xAILAB
EXPLAINABLE MACHINE LEARNING

Development of a dataset and AI-based proof-of-concept algorithm for the classification of digitized whole slide images of gastric tissue

Bachelor Thesis

Bachelor of Science in Software Systems Science

Tom Hempel

September 29, 2023

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Dr. Bettina Braunecker (Institute of Pathology of Paracelsus Medical Private University Nuremberg Clinic)

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

In the domain of medical imaging, digitized whole slide images (WSIs) of gastric tissue present a unique intersection of traditional medical expertise and modern computational capabilities. This thesis primarily focuses on two pivotal aspects: the creation and meticulous annotation of a comprehensive dataset, and the development of an AI-based algorithm for the classification of these WSIs.

The process of curating the dataset and its subsequent annotation stands as a cornerstone of this and any further work on the project. The dataset serves as the bedrock upon which the algorithm learns, making its accuracy and comprehensiveness paramount. The intricacies involved in data collection, anonymization, transportation, and storage will be discussed in detail, emphasizing the challenges faced and the strategies employed to maintain data integrity and privacy.

Two distinct models were developed, with one focusing on the classification of gastric regions and the other on the detection of inflammation. At the WSI-level, the model dedicated to gastric regions achieved an F1 score of 0.8889 for antrum and a perfect score of 1 for corpus, while presenting a lower score of 0.4615 for intermediate regions. The reduced score for intermediate regions is primarily attributed to these regions falling short of the threshold for classification as either antrum or corpus. It is noteworthy that the recall for intermediate regions was determined to be 1, indicating the precise identification of all intermediate samples. The assessment of the model for inflammation classification yielded exemplary scores across the entire test set at the WSI level, with the F1 score for both classes reaching 1. Nonetheless, these results should be interpreted with caution due to the constraints of dataset size and the inclusion of samples with unequivocal classifications.

Challenges encountered throughout the project, from dataset creation to algorithm development, offer valuable insights and set the stage for future endeavors in this domain. Building on this foundation, the subsequent master thesis by Philipp Andreas Höfling aims to further the work, refining the algorithm and expanding its applicability.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
WSI	Whole Slide Image
DPA	Digital Pathology Association
DP	Digital Pathology
GUI	Graphical User Interface

1 Introduction

In the evolving field of medical imaging, the classification of digitized whole slide images (WSIs) of gastric tissue represents a significant intersection of medical expertise and computational capabilities. This bachelor thesis is situated within a collaborative project between the Chair of Explainable Machine Learning and the Institute of Pathology of Paracelsus Medical Private University Nuremberg Clinic. The project aims to develop an AI-based algorithm capable of distinguishing between different gastric regions, namely antrum, corpus, and intermediate, as well as the identification of inflammation. The creation of the dataset of WSIs is a vital step in this thesis and any further work on this project, including Philipp Andreas Höfling's work on his master thesis, which aims on refining the system further. The careful and meticulous annotations of these images are crucial for training accurate and reliable models, which form the foundation for the development and enhancement of the AI-based algorithm.

1.1 Motivation

The motivation behind this thesis stems from the increasing need for accurate and efficient classification of WSIs in the field of pathology. By leveraging AI-based algorithms, it is possible to enhance the precision and speed of diagnosis, thereby contributing to improved patient outcomes (Nam et al., 2020). As Gastritis is likely present in over half the world's population (Sipponen and Maaroos, 2015), it takes up a significant amount of resources in everyday operations and could benefit from the better accuracy and efficiency of an automated system. The utilization of standard cases from routine diagnostics, selected by medical staff of pathology, and the subsequent digitalization and anonymization of these cases form the basis for developing the AI-based program. While it is acknowledged that the completion of this project extends beyond the scope of this thesis, the work conducted herein lays a foundational baseline for further development and refinement in future studies and applications.

1.2 Project Goal

The primary objective of this project is to lay a robust foundation for the development of an AI-based algorithm capable of classifying digitized WSIs of gastric tissue, both simply classifying the gastric region as well as inflammation. A crucial component of achieving this goal is the creation of a comprehensive dataset, which should encompass as many WSIs as possible, with a particular emphasis on including a variety of edge cases to ensure the model's adaptability and robustness. The annotations accompanying these images are expected to be of the highest accuracy, providing detailed insights into the different gastric regions and, depending on the progress of the project, potentially including information on whether the WSIs exhibit inflammation. The development of a baseline model is another pivotal aspect

of this project. This model will serve as a foundational benchmark which future work can be compared against and built upon. By showcasing the potential functionality and capabilities of the project, the baseline model aims to facilitate further research and development.

2 Technical Background

This section provides a foundational understanding of the advancements and methodologies central to digital pathology (DP). It begins by exploring the transformative shift from traditional pathology practices to more sophisticated, digital methods, highlighting the integration of technologies such as WSI and Artificial Intelligence (AI). The section delves into the intricacies of WSI, discussing its significance, application, and the technology underpinning it. Furthermore, it addresses the challenges and opportunities presented by DP, examining the constraints, limitations, and the potential for innovation and improvement in diagnostic processes.

2.1 Introduction Digital Pathology

In the realm of medical imaging, a transformative shift has emerged, moving away from traditional practices of pathology to more advanced, digitized methods. This evolution has not only enhanced diagnostic processes but also broadened the scope of pathology by seamlessly integrating cutting-edge technologies.

Often termed as WSI, this digital approach involves the conversion of pathology slides into digital formats, facilitating their examination on computer workstations (Williams et al., 2017). Such advancements have cemented their role in contemporary clinical practice, becoming indispensable within laboratory settings (Niazi et al., 2019).

The momentum towards digital methodologies has been bolstered by various catalysts, including regulatory milestones and the growing inclination of healthcare entities to incorporate DP into diagnostic modalities (Williams et al., 2017).

The implications of this digital transition in healthcare are vast. It provides a versatile platform that promises enhancements in diagnostic pathology, underpinning safety, quality, and efficiency (Williams et al., 2017). The assimilation of these digital slides into routine pathology workflows sets the stage for the development of sophisticated algorithms and computer-aided diagnostic tools, which will be delved into further in the 2.2 subsection.

Furthermore, the intersection of machine learning and artificial intelligence with this field has amplified the potential of DP. AI's promise in pinpointing unique imaging markers linked to disease processes is likely to elevate early detection, prognosis assessments, and the selection of efficacious treatments (Niazi et al., 2019).

2.2 Whole Slide Imaging (WSI)

Whole-slide digital imaging is pivotal for the high-quality digitization and storage of slides. It stands out due to its ability to store and display slides on a computer screen, contrasting with the traditional method of viewing through a microscope, as summarized by (Aeffner et al., 2019) from the Digital Pathology Association (DPA) white paper on WSI (Zarella MD, 2019).

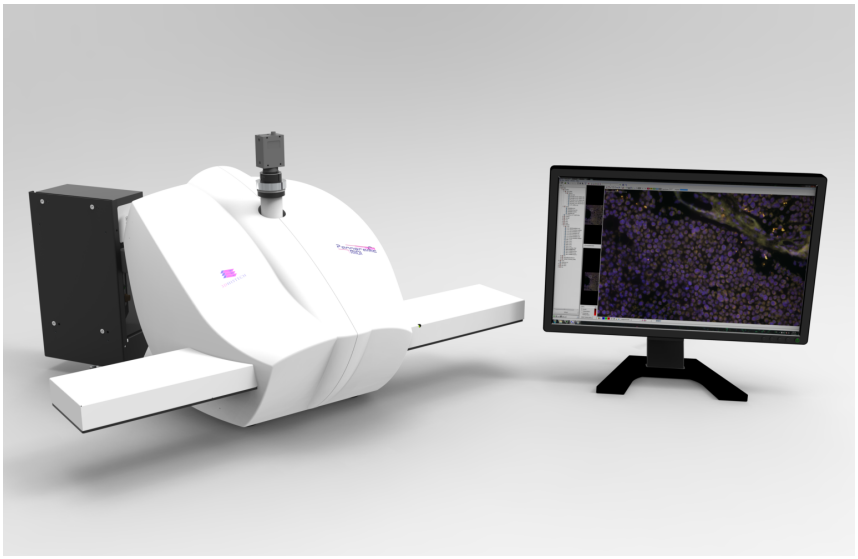


Figure 1: Pannoramic MIDI II (left) with a WSI opened on a monitor (right). Source: 3DHISTECH Ltd., `https://www.3dhistech.com/research/pannoramic-digital-slide-scanners/pannoramic-midi/` (Last accessed on 27.09.2023)

The white paper (Zarella MD, 2019) outlines several challenges, including the need for additional equipment like slide scanners, storage, workstations, trained personnel, and quality checks. However, it also highlights the benefits, such as the enhanced ability to share slides for expert consultation or educational purposes (Aeffner et al., 2016). The ease of sharing WSIs facilitates the creation of uniform educational datasets, ensuring equal learning opportunities by eliminating slide-to-slide variations (J and R, 2014). Additionally, the digital nature of WSIs allows for unbiased, objective evaluation through image analysis software (Aeffner et al., 2019) and potentially simplifies long-term archiving. However, archiving necessitates substantial infrastructure due to the large size of the images, each typically ranging from 200 MB to 1 GB (J and R, 2014). The storage of approximately 2000 images, assuming an average size of 500 MB per WSI, can incur annual costs between $3000 and $10,000, accounting for backups and security (J and R, 2014). While various vendors employ different techniques such as tiling or line scanning to generate high-quality images (Ghaznavi et al.), the majority stitch together high-resolution images of smaller sections into tiles (J and R, 2014). The WSI scanner depicted in figure 1, utilized for

this project, employs said tile technique, capable of automatically scanning up to 12 slides simultaneously using 40x and 20x objectives.

2.3 Challenges and Opportunities

The field of DP has the potential to revolutionize the way pathology is practiced, from the way samples are analyzed to the way diagnoses are made. However, as with any new technology, there are a number of challenges that must be overcome in order to fully realize its potential. This section will introduce a number of these challenges as well as multiple opportunities that arise with the development of this research field.

Challenges As highlighted in section 2.2, a pivotal constraint for widespread adoption of DP is the infrastructural requirement. This encompasses the need for adequate storage, scanners, high-performance machines, and bandwidth to facilitate swift sharing and viewing of server-stored images (J and R, 2014).

A notable limitation of most systems employing artificial intelligence for analyzing WSIs is their specialization in singular tasks (Tizhoosh and Pantanowitz, 2018). While such systems may be proficient in specific assignments, their practicality diminishes when multiple systems are necessitated to analyze a single case for various diseases. A trained pathologist on the other hand can identify irregularities, even when suspecting something else. This limitation is evident in this project, where the models, though adept at classifying the gastric region and determining inflammation, lack versatility in making other classifications or observations that might otherwise be obvious.

Furthermore, the scarcity of publicly available, labeled datasets poses a significant challenge, as they are essential for optimizing the performance of AI models (Tizhoosh and Pantanowitz, 2018). The creation of new systems often mandates the development of datasets and accompanying annotations from scratch, a resource-intensive process requiring expert input to ensure quality and mitigate inaccuracies (Tizhoosh and Pantanowitz, 2018). This project also encountered this challenge, given the absence of publicly accessible datasets, necessitating the creation of a dataset without the possibility of augmentation using existing images.

Another critical aspect is the regulatory hurdles associated with the approval of DP tools for patient use beyond research (Bera et al., 2019). Achieving regulatory approval necessitates a clear understanding of the software's functionality, a challenge that has given rise to the field of explainable AI (Gunning and Yang, 2019).

Opportunities DP tools offer significant advantages, one of which is the alleviation of repetitive tasks such as counting or scanning images for specific features. This capability not only enhances the efficiency of pathologists but also concurrently reduces their workload (Tizhoosh and Pantanowitz, 2018). While these systems may

not entirely replace certain tasks performed by pathologists, they can provide valuable assistance in decision-making, especially in ambiguous cases or when a second opinion is sought (Madabhushi and Lee, 2016).

A prominent opportunity arising from digital and computational pathology is the potential for achieving higher accuracy in diagnoses (Nam et al., 2020). The digitization of pathology slides enables their analysis by computer algorithms, which, trained on extensive data, can often discern patterns and features potentially overlooked by the human eye (Nam et al., 2020). This enhanced detection capability can contribute to more accurate diagnoses, especially in cases with subtle or challenging pathology (Nam et al., 2020). Furthermore, the integration of AI in pathology is instrumental in identifying biomarkers, facilitating the customization of treatments for patients (Tizhoosh and Pantanowitz, 2018).

Telepathology represents another significant opportunity, serving as a medium for transmitting pathology images and enabling remote diagnoses (Nam et al., 2020). Advances in this field can foster easier sharing of images among pathologists and students, thereby increasing accessibility and equality (Nam et al., 2020). It facilitates consultations for second opinions from pathologists who may not be locally available and holds potential for specialized healthcare delivery in underserved regions (N. and L., 2015). Despite challenges such as infrastructure requirements, regulatory barriers, and technical failures (N. and L., 2015), telepathology unveils a plethora of opportunities worth exploring.

3 Clinical Background

This section provides an introduction to the basic anatomy and function of the stomach, along with a brief overview of gastritis. It explores the prevalence and potential implications of gastritis and outlines the different types as classified by the Sydney system.

3.1 Stomach Anatomy and Function

The information detailed in this section is derived from the book "Histology for Pathologists (2nd Edition)" (Sternberg, 1997, pp. 481-493). The stomach, situated in the upper left quadrant of the abdomen, exhibits a J-shaped structure and serves a pivotal role in digestion. It forms a connection with the esophagus at the upper end and extends to the duodenum, the initial segment of the small intestine, at its lower end.

