosis on the base of a prediction of the given clinical values, similarity searches as a diagnostic aid, quantification processes for tumor cell counts, or the extraction of relevant features. (56)

1.3 Annotation Workflow

With the recent technological developments of WSI, regarding high throughput scanning capabilities and the subsequent digitized archiving in tissue banks, computational models can be integrated in the diagnostic workflow of pathologist and assist in histopathologic analysis, prognosis, and prediction. (7) For AI-based analysis of WSIs multiple approaches of machine learning algorithms are feasible, which require proper annotations of the slide and of the underlying tissue and cellular structures. Quality of the annotations is directly linked to consistent and accurate results and lead to more comprehensible algorithms. The annotation process is labor-intensive due to large volumes of data needed to create a sufficient amount of well-annotated data. Recent approaches in ML, in particular self-supervised learning, transfer learning and domain adaption, made it possible to train algorithms even with small amounts of annotated data. A well-defined approach to the annotation process is crucial for the annotation quality and subsequently the success of computational pathology projects. Currently there is a lack of best practice recommendations of annotation projects, however a recent publication by Wahab et al. is proposing a workflow for semantic annotation in computational pathology projects, see figure 19. (59)

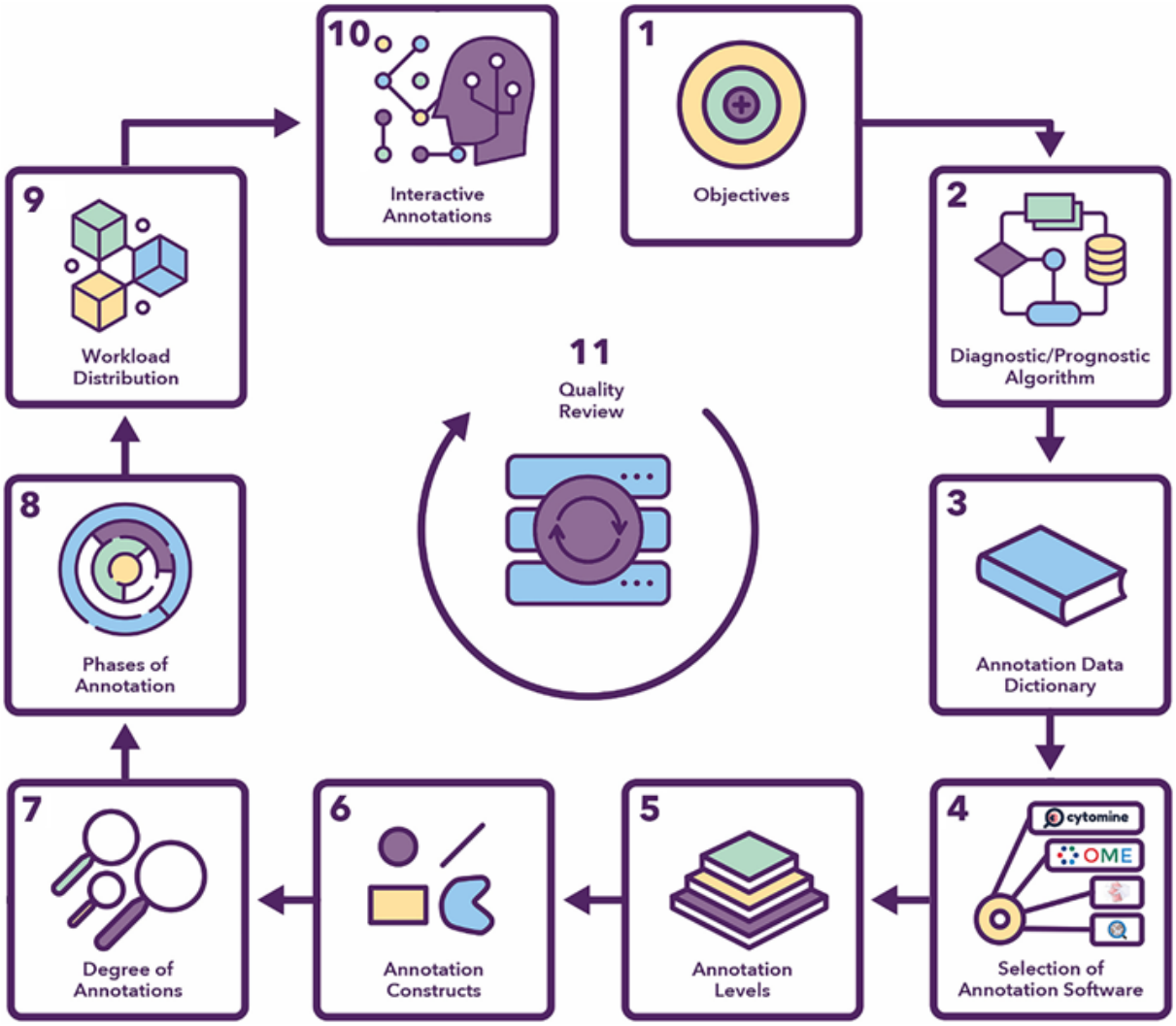

Figure 19: Proposed annotation workflow for a CPath project. Taken from The Journal of Pathology CR, Volume: 8, Issue: 2, Pages: 116-128, First published: 10 January 2022, DOI: (10.1002/cjp2.256) *(Licensed: CC-BY-NC-SA 4.0)* *(59)*

The basis of the development of AI algorithms in CPATH is the definition of project objectives. (59) For this project the specific objective is to create an algorithm for classification of colon polyps, therefore the regions of interest are histologic structures of the colon, which are further specified the upcoming chapter and noted in the annotation protocol.

1.3.1 Software

An important part for the labor-intensive annotation process is the selection of a user-friendly software that covers all project specific needs and can be implemented in the current infrastructure. Several factors regarding the storage of the annotations and associated meta-data, support for different image formats, security of the system, upgrade capability, and software-specific functionalities should be considered. (59)

1.3.1.1 QuPath

The software QuPath is the annotation tool, first introduced in 2017. It is an open-source software for digital pathology and offers comprehensive tools for the analysis and evaluation of WSI. In general, open-source tools are a driving force in research. In this case, open resources impacted the development of tissue-based analysis immensely with the exchange of customized solutions in the form of scripts, plugins or workflows. Preexisting open-source biomedical image analysis software, such as ImageJ, benefits from open-source packages, e.g., Fiji, and the implementation of plugins for scientific image analysis, but struggles with the size of WSI. Therefore, QuPath was developed to meet the evolving demands in digital pathology and as a result emerged as one of the most used bioimage analysis software. (60,61)

1.3.2 Annotation Levels

Different levels of detail of annotations should be defined (e.g., case/slide-level) to achieve the aims and objectives of an ML project. At case-level, labels are assigned to a case (e.g., benign vs malignant). At region-level, different regions in the WSI can be annotated (e.g., epithelial tissue, stroma). Cell-level annotations are more detailed, while descriptive and multi-modal annotations include pathology reports and genomics/transcriptomics. (59,60)

