Development of an AI-based algorithm for the classification of gastric tissue in computational pathology

Master Thesis

Master of Science in Applied Computer Science

Philipp Andreas Höfling

December 14, 2023

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Dr. Bettina Braunecker, Klinikum Nürnberg

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

Abstract

Computational pathology has significantly advanced in recent years, yet a notable gap exists in the specific area of gastric tissue research. Despite its importance in the daily workload of pathologists, the development of algorithms for automated and accurate classification of gastric tissue remains underexplored.

The primary objective of this work is to design, develop, and validate an AI-based algorithm tailored specifically for gastric tissue classification. To achieve this objective, two ResNet18 models are developed: one for the classification of inflammation and another for identifying specific gastric tissue types in whole slide images (WSIs). This process involves data collection, annotation revision, and implementation of a test pipeline for WSIs. Key aspects such as data augmentation and hyperparameter tuning were methodically explored to optimize the models' performance.

The revised dataset used in this research comprised 205 slides, encompassing over 34,000 tiles, small sections extracted from the WSIs. The models are trained on these annotated tiles, and during WSI analysis, the tile results were aggregated using a probabilistic averaging approach to determine the final slide-level classification. The performance of the gastric tissue model on tile level is notable, achieving a area under the curve (AUC) of the receiver operating characteristic (ROC) of 0.95, while the inflammation model achieves a AUC ROC of 0.86. At the tile level, the inflammation model exhibits limited generalizability in identifying type C gastritis, likely due to an uneven distribution of type C gastritis cases in the test and validation datasets. On the test slide level the inflammation classification model demonstrates an accuracy of 94.7%, while the gastric tissue classification model achieves a perfect accuracy of 100%, comparable with the performance of senior level pathologists. These results must be considered with reservations, as the slide test dataset consists of only 19 WSIs.

The findings of this thesis not only showcase the potential of AI in supporting pathologists in gastric tissue classification but also lay the groundwork for future research in this area. Looking ahead, it will be crucial to expand the dataset in terms of size and diversity, and to refine the gastritis type C annotations. These efforts, combined with the exploration of advanced deep learning methods for tile classification and aggregation, will be key to moving this project forward.

In conclusion, this thesis is a step forward in the integration of AI into gastric tissue classification and offers promising implications for future computational pathology research.

Acknowledgements

First, I would like to express my gratitude to Dr. Bettina Braunecker and Dr. med. Volker Mordstein for introducing me to this non-trivial topic and for their support throughout the project. Additionally, I would like to thank my supervisor, Prof. Christian Ledig, for his assistance in overcoming the many challenges encountered in this thesis. I would also like to acknowledge Tom Hempel for his cooperation on this project. Finally, I would like to express my sincere gratitude to my family, friends, and all those who have provided constant support and encouragement throughout my master thesis journey.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ML	Machine Learning
WSI	Whole Slide Image
CPath	Computational Pathology
H&E	Hematoxylin and Eosin
MSE	Mean Squared Error
BCE	Binary Cross Entropy
SGD	Stochastic Gradient Descent
ADAM	Adaptive Moment Estimation
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve
IDE	Integrated Development Environment
ResNet	Residual Neural Network

1 Introduction

Pathology, an important area of medical diagnostics, still relies predominantly on traditional manual methods to examine tissue and classify diseases, while other medical fields such as radiology are increasingly using digital methods. Due to factors such as an aging population, physician shortage, expanded cancer screening programs, and the growing complexity of diagnostic tests, pathologists are expected to experience a continual rise in their workload. (Serag et al., 2019) Computational pathology emerges as a promising answer to these challenges. By integrating artificial intelligence (AI) into pathology, diagnostic processes can become more accurate and time-efficient, transforming the landscape of disease identification and classification.

While most AI research in the field of pathology is concerned with the classification of cancer, most of a pathologist's daily work involves the classification of other diseases. (Rodriguez et al., 2022) Therefore, there is a clear need of AI research addressing broader aspects of pathology beyond cancer. In this context, the primary aim of this thesis is to design, develop, and validate an algorithm tailored for the classification of gastric tissue. This algorithm aims to distinguish between different anatomical regions of the stomach tissue and recognise the presence of inflammation - a task that corresponds to a part of the daily workload of pathologists.