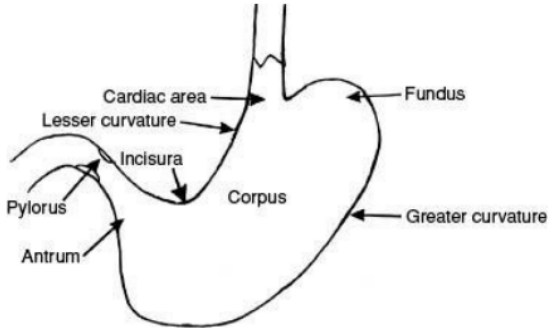


Figure 2: Gross anatomical zones of the stomach (Sternberg, 1997, pp. 481-493)

For anatomical delineation, the stomach is categorized into four primary regions:

1. **Cardia:** Adjacent to the connection point with the esophagus.
2. **Fundus:** Located above the cardia and directly beneath the left diaphragm.
3. **Corpus (or Body):** Represents the expansive central portion of the stomach.
4. **Antrum:** Constitutes the distal third of the stomach, preceding the duodenum.

In terms of its functionality, the stomach serves as a holding and mixing tank for food, thereby initiating digestion. It generates several gastric secretions, such as acid, pepsin, and electrolytes, which are pivotal in the food breakdown process. The regulation of these secretions is twofold, involving both the nervous system and gastrin, a hormone originating from the antrum of the stomach. These secretions are instrumental in preserving a sterile condition within the stomach, thereby eliminating the majority of microorganisms consumed.

Furthermore, the stomach releases gastric mucus, a vital element in safeguarding the stomach lining against its acidic surroundings and potentially aiding in lubrication. Mucin, found within the mucus, establishes a protective barrier, preventing acid from diffusing back and thus, shielding the stomach from potential harm. The formation of this mucus barrier is indispensable for the safeguarding and efficient operation of the stomach.

3.2 Gastritis

Gastritis stands as a common and frequently underestimated illness, described as a persistent and covert condition impacting the lining of the stomach (Sipponen and Maaroos, 2015). Studies estimate that over half the world's population experiences this disease in varying degrees, suggesting that hundreds of millions globally may

suffer from some form of chronic gastritis (Sipponen and Maaroos, 2015). A pivotal revelation in understanding gastritis has been the identification of the bacterium Helicobacter pylori as the culprit in a majority of gastritis cases, barring some instances of autoimmune origin (Sipponen and Maaroos, 2015). Successfully eradicating H. pylori can lead to the restoration of the gastric mucosa, especially in scenarios where the gastritis has not reached atrophic final stages (Sipponen and Maaroos, 2015). The ramifications of chronic gastritis are profound, playing a significant role in the development of peptic ulcers and gastric cancers, thereby contributing to millions of premature deaths globally each year. Despite its widespread prevalence and serious outcomes, numerous facets of chronic gastritis, including how it progresses and the specific mechanisms culminating in complications like ulcers and cancer, are still largely unexplored (Sipponen and Maaroos, 2015).

Based on the Sydney System and "Histopathologie" (Thomas, 2001, pp. 140-141), gastritis is typified into Type-A, Type-B, and Type-C, each delineating different causes and characteristics. Type-A Gastritis is immunological, representing about 3% of all chronic gastritis instances, with the majority showcasing antibodies that inhibit acid production. In contrast, Type-B Gastritis is the most prevalent, forming over 90% of all cases, and its occurrence significantly increases with age. The presence of Helicobacter pylori in the stomach's lining is a defining feature of Type-B Gastritis, causing inflammation that originates in the antrum and extends to the corpus. Type-C Gastritis, on the other hand, predominantly emerges in the antrum, especially in proximity to anastomoses of a partly resected stomach, or following exposure to substances such as non-steroidal anti-inflammatory drugs or alcohol abuse.

4 Literature Review

The literature pertinent to this project can be broadly classified into two main categories: clinical literature, which facilitated a deeper understanding of the medical aspects and was instrumental in establishing objective criteria for annotations by pathologists, and technical literature, which offered insights into the development of the dataset, annotations, and the baseline model.

4.1 Clinical Literature

Clinical literature serves as the backbone for gaining medical insights and fostering an understanding of gastritis and its classifications. The primary textbooks consulted for this purpose include "Histology for Pathologists" (Sternberg, 1997), "Histopathologie" (Thomas, 2001), and "Histologie" (Schiebler and Korf, 2007). These sources were either recommended by the collaborating pathologists, drawing from their extensive experience and familiarity with the content, or were selected after thorough discussions to ascertain the quality and relevance of the information

provided. This literature forms the foundation for medical knowledge and understanding, and it played a crucial role in aiding pathologists in developing objective annotation criteria.

4.2 Technical Literature

On the technical front, the project found a close resemblance with the study conducted by (Steinbuss et al., 2020). Focusing on the identification of gastritis subtypes through histological images of antrum and corpus biopsies, the challenges faced in this study, particularly concerning limited data size, were analogous to those encountered in our project. This similarity rendered the study a valuable foundation. It also guided towards the utilization of QuPath, which is further detailed in section 7.1.3, and experimentation with the Xception architecture, discussed in depth in section 8.1.

Although there is an abundance of research projects leveraging AI for medical imaging, the search did not yield any other paper specifically dealing with the gastric region and gastritis, with one exception. The paper by (Martin et al., 2020) also analyses the gastric region and Helicobacter pylori. However, it employs HALO-AI, a comprehensive software suite specifically designed to facilitate workflows starting from annotations and progressing to model training and analysis. This suite provides a different approach compared to the methods used in this thesis, focusing on a streamlined and integrated workflow for developing AI models in medical imaging (Hal). Unfortunately, this software suite was not publicly accessible, and therefore, utilizing it was not an option for this thesis.

5 Process

This section outlines the structured approach undertaken in the initial stages of the project, emphasizing the coordinated efforts between different stakeholders and the strategic planning involved. It further delineates the various challenges encountered during the project's execution and the solutions implemented to address them, providing a comprehensive view of the methodological aspects of this project.

5.1 Original Planning

At the inception of the project, meticulous planning was undertaken to establish a solid foundation for both the current thesis and future work beyond the current scope. This planning phase encompassed the following critical steps:

1. **Coordination between Hospital and University**

   The initial phase of the project was marked by establishing a coordinated effort between the hospital and the university. This coordination was vital

in ensuring a seamless flow of information and resources necessary for the progress of the project.

2. **Communication between Medical Staff and Developers**

   On the first day of the project, the developer and student team were given a tour around the pathology department where they were shown the process of creating slides of the same type that would later be digitized and analyzed. This experience not only provided a hands-on understanding of the initial stages of the workflow but also fostered a deeper appreciation of the intricate details involved. Following the tour, an inaugural planning meeting was held, led by Dr. Bettina Braunecker and Dr. med. Volker Mordstein. This meeting aimed to delineate the project's scope, timeframe, and to foster an initial understanding of the core topics of antrum, corpus, and gastritis.

3. **Outline of the System with Input and Output**

   In subsequent meetings, the team engaged in detailed discussions to craft a blueprint of the system, including the precise nature of its inputs and outputs. A crucial part of this phase was the training of developers in the classification of various gastric regions like the antrum, corpus, and intermediate zones, facilitated through the use of textbooks and hands-on practice with microscopes. Though a similar approach was initially intended for understanding gastritis, the overwhelming amount of information necessitated a shift in strategy. The learning process for gastritis was thus spread out over the duration of the project, utilizing both textbooks and experiences gained during the annotation to build a robust understanding gradually.

This foundational planning stage ensured a well-coordinated and informed start to the project, laying down the groundwork for the following months and for possible further work.

5.2 Project Challenges

During the course of this project, several challenges were encountered, testing the limits of available resources and expertise. Below, the most significant hurdles and the compromises that had to be made are delineated:

1. **Relatively Small Dataset**

   The dataset at our disposal was relatively small, a situation exacerbated by the limited time frame, the finite number of available slides for digitization, and the considerable time required for both scanning a slide and annotating the WSIs. This limitation imposed restrictions on the depth of analysis we could undertake, potentially affecting the robustness of our conclusions. Further details on this challenge are discussed in section 6.

2. **Complexity of Gastritis and Inflammation Details**

   The nuances of gastritis and inflammation presented a complex landscape to navigate. The intricate details were too complex to be fully integrated into the project, necessitating compromises that might have reduced the granularity of our analysis. As described in section 6, it was decided to limit the scope of the model to distinguishing between inflamed and non-inflamed tissues, foregoing the finer classification into different gastritis types and other inflammations. However, it is noteworthy that the gastritis type is indicated in the name of the WSIs, providing a pathway for future work to delve into more detailed classifications leveraging this information.

3. **Trade-off Between Resolution and Image Size**

   A critical balancing act was required to manage the trade-off between image resolution and size. High-resolution images offer more details but come at the cost of increased file size, which can hinder processing speed and resource allocation. To mitigate this, a strategic decision was made to employ a tile-based model later in the project. This approach allowed us to retain the full detail of the images while keeping the storage requirements relatively low. However, this was a challenging decision as pathologists are traditionally trained to classify using the entire slide, which facilitates a more comprehensive diagnosis. During the annotation process, as described in section 7.1, the pathologists worked at the WSI level, not the tile level. It was a substantial shift to then translate these WSI level annotations into tile-based annotations, maintaining the integrity of the original classifications while adapting to a new format.

4. **Proprietary Formats**

   The project involved handling proprietary formats, specifically the .mrxs file format which is utilized for storing high-resolution multi-layered images, often used in DP for preserving detailed information of biological samples. This format could potentially limit the interoperability with other systems and software, posing a challenge in manipulating and analyzing the data across diverse platforms. Fortunately, this hurdle was overcome during the tile creation process, where a specially developed QuPath script was employed to facilitate the handling of .mrxs files, ensuring smooth data processing and analysis. This solution is elaborated upon in section 6.4.1.

5. **Inconsistent Scanning Quality**

   The scanning process yielded images with varying quality, introducing a level of inconsistency in the dataset. Some scans showcased blurred regions, possibly due to issues encountered during the scanning process or the inherent quality of the original slides, as illustrated in figures 8 and 9. Despite these imperfections, the decision was made to retain these scans to not only augment the size of the dataset but also to incorporate a real-world variability that models might face in practical scenarios. However, this introduced a risk of the model becoming sensitive to these inconsistencies during training and validation, potentially

impacting its performance. These challenges and the strategies employed to address them will be further discussed in section 6.6.

6. **Color Variations in Scanned Slides**

   The scanned slides exhibited slight color differences compared to traditional microscopy. This variation made classification more challenging in specific cases, requiring pathologists to adapt to the altered visual cues and possibly affecting the accuracy of classifications.

7. **Communication Between Pathologists and Developers/Students**

   Effective communication between the pathologists and the developers/students was pivotal but also proved to be a challenging aspect. Bridging the gap between medical expertise and technological know-how required concerted efforts to foster understanding and collaboration. Despite numerous in-person meetings to elucidate medical terminology and the nuances of gastritis, the team faced a steep learning curve due to the overwhelming amount of information that needed to be processed. Nevertheless, these efforts were fundamental in ensuring that the project goals were met efficiently and effectively.

By navigating these challenges with strategic compromises and solutions, we aimed to maintain the integrity and accuracy of the project to the greatest extent possible. It is essential to keep these challenges in mind when interpreting the results, as they underline the conditions and limitations under which the project was executed.

6 Dataset

In this section, the intricate processes involved in compiling, handling, and preparing the dataset utilized in this project are delineated. Detailed insights into the different stages, including data collection and digitalization of samples are provided. This section also covers vital aspects such as anonymization, data transportation, and storage strategies undertaken to safeguard the privacy of the data while ensuring its utility for the project. Moreover, exploration of the preprocessing techniques, including data augmentation and dataset splitting, proves critical in refining the dataset for the training phase as well as ensuring comparability between models. An in-depth discussion of the limitations of the dataset, encompassing issues such as size constraints and quality variations, is also included, providing a comprehensive perspective on the dataset's scope and the challenges encountered.

6.1 Data Collection

The process of data collection for this project involved several crucial steps, from the selection of samples to their digitalization and subsequent classification. This section delves into the specifics of how the samples were digitized, the rationale behind the chosen stains, and the distribution of the dataset based on different classifications.

6.1.1 Digitalization of Samples

The selected samples were dispatched to the southern Nuremberg hospital, which houses a WSI scanner. This scanner is primarily employed for specific cases that necessitate the expertise of pathologists from the larger pathology department in the northern hospital within a short time frame or for archiving unique cases on the network.

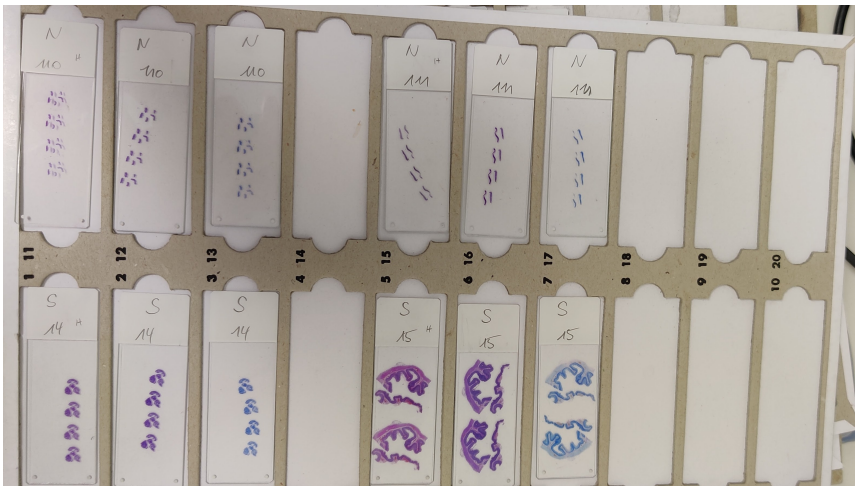


Figure 3: Four individual samples presented in a tray, with each sample prepared with the stains H&E, PAS, and modified Giemsa

After receiving instruction on the scanner's operation and the associated software, the process to scan the provided samples began. All the samples were stained using H&E, PAS, and modified Giemsa techniques. As the project progressed, the decision was made to cease the digitalization of slides stained with modified Giemsa and instead concentrate on H&E and PAS stains. Eventually, the focus narrowed further to only H&E samples, driven by the evolving requirements of the project, particularly the classification of WSIs as either inflamed or non-inflamed, but also due to time limitations that became apparent early on.