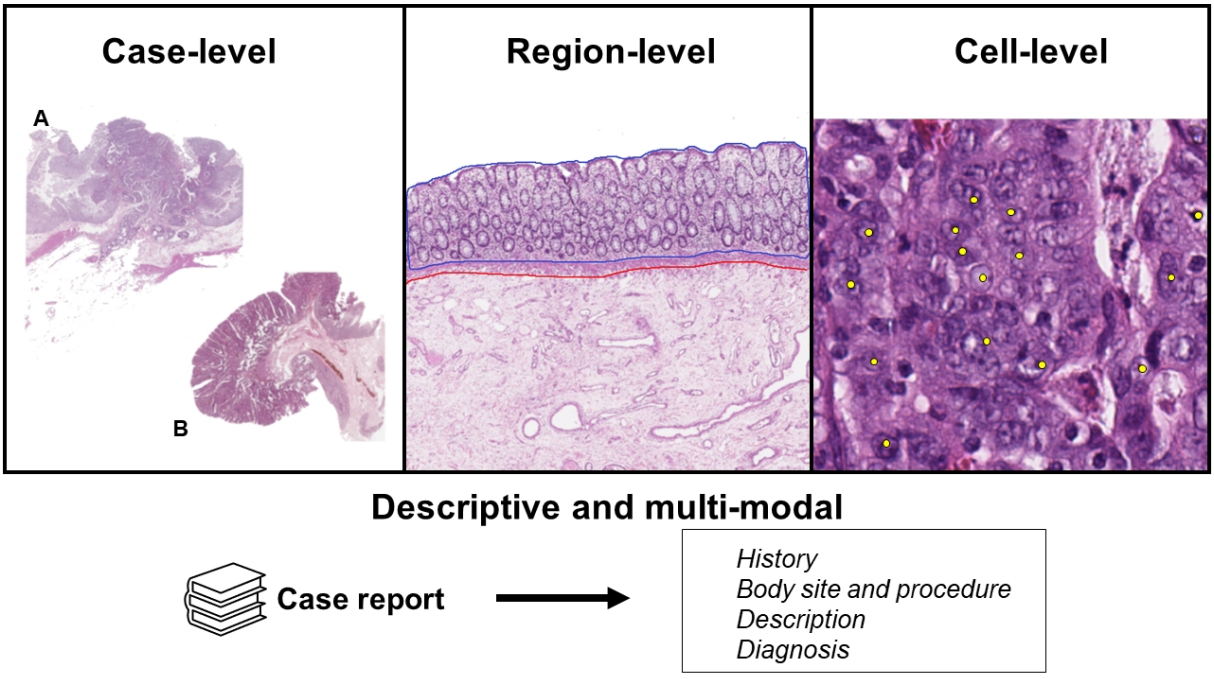

Figure 20: Levels of annotation. Case-level annotation, **A** representing invasive adenocarcinoma of the colon and **B** a colonic polyp. **Region-level annotation**, e.g., free-hand polygon-based annotation of colonic mucosa (blue) and submucosa (red). **Cell-level annotation**, for example point-based annotation of dysplastic cells.

Labels can be allotted to cases and their corresponding slides, or to each slide individually, depending on the underlying task. These labels are either binary (e.g. benign/malignant) or multi-class (such as grade 1, 2, 3). Weakly-supervised ML models, based on deep convolutional neural networks, don't require detailed annotations, and can be trained with less complex annotations, resulting in case-level predictions. (59)

In comparison to case-level annotations, region-level annotations mark other more specified regions of the WSI. This distinction may be of prognostic or diagnostic interest. In context of the downstream analysis, it is possible to have a varying degree of granularity and merge a number of region types later on and consequently the annotations can be utilized to train region segmentation models. For ML training/validation, annotated patches of the region-of-interest can be generated. Typically, polygons around areas or bounding boxes are used as annotation constructs for region-level annotation. (59)

Cell-level annotations are more detailed annotations that either mark the location of cells (point/dotting) or outline their boundaries (free-hand/polygon). They are used to extract features such as morphology, cell counts, and cell-to-cell ratios for diagnostic and prognostic purposes. (59)

Text annotations such as pathology reports can be used to extract clinically relevant information. A pathology report can provide clinical data including patient parameters, history, diagnosis, tumor size and possible targets for treatment. Natural Language Processing systems can automate this process, but accuracy might be challenging due to varying laboratory descriptions. As summarized by Wahab et al. (59), genomic, epigenomic, transcriptomic, proteomic, and imaging data from sources such as The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumour Analysis Consortium (CPTAC) can also be combined with histopathology to advance patient care. (59)

1.3.3 Annotation Constructs

Different tools can be used depending on the type of annotation. The most common annotation types are bounding box, point, polygon, line, and text. (60)

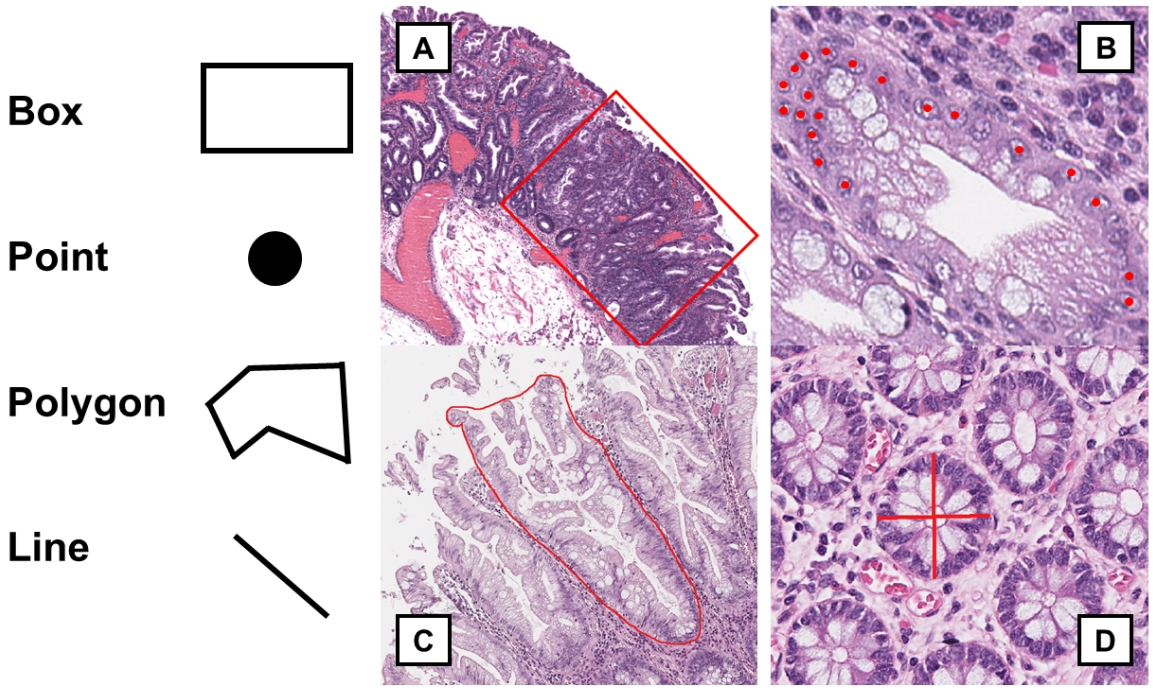

Figure 21: Annotation constructs. A Bounding box enclosing ROI (in this case a dysplastic region of an SSL). **B Point annotation** marking the centroid of cells or whole regions. **C Polygon annotation** marking the boundaries of a region or cell. **D Line annotation** marking the axis of a structure or joining/segregating annotations.