This thesis is part of a project between Klinikum Nuremberg and the chair of Explainable Machine Learning of University Bamberg. The dataset used in this research is derived from Tom Hempel's work. (Hempel, 2023) The focus of the experiments in this thesis is on model optimisation and evaluation. The development of a clinical-grade gastric tissue classification algorithm could be a step towards the modernisation of pathology. It has the potential to decrease the workload of pathologists and enhance diagnostic precision, resulting in better patient outcomes.

2 Background

2.1 Clinical Background

2.1.1 Histopathology

In the field of medicine, pathology plays a crucial role in diagnosing diseases by examining tissues obtained through biopsies or surgical procedures. (Holzinger et al., 2020, S. 4) Staining technique enables pathologists to visualize the cellular and tissue architecture and therefore scan it for possible abnormalities and diseases. One of the most commonly used staining methods in histopathology is Hematoxylin and Eosin (H&E). Hematoxylin stains cell nuclei with blue-black, while Eosin stains other structures with varying degrees of pink (see Figure 1). (Holzinger et al., 2020, S. 283).

The standard histopathology assessment begins with the interpretation of H&E stained glass microscope slides, which are prepared from formalin-fixed tissue samples. These slides are then evaluated by pathologists using a transmission light microscope. (Borowsky et al., 2020, S. 1246)

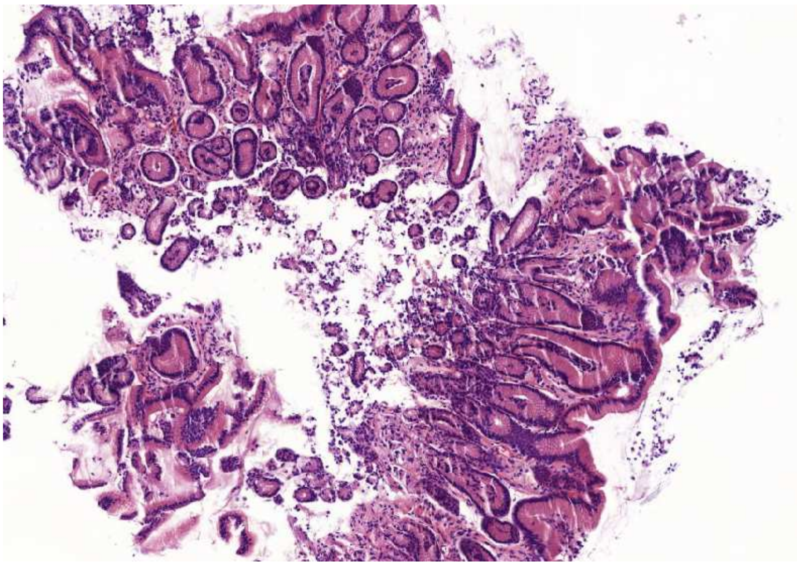


Figure 1: Example of a H&E stained stomach tissue core biopsie

2.1.2 Digital & Computational Pathology

Digital Pathology is a comprehensive domain that encompasses various techniques and systems for converting pathology slides and their associated metadata into a digital format. (Abels et al., 2019, S. 287) While several different definitions for Digital Pathology have been proposed, a common opinion is that Digital Pathology encompasses the photographic documentation of the macroscopy of the specimens ("gross pathology"), the digitization of glass slides (virtual microscopy) and telepathology. (Fraggetta et al., 2021)

Telepathology systems are generally utilized for transmitting digital images of pathology slides. Their application is often reserved for a subset of the laboratory workload where remote diagnosis is critical, such as during intraoperative consultations with frozen sections, which are rapid tissue preparations made for quick diagnosis during surgery, or when seeking second opinions in clinical practice. Also it can be used for seeking second opinions for example from specialists all over the world without the need to send the actual slide, therefore saving time and resources. (Griffin and Treanor, 2017)

Whole Slide Images (WSI) also referred to as virtual slides, are digitized histopathology glass slides that have been created on a slide scanner. The digitized glass slide represents a high-resolution replica of the original glass slide that can then be manipulated through software to mimic microscope review and diagnosis (Abels et al.,