6.1.2 Distribution of Classes

The quality and comprehensiveness of the dataset are pivotal for the success of this thesis and any subsequent endeavors related to this project. Initial project planning aimed to commence with a dataset of approximately 100 samples for the antrum, corpus, and intermediate classifications. There was also a consideration to broaden the project's scope to encompass inflammatory samples for classifying WSIs as inflamed or non-inflamed. All samples were curated and supplied by the pathologists at the Nuremberg hospital over the duration of this project, spanning from May to September 2023. The swift advancement in dataset development and the concurrent annotations led to the inclusion of more samples, notably the inflammatory ones.

Currently, the project boasts a digitalized collection of 205 WSIS. In total, there are 156 samples containing corpus particles, 108 samples with antrum, 35 with intermediate and 9 samples that could not be classified as antrum/corpus/intermediate. For the inflammatory classification, there are 97 non-inflamed and 108 inflamed samples (55 Type-B, 40 Type-C, 13 others). It is important to mention, that some WSIs have more than one gastric region present, explaining the difference between total sample count and the count per class.

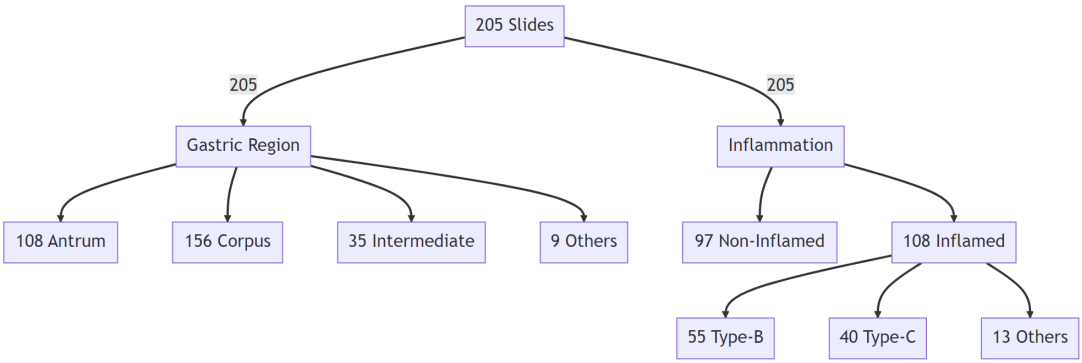


Figure 4: Class distribution visualized for both classifications

6.2 Data Anonymization

Anonymization is crucial in this project to ensure the highest level of patient privacy and confidentiality. Given that the dataset consists of genuine medical samples from individuals, it's imperative to remove any identifiable information that could trace the data back to specific patients. The complete elimination of personal identifiers was not only essential for preserving privacy but also a mandatory requirement for transporting the data outside the hospital's network. This allowed the data to be stored on the chair's file server, the developer's personal machines, and to be utilized for this project. To achieve this, every identifier from the samples was removed, and the slides were labeled with incrementing numbers starting from 1. For the inflammatory cases, the letters 'B' or 'C' were appended after the number to indicate the type of Gastritis. The letter 'S' was used for cases that show irregularities neither falling under Type B nor C, and the letter 'K' for non-classifiable cases. These non-classifiable cases were digitalized, but not fully annotated and thus not used for training and testing. Additionally, the samples were amassed throughout the project's duration, with numerous biopsies conducted daily, further diminishing the potential to trace samples back to patients. This method ensured that the dataset was entirely anonymized, allowing the data to be transferred out of the hospital without any possibility of linking it back to individual patients.

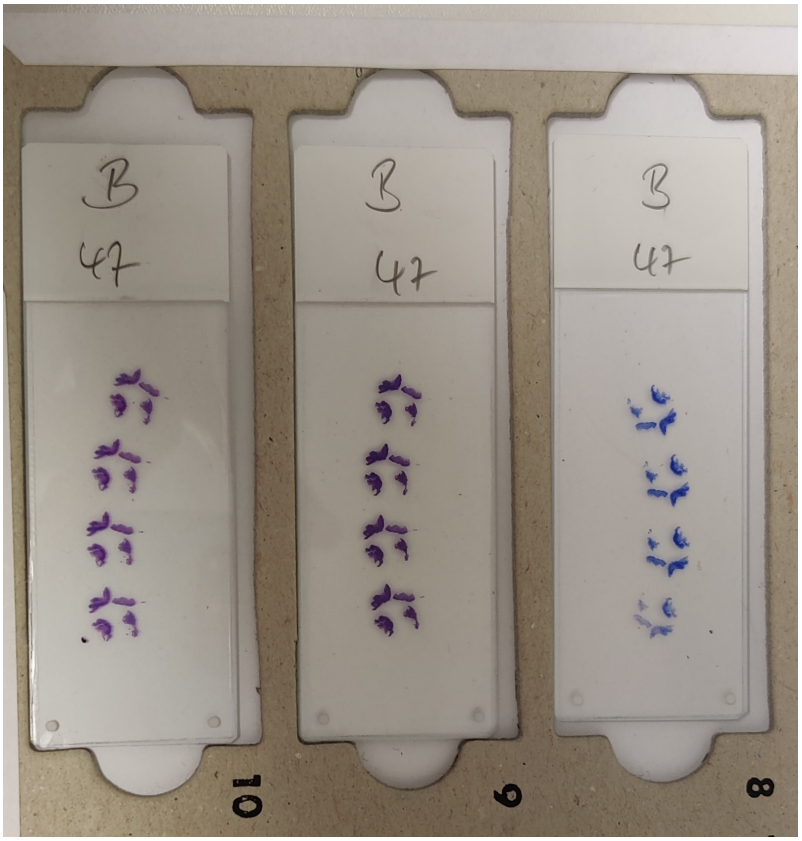


Figure 5: Three slides of the 47th Type B Gastritis case, displaying different stains (H&E, PAS, modified Giemsa), with no additional identifiers

6.3 Data Transportation

In the early stages of project planning, two primary methods for transporting data from the hospital to the chair's file server and potentially to the developer's personal machines were considered.

The first method entailed uploading the data directly from the hospital's network to a cloud service provided by the chair. The second method involved manually transporting the data using a hard drive. While the former was initially favored due to concerns about potential unauthorized access to a physical hard drive, the latter became more appealing when considering the encryption and anonymization measures that could be applied. Encrypting the hard drive combined with the dataset's anonymization would mitigate the risks associated with the drive being lost or stolen. The data, having been stripped of identifiable information, would pose minimal risk even if accessed.

However, bureaucratic challenges within the hospital meant that while obtaining the hard drive took several weeks, securing permissions to upload data from the hospital's network to the cloud would have taken even longer. Given these circumstances, the decision was made to use a single hard drive for data transportation. Contrary

to initial plans, the provided drive with the copied images was not encrypted to enhance security. Nevertheless, the thorough anonymization of the dataset meant that the risk associated with using an unencrypted drive remained relatively low, given the absence of identifiable data.

As highlighted in section 6.1.1, the slide digitization scanner is stationed in the southern pathology department. Our primary contact, Dr. Bettina Braunecker, operates mainly from the northern department. Fortunately, the hospital's integrated network, encompassing both the southern and northern departments, granted her access to the data. This interconnected infrastructure eliminated the need for additional data transportation within the hospital premises.

6.4 Data Storage

As mentioned in section 6.3, all digitalized images are originally stored on the hospital network and then transported on the chair's file server and the personal machines of the developers and the hard drive provided by the hospital, which will be returned by the end of Philipp Andreas Höfling's project. Both the hospital network and the hard drive only store the original scanned files from the WSI scanner and no processed data or processed images. The fileserver and the personal machines store the original files, the exported tiles and the annotations stored in JSON files.

Initially, an attempt was made to directly export the images using the Pannoramic Viewer in TIFF format. This approach seemed straightforward and promised a direct extraction of the required image segments. However, it soon became evident that the images exported through this method were of an exceedingly large size. Adjusting the export settings in the Pannoramic Viewer presented a dilemma: while reducing the image size to more manageable dimensions, the quality of the images deteriorated significantly. Important details within the images, crucial for accurate analysis and model training, were lost. The vast dimensions of the unadjusted images not only consumed excessive storage but also made it computationally intensive to handle, process, and train the model. Given these constraints and the inefficiencies introduced by the oversized images or the loss of quality, this method was deemed unsuitable for the project's requirements. It underscored the need for a more optimized approach, leading to the exploration and eventual adoption of the tile extraction method via QuPath.

6.4.1 Tile Creation

As discussed in section 2.2, WSIs are renowned for their vast dimensions and intricate details, capturing the full spectrum of histological features within a tissue sample. However, processing and analyzing these large-scale images in their entirety can be computationally intensive and time-consuming. To address this challenge, the practice of segmenting WSIs into smaller, manageable tiles has been adopted. Creating tiles from WSIs offers several advantages. Firstly, tiling ensures that high-resolution

details are preserved, enabling precise microscopic evaluations at the cellular or even sub-cellular level. Furthermore, by working with tiles, algorithms, especially deep learning models, can be trained more effectively, as they often require consistent input sizes. Additionally, tiles facilitate easier data storage, transfer, and sharing, as individual tiles can be accessed or transmitted without the need to load the entire WSI. In essence, tiling WSIs streamlines both the computational and analytical processes, ensuring that the richness of the data is harnessed efficiently and effectively.

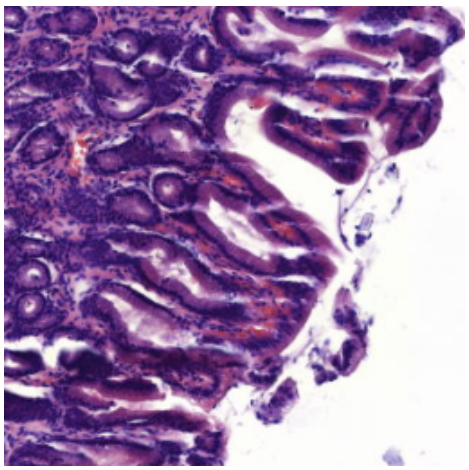


Figure 6: Tile of an antrum slide

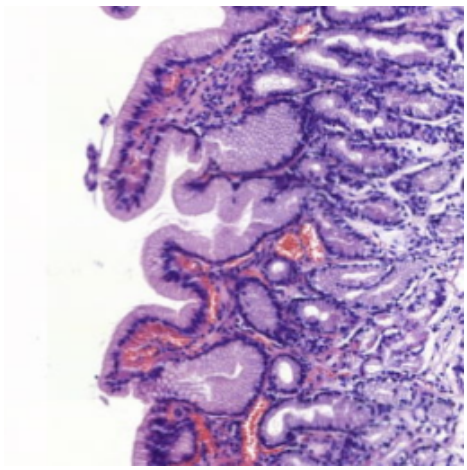


Figure 7: Tile of an inflamed slide

Leveraging the scripting capabilities of QuPath, a custom script, inspired by (Senft, 2022), was crafted to streamline the tile extraction process from WSIs. This script is designed to export image tiles in the PNG format, with a specific downsample resolution set to 10, ensuring detailed representation. Upon execution, the script identifies all annotations present in the image. For every identified annotation, it establishes a unique directory to store the corresponding tiles. A region request is subsequently formulated to encompass the precise area of the annotation. The TileExporter class, integral to the script, is configured with parameters such as a tile size of 256 pixels and an overlap of 64 pixels. This class is responsible for determining the dimensions of each tile and ensuring that only tiles with annotations are exported. Once the tiles are generated, an additional step in the script renames each tile, embedding the classification type from the annotation into the filename. This ensures a uniform naming convention across all tiles. The culmination of this process results in the segmentation of the WSI into distinct, classified regions, paving the way for efficient and organized subsequent analyses. The script along with instructions on how to use it are stored on the file server.

6.5 Preprocessing

The preprocessing stage was pivotal in enhancing the quality of the dataset and ensuring that the model was trained on reliable and diverse data. This stage involved several crucial steps, which are detailed below:

1. **Removing Empty Tiles:** The initial step in preprocessing was to remove tiles that were predominantly empty, characterized by having more than 90% of their pixels being white or colors similar to white. A Python script was employed to identify and remove these tiles, utilizing the numpy library to analyze pixel data in each tile and discard those meeting the criteria for being considered "empty." This step was vital in ensuring that the model would be trained on substantial and relevant data, focusing on meaningful patterns and structures present in the non-empty tiles.

2. **Dataset Split:** The final preprocessing step involved splitting the dataset into training, validation, and test sets. This division allowed for a structured approach to training and validating the model, facilitating a fair comparison with Philipp Andreas Höfling's model. The dataset was carefully divided to ensure a balanced representation of different classes in each subset, thereby aiming to achieve a model that is well-trained and validated across various data points. A CSV file with the detailed split is vailable on the file server.

3. **Data Augmentation through Rotation:** To augment the dataset and introduce more variability, each image in the training dataset was rotated at three different angles: 90, 180, and 270 degrees. This process generated three new augmented images for every original image in the dataset. The Python script utilized for this purpose leveraged the PIL library to apply the rotations and save the newly generated images with appropriate filenames indicating the angle of rotation applied. This augmentation strategy aimed to make the model more robust by training it on a more diverse set of data.

This comprehensive preprocessing stage laid a strong foundation for the subsequent model training and validation phases, ensuring the model had a rich and diverse dataset to learn from, which is essential in building a reliable and robust machine learning model.

6.6 Dataset Limitations

While the collected data is comprehensive in its scope, it is not without limitations. A primary constraint is its size. Due to resource and time constraints, it wasn't feasible to significantly expand the collection. A more extensive set of samples would have offered a stronger foundation for training and validation, potentially enhancing accuracy and generalizability. This challenge mirrors that of other research endeavors, such as the one documented by Steinbuss et al. (2020), where

limited data samples were available. Incorporating scans from a diverse array of institutions and scanners could potentially augment the utility of the research in real-world scenarios, given the variance in the tools and processes employed across different establishments. A richer dataset would not only enhance the accuracy of the outcomes but also foster a higher degree of generalizability.