The bounding box construct is used to enclose ROIs and other annotation constructs such as polygons or points. Bounding boxes are axis-aligned rectangles and commonly utilized in exhaustive annotations, where a ROI is marked and further annotated by multiple pathologists. Furthermore, other categorizations of bounding boxes are possible (e.g., region- or cell-box), which subsequently allows an easier implementation of ML. A point annotation is a quick and simple annotation construct, but it is not as precise as polygon or free-hand drawing. This makes it useful for regions or cells. Polygons can be used to provide more precise annotation on a region- and cellular-level. Straight lines can be used to separate or merge structures. Further use can me to measure the diameter or other distances and mark the axis of the structure. Text annotations are used to provide context and meaning to other annotations and subsequently function as a meta-annotation. (59,60)

1.3.3.1 Histologic artifacts

General histologic artifacts are formed artificially and typically not present in the examined tissue. Artifacts can occur during any processing step of the specimen, in detail during the collection, fixation, embedding, microtome sectioning, staining, and cover slipping. The result can limit the evaluation of the tissue and further misinterpret

the findings. Some specimens, including colonic polyps, are removed with electric cautery. Electrocautery can cause nuclear spindling and pseudostratification, as well as dense and amorphous connective tissue, fat, and muscle. It may also result in the separation of epithelium from basement membrane. Further difficulties at the removal can be crushing of the tissue. Excess pressure on biopsy forceps can lead to distortion of tissue, destruction and compression of nuclei, extrusion of cytoplasmic contents, and telescoping (gland-in-gland) appearance. Cryopreservation injury is a phenomenon that occurs during specimen freezing, resulting in prominent gaps and holes in tissue caused by crystal formation. Small collections of air may occur during coverslipping, which involves placing a glass slide and coverslip together. Microtome sectioning can lead to nicks, tears, and fragmentation of tissues, resulting in holes or linear streaks. This may also create thin parallel lines known as chatter. Pigments and precipitates are dark brown, yellow-brown or black granules/clusters that form during processing or staining with formalin, mercury, or chromate. Contaminants on slides may include flecks and fragments of exogenous material such as squamous epithelial cells, pollen grains, small fibers or tissue from different specimens that occur during any step from collection to coverslipping. (62)

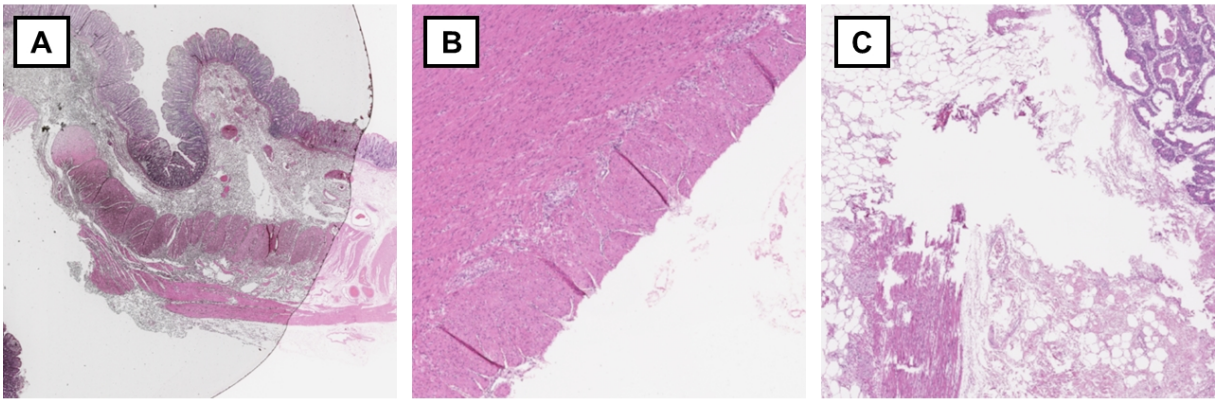

Figure 22: Common histologic artifacts. A showing an **air bubble** trapped under the coverslip after during the preparation of the slide. **B Folds and curls** are a common artifact in histology but can be prevented with careful technique. Some small folds are unavoidable in certain tissues. **C Fragmentation** of tissue can hinder proper histologic evaluation, especially in small samples, by potentially limiting or preventing adequate assessment.

Sufficient preservation of tissue integrity on the slide is imperative for precise and dependable quantitative analysis. In order to mitigate the risk of erroneous input in ML-based analyses, it is essential to thoroughly inspect slides for any signs of damage, including tears, folds, fragmentation, and other artifacts. Interfering blurriness and pen-

markings can pose major challenges to downstream analyses and should be manually excluded or automatically detected with QC of images before further annotation or computational analysis. Slides containing large artefact regions should also be directly excluded from further processing steps. (59)

1.4 Eye tracking

Eye-tracking technology is not only fostering innovative solutions and generating new data, but also spawning entirely novel areas of research. With the recording of participants and examination of eye movements, gaze behavior, and pupil dilation, it is possible to gain insight into the cognitive processes involved in visual perception and interpretation. Despite the diverse applications of eye tracking, several aspects are commonly shared, including the types of eye-tracking models and algorithms used for data processing and analysis. (63)

1.5 Aims of this study

Digital pathology is a thriving branch of pathology that has the structural capacity to introduce computational models to help pathologists in the areas of histopathological analysis and diagnosis. In the course of the project “ADOPT Dig Path Competition” at the Diagnostic and Research Institute of Pathology of the Medical University of Graz, histologic slides of colonic polyps, stored in the Biobank Graz, will be digitized and used for annotation in the open source software QuPath. (60) This dataset of annotated colon polyp WSIs is further processed as a training dataset for the development of AI-based algorithms for gland classification in a multi-step process.

1.5.1 Step 1: Annotation and export of annotated areas

The first step of this annotation project is to create a labeled dataset of physiological and dysplastic glands of colon polyps. These image sections with annotated glands are subsequently exported in different magnification levels commonly used in pathology.

1.5.2 Step 2: Eye tracking study of gland classification

With the help of an eye-tracking study we want to find out which tissue areas pathologists look at to evaluate whether a tissue is dysplastic or not, and, as a second step, enrich the glandular data set with eye tracking data. This data will be used in cooperation with DAI (TU-Berlin) for the development of a gland classifier algorithm: DAI wants to investigate whether this data can be used as (additional) input for training an attention-based neural network.

2 Materials and Methods

The required slides for this project were provided by the Biobank Graz. A pool of patients with colon polyps in the period of 1984 to 2014 was selected. These histologic slides were retrieved and digitized with high-resolution WSI scanners and subsequently anonymized. In total, 533 of these anonymized slides of colonic polyps were used in this project and further annotated in the open-source software QuPath. The dataset consisted of H&E-stained WSIs presenting colon polyps and normal colonic mucosa. (60,64) In this chapter, a comprehensive overview of the study protocol for the annotation project is provided. This includes the purpose of the project, the methodology and criteria for the annotation process and the following steps of this project.

2.1 Study protocol

The process starts with scanning the histologic slides and is followed by the annotation phase, where the regions are labeled according to a set of guidelines. Finally, the annotated data is exported in a format that can be used for further analysis or integration into other systems for a variety of purposes, including ML and data analysis.