2019) This digitization of slides is performed by a whole slide scanner (for example Figure 2), which produces high resolution replicas of the original slide with estimated file sizes ranging from 0.5 to 4GB for 40× magnification (Abels et al., 2019, S. 289) While historically digital imaging in anatomic pathology relied largely on microscope mounted cameras to produce ”static” digital images modern WSI scanner operation principle consists in moving the glass slide a small distance every time a picture is taken, in order to capture the entire tissue sample. (Pantanowitz et al., 2015, S. 24) A WSI viewer is a software that enables the user to view and zoom into specific parts of a WSI for evaluation on a compatible workstation and connected display. (Pantanowitz et al., 2018)

Figure 2: Whole Slide Scanner Pannoramic MIDI I used in Klinikum Nuremberg from 3DHISTECH

Griffin et al found that digital pathology has already delivered notable advantages in the realms of in remote and frozen section diagnosis, and in second opinion practice. (Griffin and Treanor, 2017) While Borowsky et al. and many others have shown that WSIs are non-inferior to traditional glass slides for primary diagnosis in anatomic pathology, (Borowsky et al., 2020, S. 1245) real-world data indicates that only a minority of pathology laboratories have fully implemented a digital approach to the histological workflow. (Fraggetta et al., 2021)

Computational Pathology (CPath) describes a branch of pathology that involves computational analysis of a broad array of methods to analyze patient specimens for the study of disease. These methods include the extraction of information from digitized pathology images in combination with their associated meta-data, typically using AI methods such as deep learning. (Abels et al., 2019, S. 287)

Most modern applications use image analysis to enhance the precision and consistency of measuring morphological features that pathologists traditionally assess visually or through manually count. (Jahn et al., 2020) CPath includes various applications, including image-enhancing features, measuring and quantification, graphical highlighting/preselection through heat mapping, and fully automated assessments (Jahn et al., 2020)

CPath could lead to more accurate diagnosis than traditional diagnoses and further a faster more automated diagnosis process, which also relieves pathologists. In summary digital pathology is centered around the digitization and management of pathology images, while CPath focuses on using computational methods to analyze and extract valuable information from those digital images for diagnostic, research, and clinical purposes.

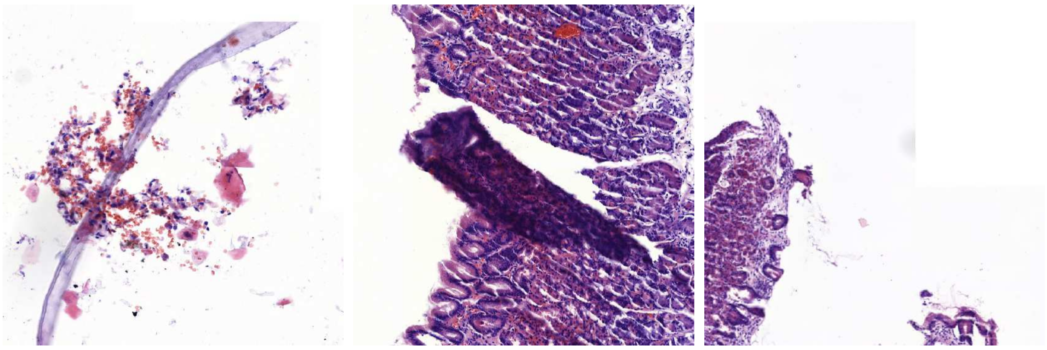
2.1.3 Challenges & Opportunities in CPath

The technological field CPath introduces a set of distinct challenges and opportunities. This subsection explains these challenges, setting the foundation for the rest of the thesis, as these need to be tackled in the goal to create a deep learning system to classify gastric tissue. Additionally, this subsection highlights potential opportunities arising from the application of CPath.

One of the biggest issues in digital pathology is the sheer data size of WSIs. For example glass slides scanned at 20 x magnification produce image files of several gigapixels (Campanella et al., 2019) Coping with this massive data requires for substantial storage capacity, high-speed data transfer solutions, and powerful computational resources.