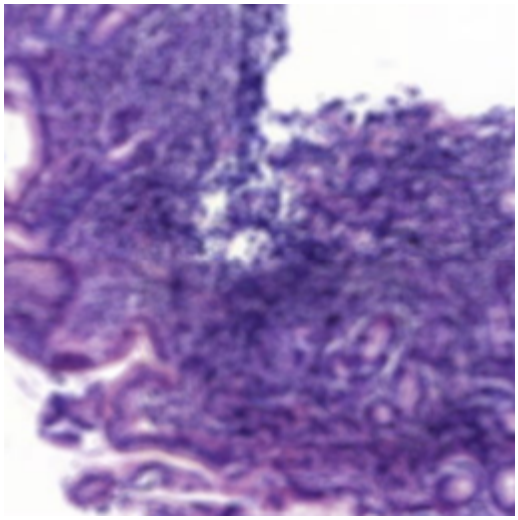


Figure 8: A tile extracted from a WSI showcasing a region with blurred details, indicative of potential issues during the scanning process or the quality of the original slide.

Another notable limitation pertains to the quality of the scans. While the majority of the scans are of high resolution and clarity, there are instances where the quality varies. Some scans appear blurred, potentially due to issues during the scanning process or the inherent quality of the original slides. Despite these imperfections, the decision was made to retain these scans in the data collection. Including them serves a dual purpose: it increases the size of the collection and introduces a level of real-world variability, which models might encounter in practical applications. However, it's essential to acknowledge that these variations in quality could impact the model's performance, especially if the model becomes sensitive to these inconsistencies during training and validation.

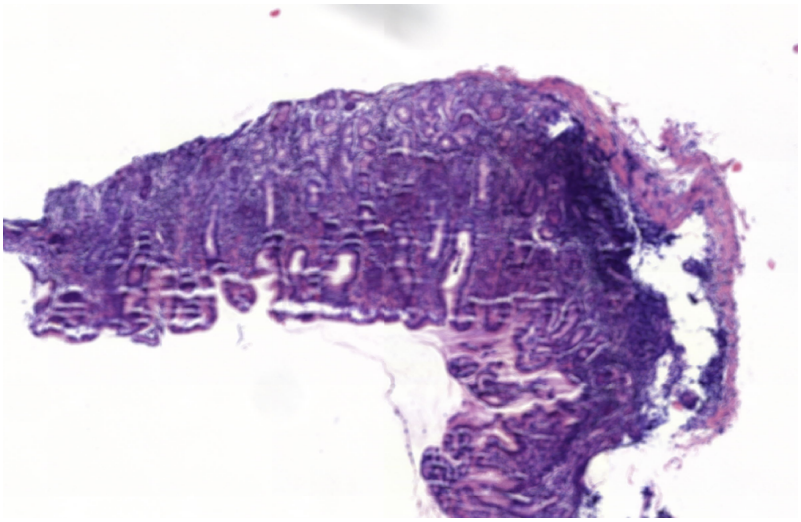


Figure 9: A region of a WSI illustrating a larger perspective of the blurred area

Furthermore, with the limited size of the collection, not all edge cases can be covered. This is especially true for capturing the full spectrum of variations in inflammations, other diseases, or regions of the stomach such as antrum, corpus, and intermediate zones. Each of these regions has distinct characteristics and functions, and a more extensive dataset would have been beneficial in representing the nuances and variations present in each region.

7 Annotations

Another crucial step to building a foundation for both this project and any future work on this project was a careful annotation process. This section will give some insights into the process itself and how classifications were tried to be added as objectively as possible. There will be a quick overview in how annotations were created using the expertise from pathologists and then added using QuPath as well as a brief discussion on the existent challenges that were faced during this process.

7.1 Annotation Process

In the wake of the digitalization of the slides, delineated in section 6.1.1, the annotation process was initiated with a clear strategy to rely on the expertise of seasoned pathologists for the most accurate results. The process was structured and executed in several meticulous stages to ensure precision and reliability. Below, the detailed procedure is outlined:

7.1.1 Collaborative Annotation with Expert Pathologists

Understanding the critical role of expertise in the endeavor, the project heavily leaned on collaborative annotation facilitated by expert pathologists, with students working closely to ensure accurate and well-documented annotations. Dr. Bettina Braunecker maintained and updated a living Word document to streamline this process, enumerating the detailed annotations corresponding to each slide, thereby serving as a reliable repository that guaranteed the most current and precise data.. This method was pivotal in upholding a high standard of accuracy from the outset. Furthermore, complex cases requiring second opinions were addressed through collaborative reviews with Dr. med. Volker Mordstein, leveraging the combined expertise of two seasoned professionals to ensure the highest possible accuracy in annotations.

7.1.2 Annotation Protocol

The annotation protocol for this project was meticulously formulated with the guidance of Dr. Bettina Braunecker and Dr. med. Volker Mordstein to foster an objective delineation of different regions of the gastric mucosa: the corpus, the antrum/pyloric, and the intermediary zone. The criteria were drawn by the pathologists from the insights presented in "Histology for Pathologists" (Sternberg, 1997) and "Histologie" (Schiebler and Korf, 2007).

The criteria outlined for each region are as follows:

Corpus

- Foveolae are shorter, less than one-third of the mucosal thickness.
- Foveolae exhibit a dense and straight structure with limited branching and a narrow lumen.
- The basal mucosa predominantly houses glands, mainly comprising chief cells secreting pepsinogen.
- The isthmus region is characterized significantly by parietal cells exhibiting eosinophilic properties and facilitating the secretion of acid and intrinsic factors.
- The neck region harbors both chief and parietal cells, alongside mucus cells.

Antrum/Pyloric

- Foveolae are longer, about half of the mucosal thickness.
- The glands are predominantly mucus-producing.
- Foveolae demonstrate a loose and convoluted structure.

- The region is devoid of chief cells, with a minimal presence of parietal cells.
- Possible presence of Brunner's gland-like cells.

Intermediary Zone The intermediary zone is defined without clear criteria, presenting characteristics partially aligned with both the corpus and antrum regions. It manifests through significant parietal cells presence yet having elongated foveolae, or shorter foveolae but with a looser structure and a higher concentration of mucus-producing cells.

This protocol, grounded in expert insights and authoritative texts, serves as a robust framework for the annotation process, ensuring high objectivity in distinguishing the distinct regions of the gastric mucosa.

During the final stages of the project, the criteria used to classify Gastritis was put into text form. While the thesis currently only distinguishes between inflamed and non-inflamed samples, these objective points can be used for future projects aiming to add finer classifications for the different types of gastritits.

Type B Gastritis Annotation Criteria:

- Presence of a neutrophil-rich infiltrate with chronic inflammation consisting of lymphocytes and plasma cells.
- Detection of rod-shaped bacteria on the surface epithelium in modified Giemsa staining.
- Absence of significant additional pathological alterations such as atrophy, glandular cysts, metaplasia, or dysplasia.

Type C Gastritis Annotation Criteria:

- Inflammation-poor antral mucosa.
- Stromal fibrosis.
- Foveolar hyperplasia.
- Signs of enhanced epithelial regeneration.
- Few eosinophilic granulocytes.
- Absence of Helicobacter bacteria in Giemsa staining.
- Absence of significant additional pathological alterations such as metaplasia or dysplasia.

7.1.3 Transferring Annotations to QuPath

With the expertly annotated Word document at disposal, the subsequent step involved transferring these annotations into the QuPath platform. In QuPath, the annotations were meticulously created using a range of tools, including the wand tool which facilitated the marking of regions of interest for annotation purposes. This tool allowed for detailed and precise annotations, translating the expert insights from the Word document into digital annotations ready for further analysis.

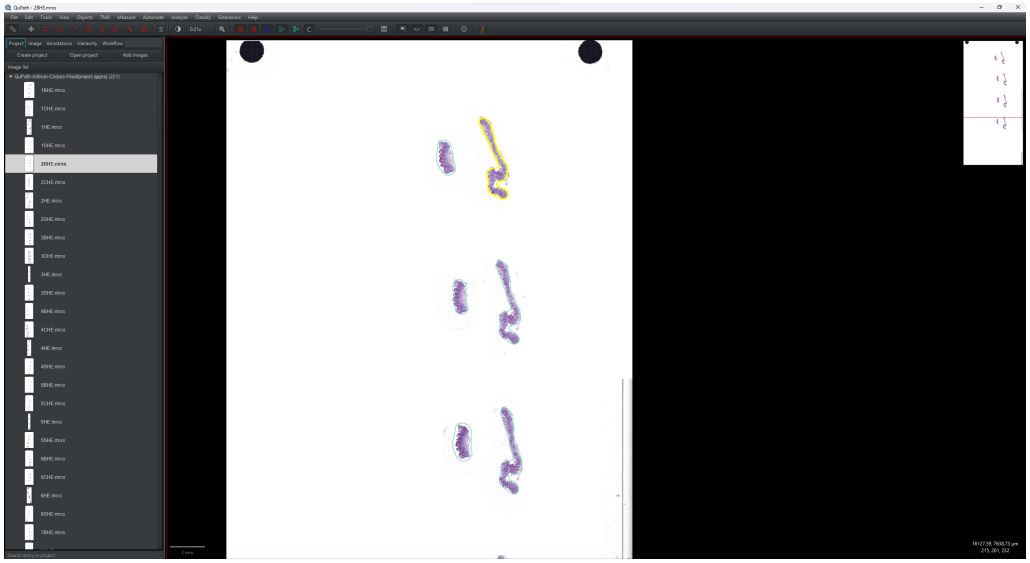


Figure 10: Open WSI in QuPath Project with an annotation selected

Through this structured and collaborative process, a robust and highly accurate set of annotations was developed. The process leveraged both the explorative efforts in the initial stages and the expert insights in the later stages to create a rich and reliable dataset for the subsequent phases of the project. The meticulous transfer of annotations into QuPath ensured that the dataset was not only accurate but also primed for detailed analysis in the forthcoming stages of the research.

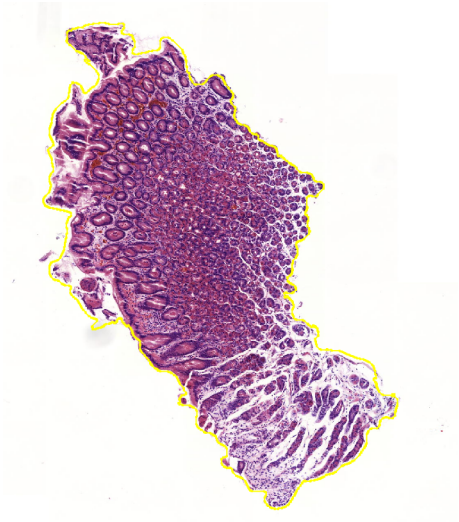


Figure 11: Detail of a particle selected for annotation using the wand tool in QuPath, showcasing the precision of the selection

7.2 Challenges in Annotation

The annotation process, a pivotal step in the development of the dataset, encountered several challenges that needed meticulous addressing to maintain the integrity and reliability of the data. These challenges stemmed from a variety of factors including the inherent subjectivity in defining classes, the time-intensive nature of the batch-wise annotation process, and logistical issues related to schedules and locations. The following subsections delve deeper into these challenges, providing a comprehensive view of the hurdles encountered during the annotation process.

7.2.1 Subjectivity in Class Definitions

A primary challenge was the absence of perfectly objective definitions for the classes. Despite leveraging established references like the books by Sternberg and Schiebler to guide the classification, the intermediary zone remained particularly challenging to define with absolute objectivity. The classifications had to rely partly on the criteria for corpus mucosa and partly on those for antrum mucosa, navigating a delicate balance to ensure accurate representation. This necessitated a protocol grounded in expert knowledge and experience to navigate the nuanced landscapes presented in the histological slides.

7.2.2 Time-Intensive Batch-wise Process

The annotation process was conducted in batches, each taking several weeks to complete. This batch-wise approach, while necessary to maintain a high degree of precision and attention to detail, inevitably prolonged the overall timeline of the

project. The slow pace of progress necessitated a sustained commitment to maintain consistency and quality across different batches, mitigating the risk of discrepancies and errors over time.

7.2.3 Logistical Challenges

Further complications arose from the different schedules and geographical locations of the individuals involved in the annotation process. These variations restricted the availability of instantaneous feedback, an element that could have fostered a more dynamic and responsive annotation process.

7.3 Storage of Annotations

In the context of QuPath projects, a pivotal advancement has been realized through the introduction of a mechanism to store annotations in JSON files, facilitated by a QuPath script. This approach not only optimizes the storage process but also significantly enhances the ease of sharing annotations across different platforms.

The initial step in this process involves a script that identifies the URI of the current image from which it extracts the name to be used in naming the JSON file. Subsequently, the script gathers all the annotation objects in the current project. Utilizing the Gson Java library, it structures these annotations into a JSON format, which is then written to a JSON file stored in a predefined directory. Importantly, this script can be executed on individual images or applied to all images selected in QuPath, iterating over each one to facilitate a batch processing of annotations.

```
// Get the URI of the current image and extract the file name
def imageData = getCurrentImageData()
def imageURI = imageData.getServer().getURIs()[0].toString()
def imageName = imageURI.substring(imageURI.lastIndexOf('/') + 1,
    imageURI.lastIndexOf('.'))

// Get annotations
def annotations = getAnnotationObjects()
boolean prettyPrint = true
def gson = GsonTools.getInstance(prettyPrint)

// Define the output directory and construct the output file name
def outputDirectory = "path\\to\\your\\output\\directory"
def outputFileName = "${imageName}_annotations.json"
def outputPath = new File(outputDirectory, outputFileName)

// Write the annotations to the JSON file
outputPath.write(gson.toJson(annotations))
print "Done!"
```

Listing 1: Export Script that converts annotations into JSON files

An essential facet of this approach is the structured and hierarchical organization of the JSON files. These files encapsulate a wealth of details associated with each annotation, including but not limited to spatial coordinates and classification names. This meticulous organization of details not only facilitates a streamlined storage process but also ensures an efficient retrieval of information.

The JSON structure is organized as an array of objects, each representing a distinct annotation. Within each object, there are several key-value pairs that hold the vital information about the annotation. The "type" key indicates the feature type, and the "id" key holds a unique identifier for the annotation. The "geometry" key contains an object that describes the geometric details of the annotation, including the type of geometry (e.g., polygon) and the coordinates outlining the shape. The "properties" key holds an object that contains additional details about the annotation, such as the object type, classification name, and a color code representing the classification.