2.2 Dataset

The design and implementation of a high-throughput digitization workflow has been described in the master thesis and two resulting papers by one of the supervisors, Dipl.-Ing. Markus Plass, Diagnostic and Research Institute of Pathology, Medical University of Graz. The digitization process was divided in: (1) case selection and data pre-processing, (2) scanning and scan quality control, and (3) post processing and cataloging" (64) p39).

In general, the process of case selection started with specifying significant search parameters: diagnosis, required sample type, time-period and additional information about the patient. The next step was the identification of possible patients in several databases, specifically the "*Pathology information system of the Medical University Graz* (PAS System), the *Hospital Information System* (Medocs), and the Statistic Austria *Death Registry"* (64) p40), which are further structured by metadata descriptors

collected from the archived medical records. Each resulting patient was pseudonymized with a designated research code and thus relabeled before delivering the slides to the scanning laboratory. Upon arrival, a 2D barcode for the automated scanning process was attached containing the prior patient-specific research code, year of sample creation, block number and cut-number. (64)

2.2.1 Scanning Process

After receiving the slides, they were prepared for the scanning process. First, a low-resolution image of each slide was captured via a specially constructed preview station. The original condition was documented as pathologists frequently mark a region of interest with a pencil, which must be erased for the subsequent scan of the slide. Additionally, all other disruptive factors like dust or fingerprints were removed thoroughly to prevent issues during focusing. In the following step, the cleaned slides are loaded in the WSI-scanner and the focusing process was performed manually, whereas the other scanning parameters like magnification, file format and compression parameters were configured automatically. About 20 focus points were set for each slide. (64) The WSI scanners used in this workflow were the P1000 by 3D Histech and the Leica Aperio AT2. Each scanning process was completed within 30 seconds and generated a file with an average size of 8GB. For this reason, data storage was a challenge in the routine implementation of WSI scanning and made a multi-stage storage process necessary. The first step included storing the scans on a local storage. Due to the handling of big amounts of data, the integrated HDD of the scanners must have be replaced by SSDs to achieve the required read/write speed while scanning the slide. Once the scan was completed, the resulting file was stored on a local HDD RAID until the quality check and metadata generation was completed, which took up to two weeks. Afterwards, the storage process concluded in the transfer of the scans to a Ceph storage cluster, as a cheaper and scalable alternative. (64)

After the scanning process, the digitized slides were post processed and catalogued. This step included the anonymization of the slides and subsequently an extensive quality control check, both initially performed manually. For anonymization, all slides were inspected for any identifiable information at the label area or on the slide. The two main goals for quality control were to verify the completeness of the slide and look for scanning artifacts, e.g., out of focus areas, stitching errors, or color fading. (64)

2.2.1.1 Cataloguing

The goal of the cataloguing process was the automatic categorization of the WSIs and the description of its scanned structures regarding size, properties, and artifacts. Due to the nonuniformity of the tissue areas, it was necessary to apply an extensive set of metadata attributes to collect structural properties for further use in visualization applications. (64) As previously described, several steps were needed for the detection and description of complex structures. First, the extraction and analysis of areas from imagesfinding the clusters of tissues andmeasures of similarities (between single areas and between groups of areas) and storage of all this information in an adequate data structure was performed. Areas of interest (possible tissue areas) were extracted in the first step with methods described in chapter 2.2.2.3. Metrics on out of-focus errors, regions of air bubbles and removed objects were calculated and stored as quality control measures." (64) p43).

2.2.2 Web-Service Implementation

For the implementation of the digitalization project, it was required to build an infrastructure that was capable of handling high throughput whole-slide scanning. Consequently, a web-service workflow was realized using Python. The implementation has multiple tasks, which will be described further in the following subchapters. (64)

2.2.2.1 Data Management

Several tasks regarding data were included in the web-service implementation. Specifically, the key elements were data handling, data movement, data storage and data hosting. Consequently, the scanned slides were transferred to a data storage solution that can handle the generated files with the approximate size of five petabytes annually. After anonymization of the data, an access to the processed data is provided, varying privileges and levels of information detail. (64)

2.2.2.2 Anonymization

The pseudonymization process was performed by the Institute for Medical Informatics of the Medical University of Graz. For de-identification, all identifiable information was replaced with pseudonyms. Subsequently, the data was imported and pseudonymized second time, where each pseudonym was replaced with a unique ID (UUID). If the data was needed for export, a new UUID for each project was generated, so the IDs inside of each project were uncluttered from other projects. (64)

2.2.2.3 Quality Control

With the implementation of large-scale slide digitalization, it was necessary to establish a well thought out quality check. During the project, quality control progressed through multiple phases. Initially in Phase 1, a manual quality check was conducted by two employees, which reviewed the preview image for completeness of the scan and later searched for out of focus areas. Following the first phase of the quality check, in Phase 2 a semi-automated quality check was introduced with the goal to simplify the process, to make it more reproducible and to lay the foundation for a fully automated quality check, envisioned in Phase 3. (64)

2.3 *Project Specific Annotation Instructions*

The following instructions guided annotators through the process of correctly identifying and labeling colonic polyps in WSI and to make the process of WSI-annotation easier to understand. The guidelines for the annotation process were discussed before two medical students started to annotate using the open-source annotation software QuPath. (60) Figure 23 shows an annotated polyp as an example of the annotation result.

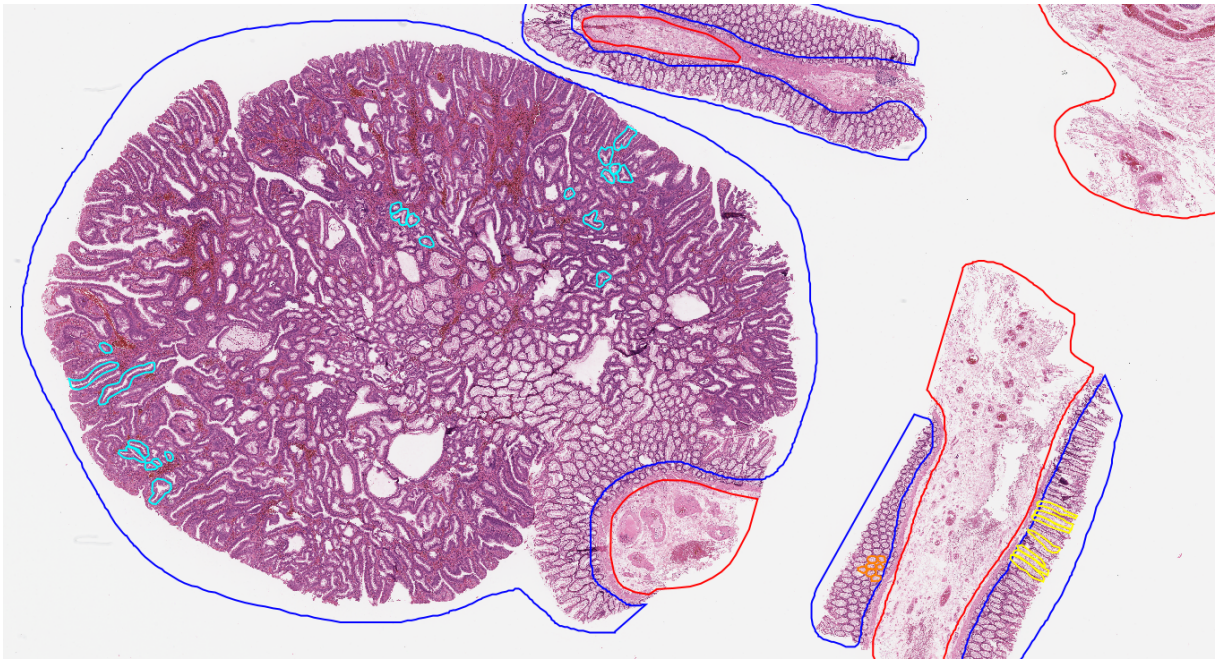

Figure 23: Example of an annotated polyp. The annotation process is described below in further detail.