Another challenge in CPath is the need for transparent, understandable and reliable algorithms, while at the same time achieving high accuracy in a short time. (Holzinger et al., 2020, S. 56) In the field of pathology the topic of transparency is important, especially pathologists and patients need to understand and trust the decision of the algorithm. The extensive mathematical computations executed within a deep Artificial Neural Network to generate an output, which is essentially a decision, do not offer a clear and explainable path to understanding why the network made that decision. In the medical field, where physicians and diagnostic experts often require justifiable explanations for decision-making, this lack of transparency is generally considered unsatisfactory. (Tizhoosh and Pantanowitz, 2018) The adoption of CPath and deep learning approaches in clinical settings faces significant regulatory challenges, primarily due to the 'black-box' nature of these methods, which complicates their interpretability and approval by regulatory agencies. Both in the USA and the EU, such AI-based devices undergo stringent classification and approval pathways, with agencies demanding a clear understanding of how the software operates, emphasizing the need for transparency, reliability, and proven clinical utility before granting them approval for clinical use. (Bera et al., 2019)

Another challenge in CPath projects for clinical use is the cooperation and communication between diverse professional worlds like medicine and data science. While data scientists are responsible for the algorithm development, engineers are in charge



(a) Example of a stain artifact (b) Example of a cut artifact (c) Example of a digitization artifact

Figure 3: Examples of encountered artifacts

of constructing and maintaining necessary physical infrastructure. Clear communication and project lead is needed in such projects to develop efficient CPath systems. (Cui and Zhang, 2021)

Since there are only a small number of publicly available datasets that contain labeled images that can be used for the purpose of building a AI model for WSI classification researchers have to rely on their own datasets. (Tizhoosh and Pantanowitz, 2018, S. 2) The lack of large annotated datasets can be explained by the fact that digital pathology is a new field and the digitization of WSIs is associated with high costs. (Campanella et al., 2019)

In addition, artifacts may appear on the slides due to the preparation process or the digitization process. These issues can emerge from various sources including delays in processing the tissue, variations in preservation durations, irregularities during slicing, staining reagent inconsistencies, or imaging irregularities including uneven illumination, focusing issues, and fluorescence deposits or bleed-through (Dimitriou et al., 2019, S. 2) These artifacts can cause unreliable raw data and produce inaccurate results (Cui and Zhang, 2021, S. 419) For an example of these artifacts see Figure 3).

Furthermore each whole slide scanner captures images using different compression types and sizes, illumination, objectives, and resolution and also outputs the images in a different proprietary file format. (Dimitriou et al., 2019, S. 2) This is a particular problem when a dataset contains WSI from different Whole Slide Scanners. (Madabhushi and Lee, 2016) Another significant issue is colour variation. This variation can arise from different factors such as the type of brand of staining reagents, tissue section thickness, staining conditions and scanner models. Not accounting for colour variation can negatively impact the performance of the machine learning algorithm.(Komura and Ishikawa, 2018) Color discrepancies are notably significant when testing CPath sytems on data from multiple laboratories, and this phenomenon is among the most studied in histopathological image preprocessing. (Kanwal et al., 2022).

While research in the field of CPath is ongoing and applications are still being explored and validated following opportunities may arise from CPath applications. CPath has the potential to be integrated into the development of a clinical decision support system aimed at facilitating precision diagnosis for patients. Campanella et al. (Campanella et al., 2019) showed that CPath systems are able to perform on a clincical gold standard level and further help pathologists to work more time efficient by excluding WSIs with no clinical relevant information.

Classical pathology data like WSIs might be integrated with radiological, genomic, and proteomic data to offer a diagnostic output that is more objective and encompasses multiple dimensions, potentially providing a functionally relevant profile of the disease.(Bera et al., 2019) ML offers the capability to detect novel features in images that could reveal morphological characteristics of clinical significance, which may have been overlooked in traditional pathology due to their prior undiscovery or subtlety beyond the scope of human visual detection.(Abels et al., 2019)

Further CPath systems can improve the performance of the human reader when used in tandem with standard protocols in detection and diagnostic scenarios and additionally help to promote consistency in diagnoses by reducing inter-observer variability often seen among pathologists (Bera et al., 2019) While several challenges in the field of CPath exist, the emerging opportunities especially with the goal of better patient care, outweigh these challenges.