Beyond storage, these JSON files find utility in being re-imported into other QuPath projects, a functionality enabled through an import script. This script works by deserializing the JSON structure back into QuPath annotation objects, which are then added to the current image in the project, thereby maintaining the structural integrity of the annotations. Similar to the export script, the import script can be run on a single image or iteratively on all selected images in QuPath, enhancing the efficiency of the process. This feature significantly enhances the collaborative potential of QuPath projects, allowing for a hassle-free sharing of annotations. The full scripts for exporting and importing can be found on the file server of the chair.

Moreover, the open-standard nature of JSON files offers an adaptability, making them compatible across various environments and platforms, thereby broadening the collaborative scope of QuPath projects.

7.4 Limitations

While the annotation process facilitated by the QuPath scripts offers a robust solution for managing and sharing annotations, it is not without its limitations. These limitations primarily pertain to the specificity and granularity of the annotations, which are detailed below:

Firstly, the current classification system, which distinguishes between inflamed and non-inflamed regions, is somewhat unspecific. This broad categorization could potentially overlook the nuanced differences that might exist within these broadly defined classes. Moreover, while the type of gastritis present on a slide is known — a detail noted before the slide's digitalization — the project currently does not classify the specific type of gastritis or other inflammations. It restricts its classification to a binary distinction between inflamed and non-inflamed regions. Therefore, there is a scope for further refining the classification system to include more classes that can capture a richer variety of features and details, thereby offering a more detailed analysis of the slides.

Secondly, the annotations are applied to entire particles on the slide, rather than focusing exclusively on areas of interest, such as inflamed areas. This approach might include regions that are not pertinent to the study, thereby potentially introducing noise into the data and affecting the accuracy of the analysis. A more targeted annotation strategy, focusing on specific areas of interest, could enhance the precision of the analysis by concentrating on the most relevant regions of the slides.

Addressing these limitations in future work on the project could significantly enhance its utility, allowing for a more nuanced and detailed analysis of histological slides.

8 Baseline Model

The inception of this project was grounded on the establishment of a baseline model, serving a pivotal role in delineating the initial sketches of the system's capabilities and functionalities. This section highlights the fundamental aspects of the baseline model, setting the stage for a detailed discussion on its development, performance evaluation, challenges, and limitations. The focal points of this introductory part are as follows:

- **Purpose of the Baseline Model**: The baseline model was conceived to serve as an initial framework to gauge the potential performance and feasibility of the project. It acts as a reference point, facilitating a preliminary understanding of the system's capabilities and laying a foundation for future work. This system is particularly significant as it sets a benchmark for forthcoming advancements, notably the master thesis of Philipp Andreas Höfling.
- **Objective of the Model**: The primary objective of the baseline model is to establish a starting point that encapsulates a rudimentary yet functional system, capable of performing basic classifications. It aims to provide insights into the project's viability, offering a tangible reference for evaluating the progress and effectiveness of subsequent developments.
- **Input and Output of the System**:
  - *Input*: The system is designed to process tiles generated through the QuPath export script, a process detailed in section 6.4.1.
  - *Output*: The output mechanism is bifurcated, offering two scripts that facilitate classification at different levels - tile and WSI. These scripts and their functionalities will be elaborated upon in section 8.3.

Further into this section, each of these aspects will be explored in detail, providing a comprehensive understanding of the baseline model and its integral role in the project's trajectory.

A complete version of this project, inclusive of all scripts, datasets, and documentation, is archived on the file server of the chair. Additionally, the project will be made available on GitHub (https://github.com/TomH1004/Baseline-Model, however, due to GitHub's file size constraints, only the scripts and documentation will be hosted there. The trained models and datasets, given their substantial sizes, will be exclusively accessible via the file server of the chair.

8.1 Model Architectures

In the development of the baseline models, two distinct and renowned architectures, ResNet-18 and Xception, were employed for both development and comparative evaluation. ResNet-18, recognized for its efficiency and versatility in addressing a myriad of computer vision tasks, acted as a foundational model for the study. On the other hand, the Xception architecture, known for its unique design and adaptability, was deliberately chosen. Its selection was influenced by its demonstrated excellence in medical imaging and its successful utilization in a study closely related to the research domain (Steinbuss et al., 2020). Subsequent sections will delve into a comprehensive overview of the features and advantages of both the ResNet-18 and Xception architectures, elucidating their significance in the context of the present research.

ResNet-18 ResNet-18 is a variant of the Residual Network (ResNet) architecture, introduced to address the vanishing gradient problem and facilitate the training of deeper neural networks. The information for this paragraph is based on the initial paper "Deep Residual Learning for Image Recognition" by (He et al., 2015). The architecture of ResNet-18 is characterized by the use of skip connections or shortcuts, which bypass one or more layers during the forward and backward passes. These shortcuts are essential for preventing the degradation of the training accuracy as the network depth increases.

The ResNet-18 model consists of an initial convolutional layer followed by several residual blocks and finally, a fully connected layer. Each residual block in ResNet-18 contains two convolutional layers with batch normalization and ReLU activation functions. The use of batch normalization ensures that neither forward nor backward signals vanish, making it possible to train very deep networks.

One of the key features of ResNet-18 is its efficiency. The architecture is designed to have lower complexity compared to other deeper ResNet models, such as ResNet-50, ResNet-101, and ResNet-152, while still maintaining competitive accuracy. The ResNet-18 model achieves comparable accuracy to its 34-layer counterpart, but with faster convergence, demonstrating the effectiveness of residual learning in optimizing deep networks.

In conclusion, ResNet-18 is a powerful and efficient deep learning architecture that leverages residual learning to train deep networks, making it suitable for a variety of computer vision tasks.

Xception The Xception architecture emerges as a pivotal and competent model for medical imaging, attributed to its unique design and adaptability (Steinbuss et al., 2020). It represents a modification of the Inception architecture, where, notably, "the Inception modules have been replaced with depthwise separable convolutions" (Steinbuss et al., 2020). This strategic alteration amplifies the model's capacity to dissect and interpret intricate image data, making it exceptionally suited for in-depth medical imaging studies (Steinbuss et al., 2020).

With its proven excellence on diverse datasets like ImageNet and its success in classifying various clinical images including tghe risk grading of skin tumors (yu Zhao et al., 2019), Xception highlights its adaptability and effectiveness in the medical field (Steinbuss et al., 2020). The architecture's incorporation in the research by Steinbuss et al. is a testament to its invaluable role in analyzing histological images of antrum and corpus biopsies and advancing medical imaging research (Steinbuss et al., 2020).

8.2 Training and Validation Process

In the implementation phase, flexibility was a key consideration, ensuring the ability to swiftly alternate between different neural network architectures, such as ResNet18 and Xception, by adjusting the code. This adaptability facilitated exploration and evaluation of various architectures.

To modify the classification task, changing the directories in the script sufficed, and the classes were automatically discerned from the data within the specified directories. This approach streamlined the adaptation to different datasets and classification tasks, essential for thorough model exploration.

Each model and classification underwent training for a total of 10 epochs, maintaining uniformity across different scenarios and enabling fair comparison. During these phases, a variety of metrics were tracked to assess performance, with further details to be provided in Section 8.5. The models employed were all pretrained, providing a solid basis for the subsequent training process.

Importantly, the script was exclusively tested with binary classifications. Although the gastric classification includes an intermediate class, this class was not included in the model training. Instead, it was evaluated later based on a threshold among tiles for each particle. This decision stemmed from insights gained during the developmental testing, aiming for an optimized approach to classification.

For further details and comprehensive insights, additional documentation will be provided both on the GitHub repository and on the file server of the chair, facilitating a deeper understanding and enabling future work on the project.

8.3 Tile-level and WSI Classification

The process of classification in the project is bifurcated into two principal paths: classification at the tile level and classification at the WSI level. Both paths are

driven by scripts that are structured to manage specific levels of classification, leveraging the two architectures mentioned in section 8.1 to predict classifications based on trained models.

8.3.1 Tile-level Classification

The tile-level classification script begins by setting up the computational device and crafting the transformation pipeline for the images. Following this, a graphical user interface (GUI) is used to facilitate the selection of multiple files for classification. Once the files are selected, each image undergoes individual classification where its RGB channels are isolated, and transformations are applied to comply with the input specifications of the pre-trained model.

In the output phase, each tile is assigned a predicted class along with associated probabilities for each class category, derived from the softmax function applied to the model's output. To bolster the reliability of the predictions, a threshold of 0.6 is instituted. If a prediction falls below this threshold, it is categorized as uncertain for classifications involving inflamed and non-inflamed categories, and as intermediate when distinguishing between corpus, antrum, and intermediate categories. It is essential to tailor this setting to align with the specific type of classification being undertaken. Once all the selected images have been processed, the script generates a detailed report that outlines the predictions for each image and provides a summary of the total occurrences for each class.

It is important to note, that the evaluation of the script to classify individual tiles is very limited beyond the evaluation on the entire test set, because the utility of classifying individual tiles is very limited and was mostly used for testing purposes during development.

8.3.2 WSI-Level Classification

In contrast to tile-level classification, WSI-level classification operates on a directory containing folders that represent different annotations of a WSI. Once the main directory is selected through the GUI, the script iterates over each annotation folder, individually classifying each image within these folders, and stores the predicted classes in a list.

Gastric Region Classification To achieve an overall classification for each annotation folder, a majority vote is applied with a threshold of 0.75. If no class reaches the threshold in terms of proportion, the classification defaults to "intermediate" for the classification of gastric regions. Following the classification of all annotation folders, the script leverages a counter to derive the classes, that were predicted at least 3 times to filter out misclassifications of single annotation folders, presenting a comprehensive view of the cell types present in the WSI based on the identified classes.

Inflammation Classification While the majority of this classification is similar to the gastric region classification, there are some noteworthy differences. Because there is no intermediate class between inflamed and non-inflamed, there is also no need for a threshold that needs to be met. Instead, a simple majority vote decides whether a annotation folder is classified into either inflamed or non-inflamed. For the overall classes present in the WSI, the same counter is used to filter out single misclassifications.

8.3.3 Input and Output

Tile-level Classification

- **Input:** Paths to the individual tile images selected through the GUI.
- **Output:** Predicted class and the associated probabilities for each possible class for each tile. A detailed report summarizing the predictions and probabilities for all selected tiles.

WSI-level Classification

- **Input:** Path to the main directory containing annotation folders selected through the GUI.
- **Output:** Predicted class for each annotation folder based on the logic explained in section 8.3.2 along with a comprehensive summary of the cell types or existence of inflammation present in the entire WSI.

8.4 Challenges

In the development and optimization of the baseline model, several challenges were encountered. These challenges, which are pivotal points of focus in ongoing research, are outlined below:

Limited Experience on Improving Convergence
One primary challenge faced was limited experience in enhancing the convergence of the model during the training phase. Limited experience in this area posed a challenge in selecting the optimal hyperparameters such as learning rate, batch size, and the appropriate activation functions that foster faster and more stable convergence. Moreover, devising strategies to overcome issues hindering convergence, proved challenging. Delving deeper into understanding the underlying mechanics of convergence and experimenting with different optimization techniques to facilitate smoother convergence in future iterations of the model is imperative.

Avoiding Overfitting
Overfitting was observed across both classifications and all models, presenting a notable challenge. It manifested as the models excelling on the training data but

struggling to generalize effectively to unseen data. Overcoming this issue proved to be a complex task and will be a significant focus of further work, necessitating exploration and the development of strategies to ensure model robustness and performance across diverse datasets.

8.5 Evaluation and Results

To comprehensively evaluate the performance of the models, a multifaceted approach was adopted, utilizing a variety of metrics. During the training phase, the models were scrutinized based on training and validation losses, as well as accuracy on the validation set. This initial assessment provided insights into the models' learning efficacy and generalization capabilities. For the test set, a more extensive set of metrics was reordered and employed, including Accuracy, Precision, Recall, F1 Score, ROC Curve, and a Confusion Matrix. As mentioned in section 8.2, all models were subjected to 10 epochs of training and validation to have comparable results.

These metrics are calculated using the sklearn.metrics and PyTorch libraries. Precision, Recall, and F1 Score are computed based on a weighted average, providing a balanced view of the model's performance across different classes. Additionally, the Confusion Matrix is used to gain insights on a class basis, helping to identify areas of strength and potential improvement for each class.

8.5.1 Performance Analysis: Classification of Gastric Regions

This section delves into a comparative analysis of the ResNet18 and Xception models, focusing on their proficiency in classifying Antrum, Corpus, and Intermediate regions, which are pivotal for accurate medical diagnoses.

ResNet18 Model: The ResNet18 model showcased a commendable learning curve, with the training accuracy witnessing a consistent ascent from 0.88 to 0.98 across the epochs, as illustrated in Figure 12. However, the validation accuracy exhibited fluctuations, oscillating between 0.86 and 0.83, before settling around 0.85. This variability in validation accuracy, juxtaposed with the steady decline in training loss from 0.3 to 0.05 and the concomitant increase in validation loss from 0.35 to 0.65, as depicted in Figure 13, raises pertinent questions regarding the model's generalization capabilities and susceptibility to overfitting.

The Confusion Matrix, presented in Figure 14, and the ROC curve, with an area under the curve (AUC) of 0.95 as shown in Figure 15, further corroborate the model's robust classification performance. However, the uniformity in Accuracy, F1 Score, Precision, and Recall at 0.89 necessitates a nuanced examination of the model's discriminative power and its ability to balance sensitivity and specificity effectively.