Several regions of interest were specified and then annotated with a surrounding polygon. The predefined areas are:

1. Mucosa. The whole layer should be enclosed, using the basal lamina as a border between the mucosa and submucosa. The preset color is blue.
2. Submucosa. Tissue beneath the mucosa consists of the submucosal layer (e.g., connective tissue including blood vessels and the submucous plexus). The chosen color for this annotation is red. The annotation of mucosa and submucosa is not relevant for creation of a gland dataset

3. Physiological glands. Several glands of intact colonic mucosa were annotated. Up to 21 physiological glands should be annotated on each slide. Three distinctions were made because of the typical histologic structure:
    a. *Physiological glands cut along.* Up to 7 glands of this kind should be annotated. The preset color is yellow.
    b. *Physiological glands cut diagonally.* Up to 7 glands of this kind should be annotated. The preset color is orange.
    c. *Physiological glands cut across.* Up to 7 glands of this kind should be annotated. The preset color is green.
4. Dysplastic glands. Altogether 20 dysplastic glands were annotated on each slide. The different types of dysplastic glands are not considered. Due to the nature of dysplasia, it is not possible to determine the cutting direction for dysplastic glands. The chosen color for this annotation class is cyan.

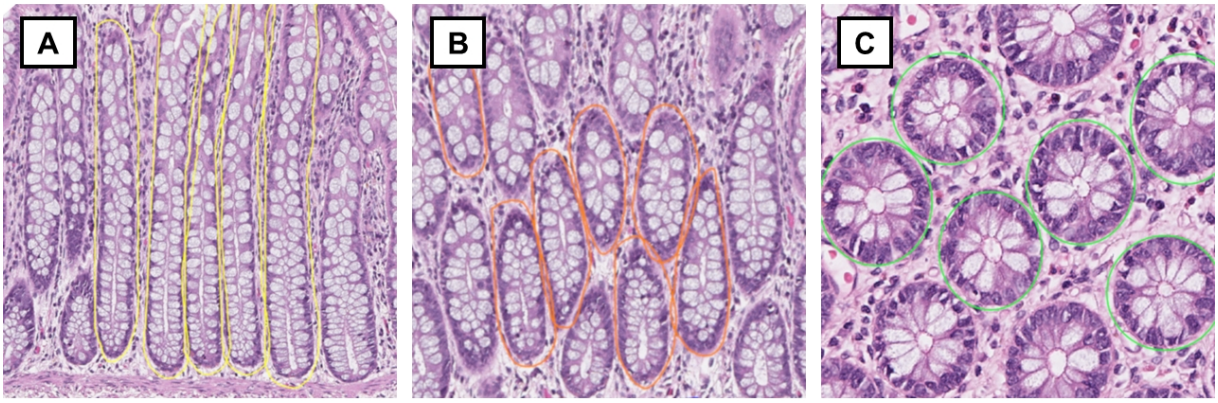

Figure 24: Annotated physiological glands. A Physiological glands cut along. **B** Physiological glands cut diagonally. **C** Physiological glands cut across.

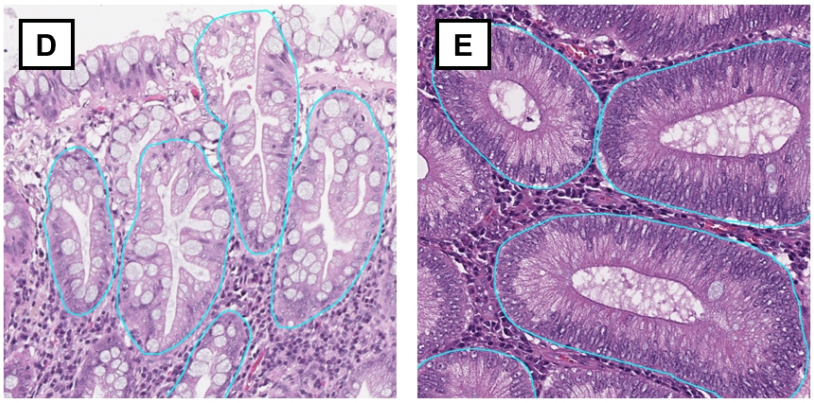

Figure 25: Annotated dysplastic glands. A Glands of a hyperplastic polyp. **B** Dysplastic glands of a tubular adenoma.

2.3.1 Exclusion criteria

In general, histologic artifacts (see chapter 1.3.3.1) were not annotated. Only H&E-stained slides were included, all other misplaced slides (e.g., IHC stained slides) and WSIs of other tissues were manually removed. In total, 33 slides were not considered for annotation.

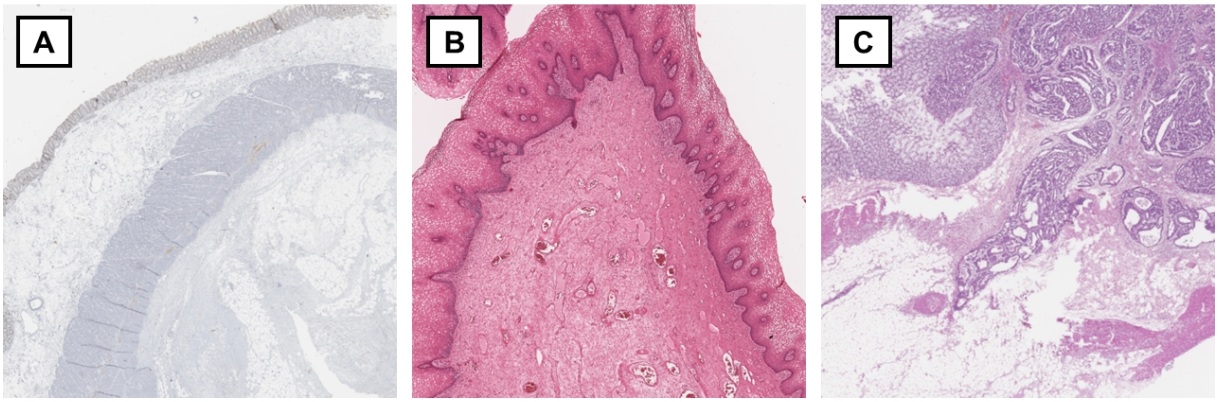

Figure 26: Examples of manually removed slides. A Only H&E stained slides were used for this project and the given slide shows an IHC stained slide. **B** Showing a misplaced slide with squamous epithelial tissue. **C** Only colonic polyps should be used, and this slide shows an invasive colorectal carcinoma.