2.1.4 Gastric Tissue

The human stomach is embedded within the abdominal cavity and initiates the process of digestion, serving as both a reservoir and a mixer of food. (Mills, 2012). The stomach is in literature categorized in four anatomical zones: cardia, fundus, body (called corpus in the following work) and the antrum. (Whitehead, 1995) The stomach's internal environment is characterized by its strong acidity. The acid in the stomach has an extremely low pH, close to 1, making it over a million times more acidic than our blood. While this high acidity helps with digestion, it can also damage the stomach lining (Kumar et al., 2018, S. 593) Gastric acid consists of high concentrations of hydrochloric acid, so most ingested microorganisms are killed and the stomach contents are normally sterile. (Mills, 2012, S. 1657)

To protect the stomach lining from the damaging effects of hydrochloric acid, the gastric mucus forms a protective barrier, safeguarding the stomach lining. This barrier not only protects the stomach lining but also provides resistance against harmful substance and facilitates the smooth passage of food through the digestive system(Mills, 2012) The gastric mucus is produced by the gastric glands as well as the surface/foveolar lining cells, constituting a complex glycoprotein characterized by a protein core and branched carbohydrate side chains, predominantly of a neutral nature (Mills, 2012)

Further examination of the stomach's lining uncovers the gastric mucosa, a mucous membrane layer with a consistent pattern throughout the stomach (Whitehead, 1995, S. 15) It consists of a superficial layer containing foveolae (pits), which represent invaginations of the surface epithelium, and a deep layer consisting of coiled glands that empty into the base of the foveolae (see Figure 4). (Mills, 2012)

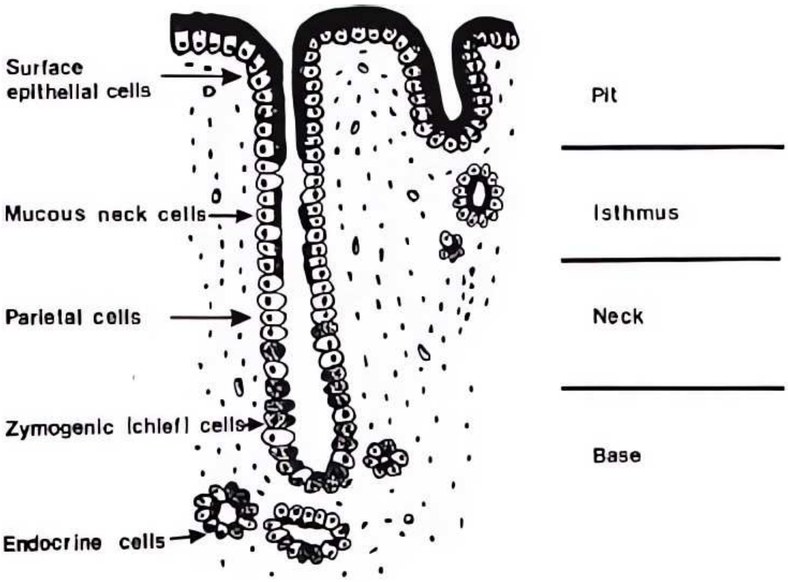


Figure 4: Diagrammatic representation of gastric oxyntic mucosa. (Mills, 2012)

Gastritis is one of the most common diseases affecting the gastric mucosa. In gastritis there is a disturbance in the balance between the local aggressive and defensive factors. (Thomas and Adler, 2001, S. 140) Chronic gastritis, a prevalent and persistent illness, affects a significant portion of the global population, with estimates suggesting that more than half of people worldwide experience this condition to varying degrees. (Sipponen and Maaroos, 2015, S. 657)

Chronic gastritis poses a lifelong risk of aggressive inflammation leading to the gradual destruction of stomach mucosa (atrophic gastritis), which, over time, can result in dysfunction and, in extreme cases, a permanently acid-free stomach. (Sipponen and Maaroos, 2015, S. 657) This severe atrophic gastritis and acid-free stomach are the most significant independent risk factors for gastric cancer. (Sipponen and Maaroos, 2015, S. 657)

There are several types of gastritis, each characterized by distinct causes and manifestations. Type B gastritis, which is the most common, accounts for up to 90% of all gastritis cases worldwide and increases in frequency with age. (Thomas and Adler, 2001, S. 140) Type B gastritis is mainly caused by the infection with the bacterium helicobacter pylori, which starts in the antrum and can spread towards the corpus. (Thomas and Adler, 2001, S. 140) Helicobacter pylori infection is well known to be the most common human infection worldwide on the basis of the fact that approximately 50% of the world's populations are infected and that human beings are the main reservoir. (Salih, 2009) Helicobacter pylori is classified by the World Health Organization (WHO) as a major carcinogen and a leading cause of stomach cancer worldwide. Helicobacter pylori is typically transmitted through direct contact between family members, as well as via contaminated water and food. (Salih, 2009)