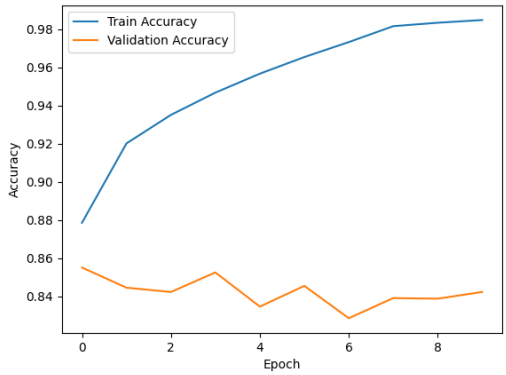


Figure 12: Training and Validation Accuracy for ResNet18 Model in gastric classification

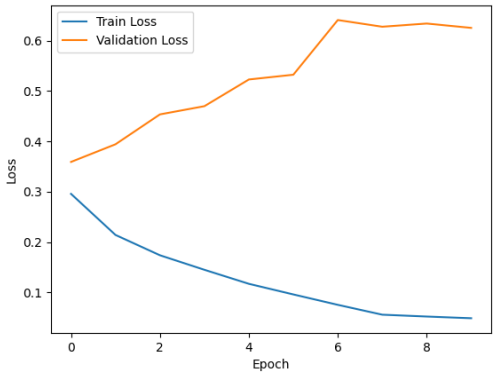


Figure 13: Training and Validation Loss for ResNet18 Model in gastric classification

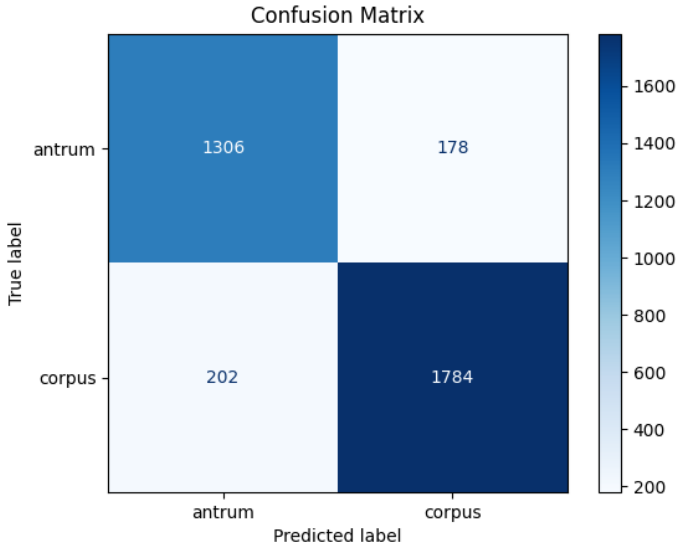


Figure 14: Confusion Matrix for ResNet18 Model in gastric classification

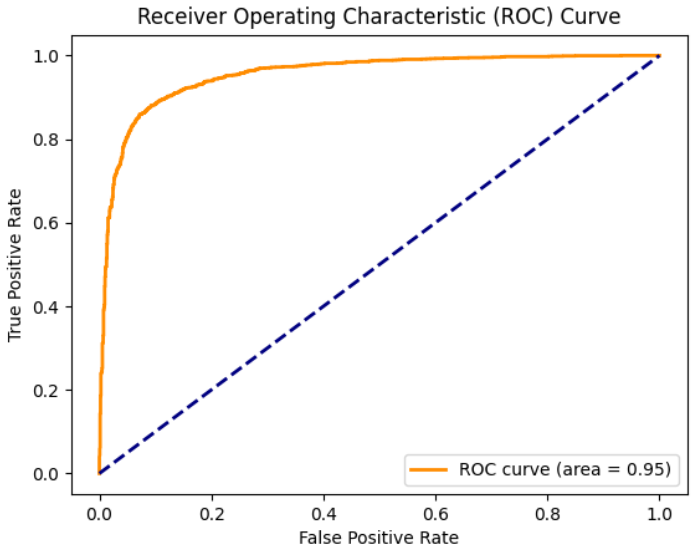


Figure 15: ROC Curve for ResNet18 Model in gastric classification

The ResNet18 model, despite its high training accuracy and commendable performance on test set metrics, exhibits signs of potential overfitting, as evidenced by the discrepancies between training and validation metrics. This necessitates further exploration into regularization techniques and model refinement to enhance its generalization capabilities and mitigate overfitting.

Xception Model: The Xception model demonstrated a progressive enhancement in training accuracy, escalating from approximately 0.87 to 0.99, as depicted in Figure 16. The validation accuracy maintained stability, hovering between 0.87 and 0.88, indicative of the model's consistent generalization performance. The training loss experienced a substantial reduction from an initial 0.3 to 0.05, while the validation loss witnessed an increment from about 0.3 to 0.65, as illustrated in Figure 17.

The Confusion Matrix and the ROC curve for the Xception model, presented in Figures 18 and 19 respectively, reveal a performance profile analogous to the ResNet18 model. This similarity in performance metrics underscores the models' comparable classification prowess, but also prompts a deeper investigation into their nuanced differences and potential areas for optimization.

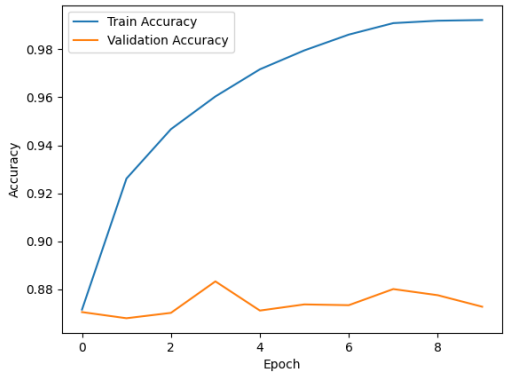


Figure 16: Training and Validation Accuracy for Xception Model in gastric classification

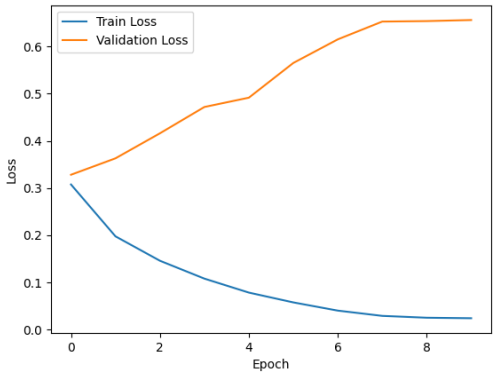


Figure 17: Training and Validation Loss for Xception Model in gastric classification

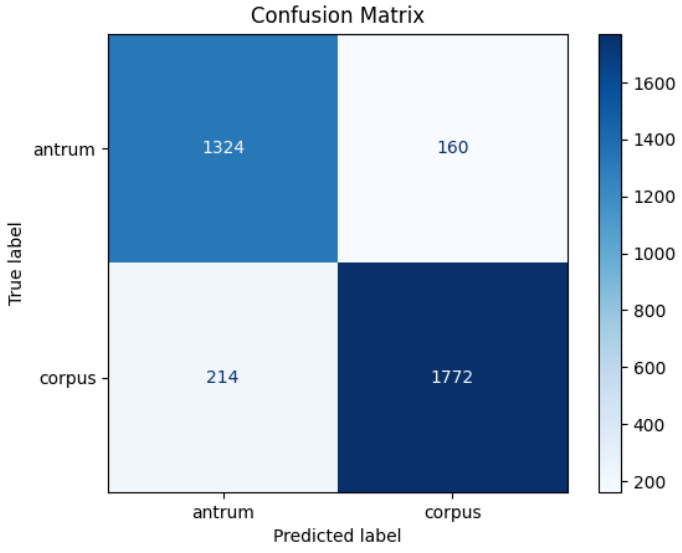


Figure 18: Confusion Matrix for Xception Model in gastric classification

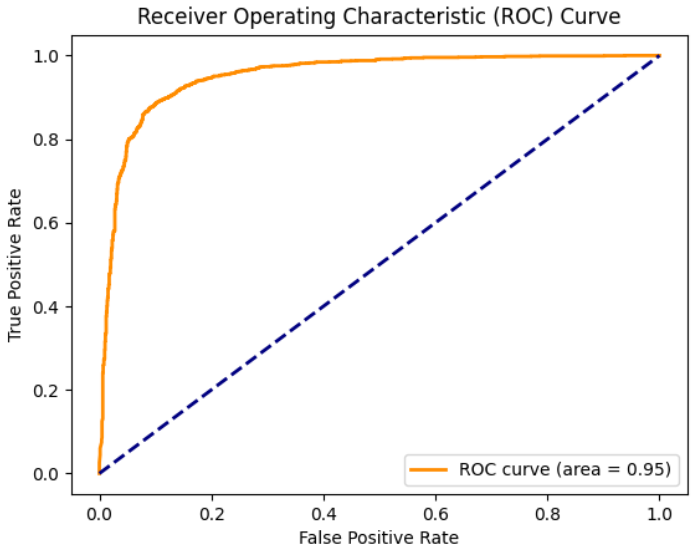


Figure 19: ROC Curve for Xception Model in gastric classification

The Xception model, akin to the ResNet18 model, showcased high training accuracy and consistent test set performance. The subtle variances in the confusion matrix between the two models hint at the Xception model's slight edge in differentiating between Antrum and Corpus. However, the overarching similarity in performance metrics and the shared challenge of overfitting necessitate a meticulous exploration of model refinement strategies and optimization techniques.

Table 1: Model Comparison: ResNet18 vs. Xception

Metric	**ResNet18**	**Xception**
Training Accuracy	0.88 - 0.98	0.87 - 0.99
Validation Accuracy	0.83 - 0.86	0.87 - 0.88
Training Loss	0.3 - 0.05	0.3 - 0.05
Validation Loss	0.35 - 0.65	0.3 - 0.65
Area under ROC	0.95	0.95

This detailed comparison between the ResNet18 and Xception models elucidates their respective strengths and areas necessitating improvement in classifying Antrum, Corpus, and Intermediate regions. The convergence in their performance metrics signifies their efficacy as classifiers, while the subtle distinctions in their outcomes illuminate avenues for further refinement and enhancement. The insights gleaned from this analysis lay the foundation for future work aimed at optimizing model performance and advancing the field of DP.

8.5.2 Performance Analysis: Inflamed/Non-Inflamed Classification

This section provides an in-depth analysis of the ResNet18 and Xception performance in classifying Inflamed and Non-Inflamed regions, which is crucial for accurate medical diagnoses and treatment planning.

ResNet18 Model: The ResNet18 model demonstrated a notable learning progression, with the training accuracy steadily increasing from 0.84 to 0.99, as illustrated in Figure 20. The validation accuracy experienced a slight enhancement, starting from approximately 0.87 and concluding around 0.9, showcasing the model's stable generalization capabilities. The training loss observed a significant reduction from about 0.35 to 0.03, while the validation loss experienced fluctuations between 0.32 and 0.46, ultimately stabilizing at 0.46 in the final epoch, as depicted in Figure 21.

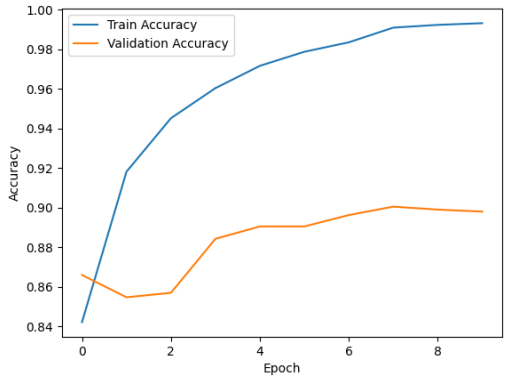


Figure 20: Training and Validation Accuracy for ResNet18 Model in inflammatory classification

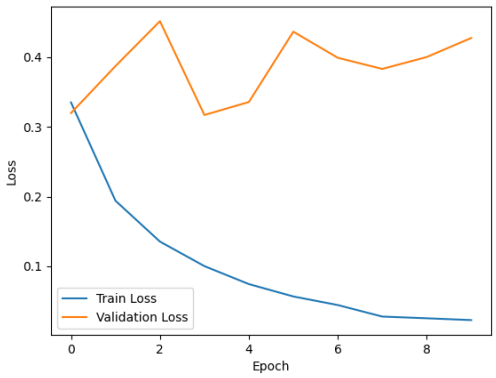


Figure 21: Training and Validation Loss for ResNet18 Model in inflammatory classification

The Confusion Matrix, presented in Figure 22, reveals a strong classification performance, with a particularly high true positive rate for both Inflamed and Non-Inflamed classes. The ROC curve, with an AUC of 0.98 as shown in Figure 23, further attests to the model's discriminative power and its ability to effectively differentiate between the two classes.

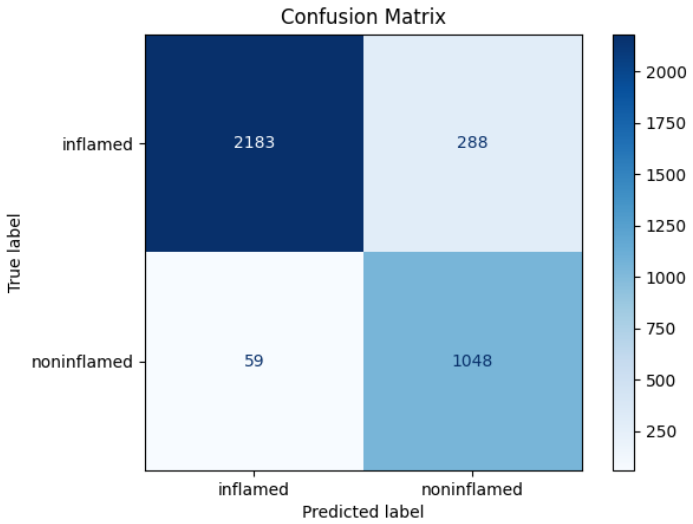


Figure 22: Confusion Matrix for ResNet18 Model in Inflamed/Non-Inflamed Classification

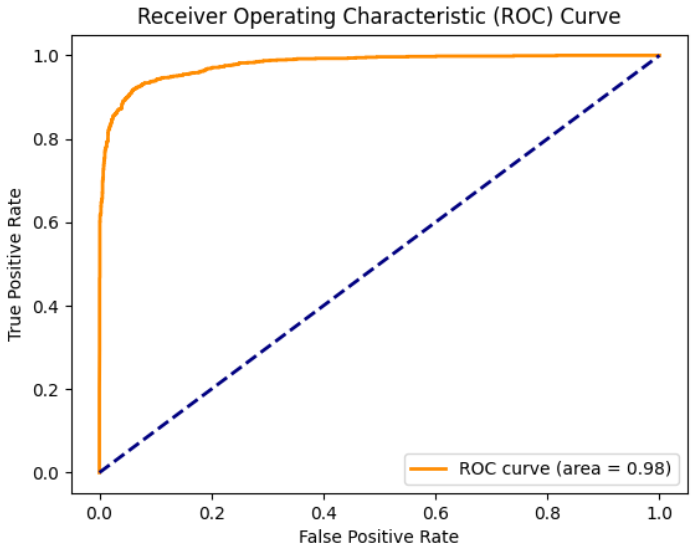


Figure 23: ROC Curve for ResNet18 Model in Inflamed/Non-Inflamed Classification

The model achieved an Accuracy of 0.90, F1 Score of 0.91, Precision of 0.92, and Recall of 0.90, indicating a balanced and robust performance across different aspects of classification.