2.3.2 Export of annotated areas

The annotated areas must be exported and stored for further development of ML algorithms. Using QuPath and a custom developed groovy script, the annotated areas are exported as rectangular images, where the center of the annotation-polygon is in the center of the exported image. The following attributes are variables for the export that can be set by the user:

1. Image resolution in x-direction (in px)
2. Image resolution in y-direction (in px)
3. File format (for this study we export .jpeg images)
4. Magnification at which the image should be exported

The following metadata is stored in a json file, separately for each image:

1. slidename - name of the slide (WSI) the image was exported from
2. scanner - which scanner digitized the slide in question
3. annotatorId - who annotated the slide (WSI) in question
4. maxMagSlide - the maximum magnification the slide was scanned with

5. resolutionSlide - the resolution of the scan
6. currentMagTile - the magnification at which the image was exported
7. centerX - the x coordinate of the image center at the WSI (WSI (0,0) is at the upper left corner)
8. centerY - the y coordinate of the image center at the WSI (WSI (0,0) is at the upper left corner)
9. geometry_xycomma - points that describe the annotation (usually a polygon)
10. comment - comment that can be added with the export
11. class - annotation classification

The mask, generated from the annotation polygon is stored in a separate file.

3 Results

For this project, physiological and dysplastic glands of the colon were annotated. In total 17937 image samples were collected, consisting of 10.088 physiological glands and 7.848 dysplastic glands. The number of the annotated glands including the absolute and relative share are presented in Table 1. The crop images were exported as rectangular images with the annotation-polygon marking the center of the image. Each image had the resolution of 1024x1024 pixels.

	Absolute share	**Relative share**
Physiological glands cut along	3211	17,90%
Physiological glands cut diagonally	3475	19,37%
Physiological glands cut across	3402	18,97%
Dysplastic glands	7848	43,75%
Total	17937	100,00%

Table 1: List of the collected crop images and their shares

3.1 Dataset for the Eye Tracking Study

The following classification of dysplastic/physiologic refers to the WSI annotation described above. For the number of images used in the eye tracking study, see the table below:

	Total Number of annotated images	**Images per pathologist**	
Category 1	1820	250	
Category 2	900	150	
Category 3	0		
Category 4	1456	200	50 sets à 4 different magnifications
Planned Total number of annotated images	**4176**	**600**	
total number of annotated images from individual pools	3720	372	
total number of annotated images from common pools	456	228	

Table 2: Summary of the dataset of the eye tracking study

If not stated otherwise, the order of images shown to the participants was randomized and the same image was not shown to the same participant multiple times. In addition,

if not stated otherwise, the number of physiological glands was equally spread between those cut across, along and diagonally (⅓, ⅓, ⅓)

For this study, we used four categories of images:

Category 1

1. images with a magnification of 10 and a resolution of 1024px*1024px with the gland in question in the center
2. from those images, a common pool and an individual pool was selected. The common pool was classified by multiple participants whereas the individual pool was exclusive for each participant.

Category 2

1. images with a magnification of 40 and a resolution of 1024px*1024px with the gland in question in the center
2. using the annotation-mask, the surroundings of the gland were blurred or blackened.

Category 3

1. Whole Slide Images (WSIs) from the digitized set of 533 histological slides of colon polyps that were diagnosed.

Category 4

1. sets of images with the gland in question in the center; all images of such a set have got a resolution of 1024px*1024px but at different magnifications (40x, 20x, 10x, 5x), meaning that the same gland is shown 4 times with different levels of magnification.

3.1.1 Study procedure

The participants were shown images of glands which should then be annotated. During the study, the participants' gaze was recorded using a Tobii Eye-Tracker and the iMotions suite. Annotation is done in a 2-step process:

1. Classification of the tissue as:
    a. Dysplastic
    b. Physiologic
    c. Not classifiable
2. Statement of confidence as:

a. Very sure
b. Sure
c. Unsure

After an image is annotated, the next image is shown automatically.

3.1.2 Participants

Number of participants included in the study was up to 12 (10 pathologists and 2 medical students).

3.1.3 Implementation

Iteration – 1

Participants: 10 pathologists

Tasks: classify images of category 1, category 2 and additionally diagnose WSIs from category 3.

Iteration – 2

Participants: 3 pathologists

Tasks: classify images of categories 2 and 4.

Iteration – 3

Participants: 2 medical students

Tasks: classify images of category 1, 2, 3 and 4.

4 Discussion

The objective of this diploma thesis was the creation of an annotated dataset for further use in the development of classification algorithms. Data is a crucial component of AI algorithms, as it provides the model with hidden insights and statistics that it can learn from. It is crucial to verify that the data utilized for training an artificial intelligence model is pertinent to the specific problem at hand and closely mirrors real-world data. A large amount of diverse data is required to train an AI model effectively, as it allows the model to recognize invariant features and variations in the input samples, leading to better accuracy. Our dataset consisted of 533 H&E-stained WSIs.

Labeled datasets of histopathologic images that are needed to train complex networks are sparse. Typically, experienced pathologists have annotated specific characteristics in these datasets, which offer comprehensive imagery of both colorectal cancer tissue and tumors. Several publicly available datasets, which were used for studies and contests are summarized below in Table 3. In comparison, our colon gland dataset is of considerable size and quality. There are several limitations in the direct comparison of these datasets as the datasets were created with different goals. For example, the Kather texture dataset (65) is a collection of several histological textures found in CRC. In some, the annotation modality differs. Graham et al. presented an unannotated dataset consisting of 41 H&E-stained images, which was further annotated by their CNN for nuclear segmentation and classification that produced in total 24.319 annotated and classified nuclei. (66) In contrast, Byeon et al. annotated the images of colonic polyps for their automated histological classification project on a case-level. (67) The most similar datasets in terms of region-level annotation of glands are the GlaS datatset by Sirinukunwattana et al. (68) and the CRAG dataset by Graham et al. (69). However, it should be mentioned, that the input for these datasets were WSIs of CRC, while our colon gland dataset is based on the precursor of CRC, non-invasive colonic polyps (adenomas).

Name	**Release Date**	**Number of Samples**	**Average Dimensions (in Pixels)**
Kather texture dataset (65)	2016	5000 patches	150 × 150
GlaS dataset (68)	2016	165 images	775 × 522
CRC-TIA (70)	2017	139 WSI	1792 × 1792

Deep Learning for Classification of Colorectal Polyps on Whole-slide Images (7)	2017	2074 images	811 x 984
Colorectal nuclear segmentation and phenotypes (CoNSeP) (66)	2018	41 images	1000 × 1000
Colorectal adenocarcinoma gland (CRAG) (69)	2019	213 images	1512 × 1516
Histological images for tumor detection in gastrointestinal cancer (70)	2019	11,977 patches	512 × 512
Pathology AI platform (PAIP) (71)	2019	118 WSI	29,879 × 23,066
UniToPatho (72)	2021	9536 patches	224 × 224
Automated histological classification for digital pathology images of colonoscopy specimen via deep learning (67)	2022	1865 images	613 x 408
EMPAIA Eye tracking - Colon gland dataset	***2022***	***17937 images***	***1024x1024***

Table 3: Several datasets that are readily accessible to the public. Own table based on Tamang et al. (2021; p6) (73)

The presence of diverse structures in biological tissue creates a difficulty for both manual and automated examination of histopathology slides. In the past few years, automated analysis has become crucial for precise quantitative morphology evaluation and cancer grading. Although virtual microscopy has already become viable in some pathology departments, the issue of interobserver variability persists due to the subjective nature of slide evaluations. Consequently, relying solely on expert scoring as the gold standard for histopathological assessment may be inadequate. As a result, there is an increasing need for robust computational methods to enhance diagnostic reproducibility.

In this regard, medical application-focused DL models have the potential to decrease the time required for analyzing, diagnosing, and prognosing histologic entities. Additionally, it can reduce the workload of clinicians and pathologists, as well as minimize potential errors that may occur during histopathological characterization of colorectal polyps, which is crucial for follow-up recommendations and risk assessment. More advanced technology can significantly enhance diagnostic and prognostic

accuracy, thus progressing to precision medicine. Furthermore, digital pathology projects on colonic polyp classification can enhance clinical training by creating a platform for better quality assurance of colorectal cancer screening and a deeper comprehension of common error patterns in the histopathological characterization of colorectal polyps. Summarized, the implementation of digital pathology in daily practice can improve the accuracy of colorectal cancer screening, alleviate the cognitive burden on pathologists, enhance patient health outcomes, and promote early preventive measures to lower colorectal cancer mortality rates.

5 References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, u. a. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. Mai 2021;71(3):209–49.

2. Aarons CB, Shanmugan S, Bleier JI. Management of malignant colon polyps: Current status and controversies. World J Gastroenterol. 21. November 2014;20(43):16178–83.

3. Remmele W. Pathologie. 3., neubearbeitete Aufl. Berlin: Springer; 2013.

4. Arnold CA, Lam-Himlin DM, Montgomery E. Atlas of gastrointestinal pathology: a pattern based approach to neoplastic biopsies. 2019.

5. Rye C, Wise R, Jurukovski V, DeSaix J, Choi J, Avissar Y. Biology. Houston, Texas: OpenStax; 2016.

6. Lüllmann-Rauch R, Asan E. Taschenlehrbuch Histologie. 6., vollständig überarbeitete Auflage. Stuttgart New York: Georg Thieme Verlag; 2019. 781 S.

7. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, u. a. Deep Learning for Classification of Colorectal Polyps on Whole-slide Images. J Pathol Inform. 2017;8:30.

8. Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology Guidelines for Colorectal Cancer Screening 2008. Official journal of the American College of Gastroenterology | ACG. März 2009;104(3):739–50.

9. Participants in the Paris Workshop. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. Gastrointestinal Endoscopy. Dezember 2003;58(6):S3–43.

10. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial colon cancer. Gastroenterology. Juni 2010;138(6):2044–58.

11. Bailie L, Loughrey MB, Coleman HG. Lifestyle Risk Factors for Serrated Colorectal Polyps: A Systematic Review and Meta-analysis. Gastroenterology. Januar 2017;152(1):92–104.

12. Durno CA. Colonic polyps in children and adolescents. Can J Gastroenterol. April 2007;21(4):233–9.

13. Zbuk KM, Eng C. Hamartomatous polyposis syndromes. Nat Rev Gastroenterol Hepatol. September 2007;4(9):492–502.

14. Carlsson G, Petrelli NJ, Nava H, Herrera L, Mittelman A. The Value of Colonoscopic Surveillance After Curative Resection for Colorectal Cancer or Synchronous Adenomatous Polyps. Archives of Surgery. 1. November 1987;122(11):1261–3.

15. Organisation mondiale de la santé, Centre international de recherche sur le cancer, Herausgeber. Digestive system tumours. 5th ed. Lyon: International agency for research on cancer; 2019. (World health organization classification of tumours).

16. Fenoglio-Preiser CM, Herausgeber. Gastrointestinal pathology: an atlas and text. 3rd ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins; 2008. 1296 S.

17. Pendergrass CJ, Edelstein DL, Hylind LM, Phillips BT, Iacobuzio-Donahue C, Romans K, u. a. Occurrence of colorectal adenomas in younger adults: an epidemiologic necropsy study. Clin Gastroenterol Hepatol. September 2008;6(9):1011–5.

18. Williams AR, Balasooriya BA, Day DW. Polyps and cancer of the large bowel: a necropsy study in Liverpool. Gut. Oktober 1982;23(10):835–42.

19. Sanchez NF, Stierman B, Saab S, Mahajan D, Yeung H, Francois F. Physical activity reduces risk for colon polyps in a multiethnic colorectal cancer screening population. BMC Research Notes. 20. Juni 2012;5(1):312.

20. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1. Juni 1990;61(5):759–67.

21. Smith G, Carey FA, Beattie J, Wilkie MJV, Lightfoot TJ, Coxhead J, u. a. Mutations in APC, Kirsten-ras, and p53--alternative genetic pathways to colorectal cancer. Proc Natl Acad Sci U S A. 9. Juli 2002;99(14):9433–8.

22. White BD, Chien AJ, Dawson DW. Dysregulation of Wnt/β-catenin signaling in gastrointestinal cancers. Gastroenterology. Februar 2012;142(2):219–32.

23. Nakamura M, Zhou XZ, Lu KP. Critical role for the EB1 and APC interaction in the regulation of microtubule polymerization. Curr Biol. 1. Juli 2001;11(13):1062–7.

24. Rubio CA, Rodesjö M, Duvander A, Mathies M, Garberg L, Shetye J. p53 Up-regulation During Colorectal Carcinogenesis. ANTICANCER RESEARCH. 2014;7.

25. Shinya H, Wolff WI. Morphology, anatomic distribution and cancer potential of colonic polyps. Ann Surg. Dezember 1979;190(6):679–83.

26. Wong S, Lidums I, Rosty C, Ruszkiewicz A, Parry S, Win AK, u. a. Findings in young adults at colonoscopy from a hospital service database audit. BMC Gastroenterol. 19. April 2017;17(1):56.

27. Zhou H, Shen Z, Zhao J, Zhou Z, Xu Y. [Distribution characteristics and risk factors of colorectal adenomas]. Zhonghua Wei Chang Wai Ke Za Zhi. 25. Juni 2018;21(6):678–84.

28. Odze RD. “Sessile Serrated Lesion”: The Art and Science of Naming a Disorder. Archives of Pathology & Laboratory Medicine. 27. September 2021;145(10):1190–1.

29. Sacco M, Palma FDED, Guadagno E, Giglio MC, Peltrini R, Marra E, u. a. Serrated lesions of the colon and rectum: Emergent epidemiological data and molecular pathways. Open Medicine. 1. Januar 2020;15(1):1087–95.

30. Crockett SD, Nagtegaal ID. Terminology, Molecular Features, Epidemiology, and Management of Serrated Colorectal Neoplasia. Gastroenterology. 1. Oktober 2019;157(4):949-966.e4.

31. Carr NJ, Mahajan H, Tan KL, Hawkins NJ, Ward RL. Serrated and non-serrated polyps of the colorectum: their prevalence in an unselected case series and correlation of BRAF mutation analysis with the diagnosis of sessile serrated adenoma. J Clin Pathol. Juni 2009;62(6):516–8.

32. Abdeljawad K, Vemulapalli KC, Kahi CJ, Cummings OW, Snover DC, Rex DK. Sessile serrated polyp prevalence determined by a colonoscopist with a high lesion detection rate and an experienced pathologist. Gastrointest Endosc. März 2015;81(3):517–24.

33. Kahi CJ, Rex DK. Sessile serrated lesions: Searching for the true prevalence. Endosc Int Open. April 2021;9(4):E635–6.

34. Haque TR, Bradshaw PT, Crockett SD. Risk factors for serrated polyps of the colorectum. Dig Dis Sci. Dezember 2014;59(12):2874–89.