Xception Model: The Xception model, leveraging its depthwise separable convolutions, exhibited a consistent improvement in training accuracy, starting from 0.84 and culminating at 0.99, as visualized in Figure 24. The validation accuracy

remained stable, fluctuating between 0.88 and 0.9, and concluding at 0.9, showcasing the model's adeptness in generalization. The training loss underwent a substantial decrease from an initial 0.45 to 0.03, while the validation loss experienced an increase to 0.45 in the sixth epoch before reducing to 0.38 in the final epoch, as depicted in Figure 25.

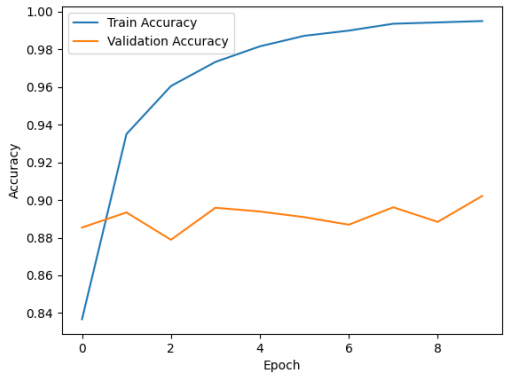


Figure 24: Training and Validation Accuracy for Xception Model in Inflamed/Non-Inflamed Classification

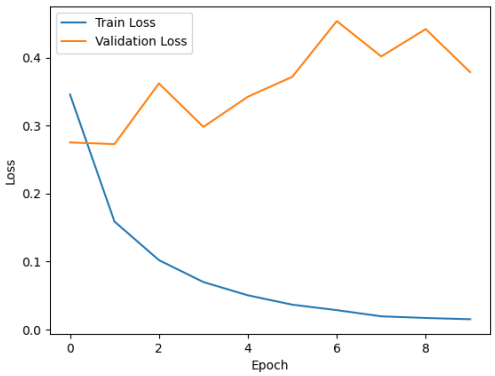


Figure 25: Training and Validation Loss for Xception Model in Inflamed/Non-Inflamed Classification

The Confusion Matrix, illustrated in Figure 26, indicates a high true positive rate for both Inflamed and Non-Inflamed classes, with a slight improvement in classifying Inflamed regions compared to the ResNet18 model. The ROC curve, with an AUC of 0.98, is presented in Figure 27, affirming the model's strong discriminative ability.

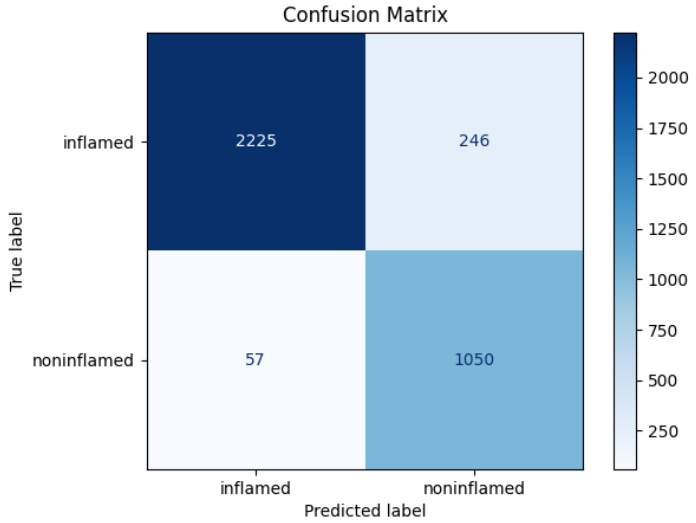


Figure 26: Confusion Matrix for Xception Model in Inflamed/Non-Inflamed Classification

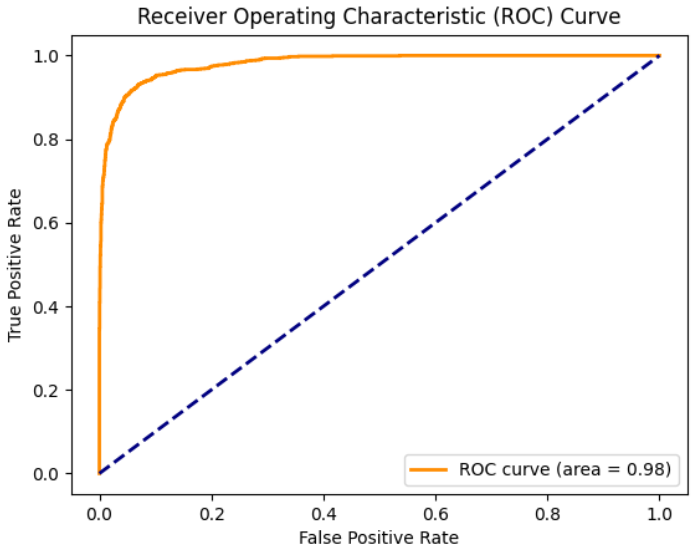


Figure 27: ROC Curve for Xception Model in Inflamed/Non-Inflamed Classification

The Xception model achieved an Accuracy of 0.90, F1 Score of 0.91, Precision of 0.92, and Recall of 0.90, mirroring the performance of the ResNet18 model but with subtle improvements in certain aspects. The nuanced differences in the confusion matrix and the stability in validation metrics suggest that the Xception model may offer slight advantages in classifying Inflamed and Non-Inflamed regions, warranting further exploration and optimization.

Table 2: Model Comparison: ResNet18 vs. Xception in Inflamed/Non-Inflamed Classification

Metric	**ResNet18**	**Xception**
Training Accuracy	0.84 - 0.99	0.84 - 0.99
Validation Accuracy	0.87 - 0.90	0.88 - 0.90
Training Loss	0.03 - 0.35	0.03 - 0.45
Validation Loss	0.32 - 0.46	0.38 - 0.45
Area under ROC	0.98	0.98

This comparative analysis between the ResNet18 and Xception models for Inflamed/Non-Inflamed classification provides valuable insights into their respective capabilities and areas for enhancement. The similarities in performance metrics highlight their effectiveness, while the minor variations offer avenues for further refinement to achieve optimal classification results in DP.

8.5.3 Performance Analysis: Particle-level and WSI-level

Classifying on particle and WSI level uses the same models as for the classification on tile-level and then apply certain thresholds and logic as described in sections 8.3.1

and 8.3.2 to classify on these higher levels. Due to the very similar performance between the ResNet18 and Xception based models, this evaluation will solely focus on ResNet18 models for the tile-level classification. For each level, the focus will first be on the gastric and then the inflammation classification performance.

Particle-level Assessing the gastric classifications at the particle level offers significant insights into the implications of misclassifications at the tile level. Due to the majority voting mechanism, there were no instances of antrum particles being incorrectly classified as corpus, and vice versa. The only observable errors in the test set involved classifying antrum and corpus particles as intermediate, as these did not meet the requisite threshold. These observations are illustrated in Figure 28. The accuracy for antrum particles was 80.72%, with corpus achieving 89.90%, and intermediate displaying the lowest accuracy at 75.00%, attributable to misclassifications arising from unmet thresholds.

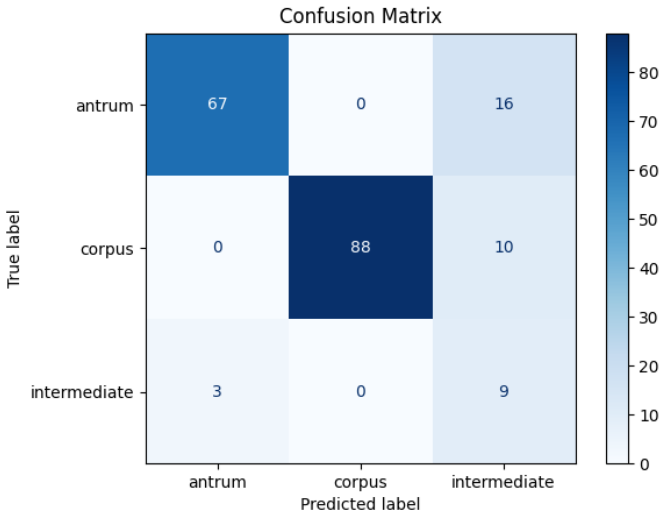


Figure 28: Confusion Matrix for Tile-level Gastric Classification

Evaluating the performance for the inflammatory classification, the results show promising results by accurately classifying all non-inflamed particles as such, while also showing a 91.91% accuracy for classifying inflamed particles. This shows an overall accuracy of 94.61% on the test set. This will also contribute to the performance described for the classification on WSI-level.

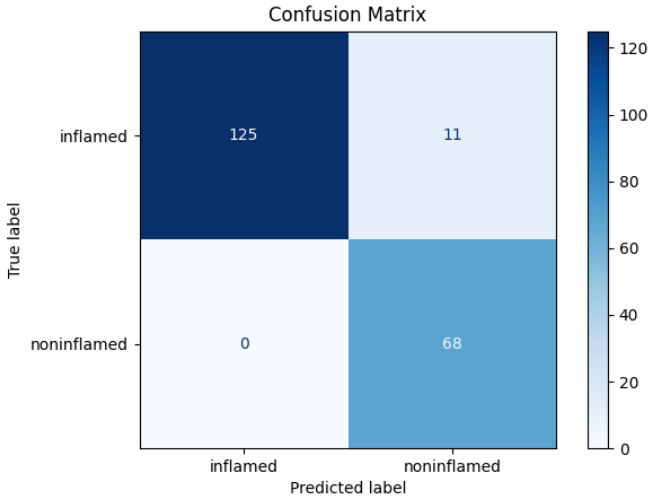


Figure 29: Confusion Matrix for Tile-level Inflammatory Classification

WSI-level Building upon the particle-level performance, the final output that is most relevant for real-world applications is the classification at the WSI-level. This level of analysis is crucial as it aggregates the results of individual particle classifications to generate a comprehensive overview, allowing for more accurate and applicable insights in medical diagnostics.

For the gastric classification, the results exhibit remarkable precision across all categories. The antrum classification yielded a precision of 1.0000, albeit with a recall of 0.8000, resulting in an F1 Score of 0.8889. This suggests that while every classification of antrum by the model was accurate, there was a 20% chance that some antrum classifications were missed, affecting the overall F1 Score.

Similarly, the corpus classification showcased a flawless performance with both precision and recall at 1.0000, leading to a perfect F1 Score of 1.0000. This indicates that the model was able to identify and classify all corpus particles accurately, demonstrating its reliability in detecting this category.

Table 3: Gastric Classification Metrics

Category	Precision	Recall	F1 Score
Antrum	1.0000	0.8000	0.8889
Corpus	1.0000	1.0000	1.0000
Intermediate	0.3000	1.0000	0.4615

In contrast, when we examine the intermediate classification within the test dataset, a noteworthy pattern emerges. The model achieved a perfect recall score of 1.0000, indicating its ability to correctly identify every genuine intermediate case in the test dataset. However, precision for this class is marked at 0.3000, highlighting a discrepancy between true intermediate cases and those falsely classified as such.

A crucial factor contributing to the lower precision in the intermediate classification is the composition of the test dataset. There are only three WSIs containing intermediate particles in the test dataset, a relatively small number compared to the more prevalent antrum and corpus classes. Consequently, any misclassification within these dominant classes has a more pronounced impact on the overall precision metric, leading to a reduced precision score for intermediate classification.

The resulting F1 Score of 0.4615 underscores the need for model refinement. Striking a better balance between precision and recall is essential to reduce false positives and enhance the model's suitability for real-world applications within the context of the test dataset.

It's worth noting that while confusion matrices are valuable tools for many classification tasks, they do not provide meaningful insights in this context of WSI-level classifications. These classifications are essentially enumerations of which classes are present on the WSI, meaning that while some classes may be missing, the presence of other classes can still be accurate. Therefore, a confusion matrix doesn't offer relevant information for this specific analysis.

Switching the focus to inflammatory classification, the model demonstrated exemplary performance at the WSI-level, achieving perfection in all metrics — Accuracy, Precision, Recall, and F1 Score were all measured at 1.0. The confusion matrix further substantiates this success, with 12 inflamed and 8 non-inflamed cases correctly classified, and no instances of misclassification. This implies that the model has outstanding capabilities in differentiating between inflamed and non-inflamed particles, making it a highly reliable tool for inflammatory classification in medical diagnostics.

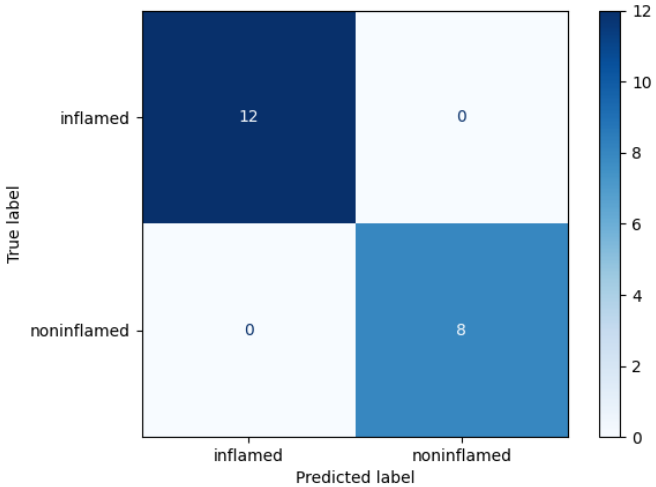


Figure 30: Confusion Matrix for WSI-level Inflammatory Classification

While the results presented above illustrate a highly promising performance of the model in inflammatory classification at the WSI-level, it is imperative to approach

these findings with a degree of caution. The perfection observed in the model's performance metrics needs to be contextualized within the limitations of the limited size of the test dataset as described in section 6.6.