35. Figueiredo JC, Crockett SD, Snover DC, Morris CB, McKeown-Eyssen G, Sandler RS, u. a. Smoking-associated risks of conventional adenomas and serrated polyps in the colorectum. Cancer Causes Control. März 2015;26(3):377–86.

36. Wallace K, Grau MV, Ahnen D, Snover DC, Robertson DJ, Mahnke D, u. a. The association of lifestyle and dietary factors with the risk for serrated polyps of the colorectum. Cancer Epidemiol Biomarkers Prev. August 2009;18(8):2310–7.

37. Hashimoto T, Yamashita S, Yoshida H, Taniguchi H, Ushijima T, Yamada T, u. a. WNT Pathway Gene Mutations Are Associated With the Presence of Dysplasia in Colorectal Sessile Serrated Adenoma/Polyps. Am J Surg Pathol. September 2017;41(9):1188–97.

38. Sekine S, Ogawa R, Hashimoto T, Motohiro K, Yoshida H, Taniguchi H, u. a. Comprehensive characterization of RSPO fusions in colorectal traditional serrated adenomas. Histopathology. 2017;71(4):601–9.

39. Cenaj O, Gibson J, Odze RD. Clinicopathologic and outcome study of sessile serrated adenomas/polyps with serrated versus intestinal dysplasia. Mod Pathol. April 2018;31(4):633–42.

40. Bettington ML, Walker NI, Rosty C, Brown IS, Clouston AD, McKeone DM, u. a. A clinicopathological and molecular analysis of 200 traditional serrated adenomas. Mod Pathol. März 2015;28(3):414–27.

41. Broich K, Löbker W, Lauer W. Beitrag des BfArM zur Potenzialentfaltung der Digitalisierung im Gesundheitswesen – digital readiness@BfArM. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2021;64(10):1292–7.

42. Weinstein RS. Prospects for telepathology. Human Pathology. 1. Mai 1986;17(5):433–4.

43. Pallua JD, Brunner A, Zelger B, Schirmer M, Haybaeck J. The future of pathology is digital. Pathol Res Pract. September 2020;216(9):153040.

44. Pantanowitz L. Digital images and the future of digital pathology. J Pathol Inform. 10. August 2010;1:15.

45. Pantanowitz L, Sharma A, Carter AB, Kurc T, Sussman A, Saltz J. Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives. J Pathol Inform. 21. November 2018;9:40.

46. Jahn SW, Plass M, Moinfar F. Digital Pathology: Advantages, Limitations and Emerging Perspectives. J Clin Med. 18. November 2020;9(11):E3697.

47. García-Rojo M, Ordi J. Trying to Understand Digital Pathology before We Move to Computational Pathology. PAT. 2016;83(2–3):57–60.

48. Marini N, Otálora Montenegro J, Podareanu D, Rijthoven M, van der Laak J, Ciompi F, u. a. Multi_Scale_Tools: A Python Library to Exploit Multi-Scale Whole Slide Images. Frontiers in Computer Science. 1. August 2021;3:684521.

49. Wetteland R, Kvikstad V, Eftestøl T, Tossebro E, Lillesand M, Janssen E, u. a. Automatic Diagnostic Tool for Predicting Cancer Grade in Bladder Cancer Patients Using Deep Learning. IEEE Access. 13. August 2021;PP:1–1.

50. Kumar N, Gupta R, Gupta S. Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. J Digit Imaging. 1. August 2020;33(4):1034–40.

51. Iyengar JN. Whole slide imaging: The futurescape of histopathology. Indian Journal of Pathology and Microbiology. 1. Januar 2021;64(1):8.

52. Cheng WC, Saleheen F, Badano A. Assessing color performance of whole-slide imaging scanners for digital pathology. Color Research & Application. 2019;44(3):322–34.

53. Lahrmann B, Valous N, Eisenmann U, Wentzensen N, Grabe N. Semantic Focusing Allows Fully Automated Single-Layer Slide Scanning of Cervical Cytology Slides. PloS one. 9. April 2013;8:e61441.

54. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest. April 2021;101(4):412–22.

55. Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, u. a. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. November 2019;249(3):286–94.

56. Sakamoto T, Furukawa T, Lami K, Pham HHN, Uegami W, Kuroda K, u. a. A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. Transl Lung Cancer Res. Oktober 2020;9(5):2255–76.

57. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointestinal Endoscopy. 1. Oktober 2020;92(4):807–12.

58. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 19. März 2019;19:64.

59. Wahab N, Miligy IM, Dodd K, Sahota H, Toss M, Lu W, u. a. Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. J Pathol Clin Res. 10. Januar 2022;8(2):116–28.

60. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, u. a. QuPath: Open source software for digital pathology image analysis. Sci Rep. 4. Dezember 2017;7(1):16878.

61. Humphries MP, Maxwell P, Salto-Tellez M. QuPath: The global impact of an open source digital pathology system. Computational and Structural Biotechnology Journal. 1. Januar 2021;19:852–9.

62. Lindberg MR, Herausgeber. Diagnostic pathology. Normal histology. Third edition. Philadelphia, PA: Elsevier; 2023.

63. Holmqvist K, Örbom SL, Hooge ITC, Niehorster DC, Alexander RG, Andersson R, u. a. Eye tracking: empirical foundations for a minimal reporting guideline. Behav Res Methods. 2023;55(1):364–416.

64. Plass M. Large Scale Slide Digitalisation for Machine Learning in Computational Pathology [Master's Thesis]. 2020.

65. Kather JN, Weis CA, Bianconi F, Melchers SM, Schad LR, Gaiser T, u. a. Multi-class texture analysis in colorectal cancer histology. Sci Rep. 16. Juni 2016;6:27988.

66. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, u. a. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis. 1. Dezember 2019;58:101563.

67. Byeon S ju, Park J, Cho YA, Cho BJ. Automated histological classification for digital pathology images of colonoscopy specimen via deep learning. Sci Rep. 27. Juli 2022;12:12804.

68. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng PA, Guo YB, u. a. Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest [Internet]. arXiv; 2016 [zitiert 25. Februar 2023]. Verfügbar unter: http://arxiv.org/abs/1603.00275

69. Graham S, Chen H, Gamper J, Dou Q, Heng PA, Snead D, u. a. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Med Image Anal. Februar 2019;52:199–211.

70. Shaban M, Awan R, Fraz MM, Azam A, Tsang YW, Snead D, u. a. Context-Aware Convolutional Neural Network for Grading of Colorectal Cancer Histology Images. IEEE Trans Med Imaging. Juli 2020;39(7):2395–405.

71. Kim YJ, Jang H, Lee K, Park S, Min SG, Hong C, u. a. PAIP 2019: Liver cancer segmentation challenge. Medical Image Analysis. 1. Januar 2021;67:101854.

72. Barbano CA, Perlo D, Tartaglione E, Fiandrotti A, Bertero L, Cassoni P, u. a. Unitopatho, A Labeled Histopathological Dataset for Colorectal Polyps Classification and Adenoma Dysplasia Grading. 2021 IEEE International Conference on Image Processing (ICIP). 19. September 2021;76–80.

73. Tamang L, Kim B. Deep Learning Approaches to Colorectal Cancer Diagnosis: A Review. Applied Sciences. 19. November 2021;11:10982.