8.5.4 Final Evaluation Reflections

In conclusion, both ResNet18 and Xception models demonstrated proficient and comparable performance in classifying gastric regions and inflammatory conditions, with slight nuances highlighting areas for refinement. The detailed insights from this analysis underscore the potential of these models in DP, yet also emphasize the need for continuous optimization to address challenges such as overfitting and misclassifications. The results at the WSI-level, especially in inflammatory classification, are promising, yet they call for careful interpretation due to the limitations stemming from the dataset size and the selection of explicitly inflammatory and non-inflammatory slides.

9 Discussion

This section delineates the potential avenues for alternative approaches that could have been employed during the different phases of the project. The objective is to critically analyze the chosen methodologies and to identify areas where improvements could be instituted for future research endeavors.

In retrospect, several aspects of the project could have been approached differently to potentially enhance the outcomes and streamline the process.

9.1 Annotation and Scanning Process

The annotation process represents a critical area where improvements could have been instituted. Initially, annotations were documented in an unstructured manner within a Word document, which contained brief descriptions on how to annotate the WSIs. A more structured approach could have involved direct annotations on the WSIs through the guidance of a pathologist, utilizing appropriate annotation software from the onset. This structured approach was not adopted due to indecisiveness regarding the choice of annotation software and a lack of experience with high-resolution images derived from WSIs. Moreover, an attempt to annotate the images prior to consulting with the pathologist consumed a considerable amount of time, a process that, in hindsight, could have been conducted in parallel to save time. Despite this, the preliminary attempt at annotation facilitated a deeper understanding of the classification from a human perspective, offering insights into the potential criteria that could aid in the classification of the images.

Furthermore, the quality of the scans presented a notable area for potential improvement. While the majority of the slides were of satisfactory quality, a subset

exhibited blurriness, detracting from the overall consistency in the dataset. Enhancing the quality of these slides could have been possible with a deeper understanding and expertise with the scanner and its accompanying software. It is worth noting that the initial setup and guidance provided a foundational understanding of the scanning process, which was sufficient for the majority of the slides. However, addressing the issues with the few blurry slides would have required a more advanced knowledge, which was not readily accessible at the time. This represents an area for future improvement, with a focus on ensuring consistent quality across all slides.

9.2 Sample Selection and Model Development

The selection of samples for training the model posed a considerable challenge. Ideally, a more meticulous selection process, encompassing a wide array of scenarios and edge cases, would have been pursued. However, constraints such as time, scheduling conflicts, and geographical barriers rendered this unfeasible within the scope of this bachelor thesis. Moreover, a lack of medical expertise among the developers further limited the influence on the selection of the samples, especially in the initial stages of the project. This was a significant constraint, given that a more informed selection could potentially lead to a more robust training dataset. Despite these challenges, the dataset compiled for both the classification of antrum/corpus/intermediate and for distinguishing between inflamed and non-inflamed samples serves as a good initial starting point. It lays a foundation that can be further refined and expanded upon in future work, leveraging a more detailed understanding and possibly collaborative efforts with medical experts to enhance the dataset. Despite the substantial amount of time invested in the initial planning phase, these factors collectively limited the opportunities for a more refined selection process.

Despite these challenges, the project aimed to establish a baseline model, a goal that was successfully attained. The development phase, initiated after a significant portion of the dataset was scanned and annotated, could have explored various architectures and hyperparameter tuning to enhance the system's performance. However, the primary focus remained on developing a baseline model, laying a foundation for further improvements.

9.3 Practical Implications

Looking forward, it is imperative to learn from the experiences garnered during this project. Adopting a more structured approach to annotation and leveraging expertise in scanner operations could enhance the quality of scans. Moreover, future endeavors should prioritize a well-rounded sample selection process to foster a model capable of handling a diverse range of scenarios effectively.

The development of two scripts that enable the classification of single tiles and entire WSIs marks a significant step forward in the project. These scripts facilitate the classification of images, albeit with certain limitations. Firstly, the prerequisite of

annotating areas of the WSI in QuPath and exporting the tiles for script utilization introduces a manual, time-consuming step before the system can take over. This process stands as a considerable point for future enhancements, with automation being a key focus to streamline operations. Secondly, the limited accuracy, as detailed in section 8.5, necessitates pathologist oversight, albeit potentially offering support in the classification process. Thirdly, the current state of the input and output interfaces lacks user-friendliness. Lastly, legal hurdles present a substantial barrier to real-world implementation, indicating a long pathway to legal clearance for any practical application. Despite these hurdles, the scripts represent a promising start, laying groundwork for future advancements in automated classification and analysis.

10 Further Work

As the project navigates forward, several avenues for further work emerge, aiming to enhance the system's performance and usability. Below are the focal points identified for future endeavors:

- **Dataset Expansion**: A critical step towards improving the system's performance involves expanding the dataset. Incorporating more edge cases, including scans from different WSI scanners, and exploring the use of different stains, as initially contemplated during the project planning, could potentially enhance the robustness of the model. This expansion would foster a more comprehensive understanding and classification capability, accommodating a diverse range of scenarios and variations in the samples.

- **User Experience Enhancement**: Streamlining the input and output interfaces of the system stands as a vital task in transitioning the project into a potentially real-world application. Simplifying the user experience, especially for individuals with limited technical knowledge, is essential. The current necessity to run multiple scripts and manually annotate regions for classification poses a significant hurdle, emphasizing the need for a more user-friendly system that can be seamlessly integrated into the hospital workflow.

- **Model Performance**: Enhancing the model's performance through strategies that prevent overfitting is a pertinent area for further work. By avoiding overfitting, the model can generalize better to unseen data, thereby improving its reliability and accuracy in real-world scenarios.

- **Finer Classification Levels**: Introducing finer levels of classification, such as differentiating between various gastritis subtypes, could augment the automation of the diagnostic process. While determining the inflammation status of a slide is a crucial step, it remains insufficient for a comprehensive diagnosis. Therefore, a more detailed classification schema would be a valuable addition, aiding in a more nuanced understanding and analysis of the slides.

- **Performance Comparison**: A comprehensive evaluation of the project necessitates a comparison of the classification performance between the models developed in this bachelor thesis and those created by Philipp Andreas Höfling, alongside other potential models. This comparison would offer a deeper insight into the project's efficacy, helping to gauge its standing and potential for practical implementation. Moreover, comparing the model's performance with pathologist classifications would provide a benchmark, facilitating an understanding of the model's capabilities in relation to expert human analysis.

These identified avenues for further work underscore the project's potential for growth and refinement. By addressing these areas, it is anticipated that the project can evolve into a more robust and user-friendly system, with enhanced diagnostic capabilities, paving the way for a tool that can significantly support the medical community in the future.

11 Summary

The development of the dataset and annotations, along with the creation of a baseline model, marked the central goal of this thesis. The dataset, comprising digitalized WSIs annotated by Dr. Bettina Braunecker and subsequently exported as tiles, served as the foundation for this research. However, the dataset inherently possesses limitations, including its size and the inability to capture all edge cases. The annotations, while invaluable for distinguishing between inflamed and non-inflamed samples, lack the granularity required for real-world deployment. This limitation is particularly evident as annotations are assigned to entire particles, not accounting for localized inflamed spots within a particle, thereby impacting the model's precision and accuracy.

Furthermore, a specific challenge observed was the misclassification within gastric classifications. Notably, all misclassifications between antrum and corpus were identified as intermediate due to the failure in meeting the necessary threshold. This highlights an area that necessitates further refinement.

Despite these limitations, the baseline models utilizing Resnet18 and Xception have shown promising results at tile-level, particle-level, and WSI-level. Particularly, the classification performance between inflamed and non-inflamed is noteworthy. The classification of gastric regions, though a valuable starting point, requires additional enhancement. A notable achievement was the WSI-level classification of inflamed/non-inflamed, which yielded an F1 Score of 1. However, this impressive score warrants cautious interpretation due to the limited size of the test dataset and the selection of samples with clear classification.

In conclusion, while the results are promising, especially for inflammatory classification, and the models showcase potential, there is a prevalence of overfitting across all models and classifications, necessitating further investigation and refinement in future work.

References

Halo ai. URL `https://indicalab.com/halo-ai/`. Last Accessed: 28.09.2023.

Famke Aeffner, Kristin Wilson, Brad Bolon, Suzanne Kanaly, Charles R. Mahrt, Dan Rudmann, Elaine Charles, and G. David Young. Commentary: Roles for pathologists in a high-throughput image analysis team. *Toxicologic pathology*, 44: 825–834, 8 2016. ISSN 1533-1601. doi: 10.1177/0192623316653492.

Famke Aeffner, Mark D. Zarella, Nathan Buchbinder, Marilyn M. Bui, Matthew R. Goodman, Douglas J. Hartman, Giovanni M. Lujan, Mariam A. Molani, Anil V. Parwani, Kate Lillard, Oliver C. Turner, Venkata N.P. Vemuri, Ana G. Yuil-Valdes, and Douglas Bowman. Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association. *Journal of pathology informatics*, 10, 1 2019. ISSN 2229-5089. doi: 10.4103/JPI.JPI_82_18.

Kaustav Bera, Kurt A. Schalper, David L. Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology 2019 16:11*, 16:703–715, 8 2019. ISSN 1759-4782. doi: 10.1038/s41571-019-0252-y.

Farzad Ghaznavi, Andrew Evans, Anant Madabhushi, and Michael Feldman. Digital imaging in pathology: Whole-slide imaging and beyond. doi: 10.1146/annurev-pathol-011811-120902.

Stefik-M. Choi J. Miller T. Stumpf S. Gunning, D. and G-Z Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4, 12 2019. ISSN 24709476. doi: 10.1126/SCIROBOTICS.AAY7120.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Webster J and Dunstan R. Whole-slide imaging and automated image analysis: Considerations and opportunities in the practice of pathology. *Veterinary Pathology*, 51:211–223, 2014. doi: 10.1177/0300985813503570.

Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 10 2016. ISSN 1361-8423. doi: 10.1016/J.MEDIA.2016.06.037.

David R. Martin, Joshua A. Hanson, Rama R. Gullapalli, Fred A. Schultz, Aisha Sethi, and Douglas P. Clark. A deep learning convolutional neural network can recognize common patterns of injury in gastric pathology. *Archives of pathology laboratory medicine*, 144:370–378, 2020. ISSN 1543-2165. doi: 10.5858/ARPA.2019-0004-OA.

Farahani N. and Pantanowitz L. Overview of telepathology. *Surgical Pathology Clinics*, 8:223–231, 6 2015. ISSN 1875-9181. doi: 10.1016/J.PATH.2015.02.018.

Soojeong Nam, Yosep Chong, Chan Kwon Jung, Tae Yeong Kwak, Ji Youl Lee, Jihwan Park, Mi Jung Rho, and Heounjeong Go. Introduction to digital pathology and computer-aided pathology. *Journal of Pathology and Translational Medicine*, 54:125, 2020. ISSN 23837845. doi: 10.4132/JPTM.2019.12.31.

Muhammad Khalid Khan Niazi, Anil V. Parwani, and Metin N. Gurcan. Digital pathology and artificial intelligence. *The Lancet. Oncology*, 20:e253, 5 2019. ISSN 14745488. doi: 10.1016/S1470-2045(19)30154-8.

Theodor Heinrich Schiebler and Horst-Werner Korf. *Histologie*. Steinkopff, Heidelberg, 2007. ISBN 978-3-7985-1771-4. doi: 10.1007/978-3-7985-1771-4_2.

Rebecca Senft. How to export tiles of large histology images in qupath — carpenter-singh lab, 10 2022. URL `https://carpenter-singh-lab.broadinstitute.org/blog/how-export-tiles-large-histology-images-qupath`. Last Accessed: 06.09.2023.

Pentti Sipponen and Heidi Ingrid Maaroos. Chronic gastritis. *Scandinavian Journal of Gastroenterology*, 50:657–667, 6 2015. ISSN 15027708. doi: 10.3109/00365521.2015.1019918.

Georg Steinbuss, Katharina Kriegsmann, and Mark Kriegsmann. Identification of gastritis subtypes by convolutional neuronal networks on histological images of antrum and corpus biopsies. *International Journal of Molecular Sciences 2020, Vol. 21, Page 6652*, 21:6652, 9 2020. ISSN 1422-0067. doi: 10.3390/IJMS21186652.

Stephen S. Sternberg. *Histology for Pathologists*. Lippincott - Raven, second edition edition, 1997.

C. Thomas. *Histopathologie*. Schattauer GmbH, 13. auflage edition, 2001.

Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of Pathology Informatics*, 9, 1 2018. ISSN 21533539. doi: 10.4103/JPI.JPI_53_18.

Bethany Jill Williams, David Bottoms, and Darren Treanor. Future-proofing pathology: the case for clinical adoption of digital pathology. *Journal of clinical pathology*, 70:1010–1018, 12 2017. ISSN 1472-4146. doi: 10.1136/JCLINPATH-2017-204644.

Xin yu Zhao, Xian Wu, Fang fang Li, Yi Li, Wei hong Huang, Kai Huang, Xiao yu He, Wei Fan, Zhe Wu, Ming liang Chen, Jie Li, Zhong ling Luo, Juan Su, Bin Xie, and Shuang Zhao. The application of deep learning in the risk grading of skin tumors for patients using clinical images. *Journal of Medical Systems*, 43:1–7, 8 2019. ISSN 1573689X. doi: 10.1007/S10916-019-1414-2.

Aeffner F Farahani N Xthona A Absar SF Parwani A Bui M Hartman DJ Zarella MD, Bowman D. A practical guide to whole slide imaging: A white

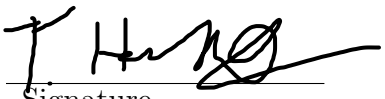
paper from the digital pathology association. *Archives of Pathology Laboratory Medicine*, 143:222–234, 2 2019. ISSN 0003-9985. doi: 10.5858/ARPA.2018-0343-RA.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Bamberg, 02.10.2023

Place, Date



Signature