Type C gastritis is the second most common subtype of gastritis and is caused by duodenogastric reflux, a condition where the duodenal contents flow back into the stomach, potentially causing irritation and inflammation(Thomas and Adler, 2001, S. 141). Common contributors to this condition include alcohol and the use of non-steroidal anti-inflammatory drugs, which are medications typically taken to alleviate pain or reduce fever, but they can also compromise the stomach's defensive mucosal barrier, making it more vulnerable to irritation from refluxed substances (Steinbuss et al., 2020)

Gastritis symptoms can include epigastric pain, nausea, vomiting, mucosal erosion, ulceration, hemorrhage, hematemesis, melena, and in rare cases, massive blood loss. (Kumar et al., 2018, S. 598) In conclusion, the complex anatomy and physiology of the stomach, coupled with the prevalence of gastritis and its potential complications, underscore the importance of the research goal of this work.

2.2 Machine Learning Background

In recent years, machine learning has emerged as a transformative force across various fields and industries. Machine learning describes the capacity of systems to learn from problem-specific training data to automate the process of analytical model building and solve associated tasks (Janiesch et al., 2021) The objective of this section is to present a concise introduction to the field of machine learning and further concepts used in this work.

2.2.1 Basics of Machine Learning

Machine Learning can be broadly categorized into different types, each with its unique approach and application areas. Supervised learning is a type of machine learning where the algorithm is trained on labeled data. In this setting, the data provided to the algorithm consists of input-output pairs. The objective is for the algorithm to learn a mapping from inputs to outputs. Common applications of supervised learning include classification, where the output is a category, and regression, where the output is a continuous value. (Müller and Guido, 2016)

In contrast, unsupervised learning involves training an algorithm on data that does not have labeled outputs. Here, the goal is for the algorithm to uncover hidden patterns or intrinsic structures within the input data. Unsupervised learning is particularly useful for exploratory data analysis, identifying hidden patterns, or understanding data's underlying structure. (Müller and Guido, 2016)

In reinforcement learning, algorithms interact with an environment rather than processing a static dataset. This interaction forms a feedback loop, where the experiences of the learning system continually inform and refine its learning process.

(Goodfellow et al., 2016, p. 25) While other types of machine learning exist the focus of this work is on supervised learning.

A fundamental aspect of machine learning is the loss function, also referred to as cost function. This function measures the difference between the actual data and the predictions made by the model. The primary goal in training a machine learning model is to reduce this difference, known as minimizing the loss function, which helps to improve the accuracy of the model's predictions.(Bishop, 2016, p. 41)

The learning rate is a important parameter in the gradient descent optimization, a method used to adapt the weights of a network with the goal to minimize the loss function. It determines the size of steps the algorithm takes towards the minimum of the loss function. An appropriately chosen learning rate ensures efficient training of the model, balancing the speed of convergence and the risk of overshooting the global minimum. (Bishop, 2016, p. 240)

Batch size determines how many training samples are used in a single iteration of model training within mini-batch gradient descent, where the training data is divided into random mini-batches for iterative parameter updates. Larger batches provide a more accurate estimate of the gradient, but also require more memory while training. On the other hand, smaller batches can potentially enhance the model's generalization capabilities due to the regularizing effect of the noise inherent in smaller batches, although they might slow down the overall training process (Goodfellow et al., 2016, p. 276)

2.2.2 Neural Network Architectures

Convolutional Neural Networks (CNNs) CNNs have emerged as a potent tool for various machine learning tasks (Krizhevsky et al., 2012) and are a specialized kind of neural network for processing data that has a known grid-like topology. (Goodfellow et al., 2016) CNNs draw inspiration from the way humans process visual information and are highly effective at detecting and learning structured patterns within images. Convolutional networks are neural networks that use convolution in place of general matrix multiplication in at least one of their layers. (Goodfellow et al., 2016) For a comprehensive understanding of CNNs, it is necessary to be familiar with the key concepts and the different layers of their architecture. At the core of CNN functionality is the convolution operation, which is instrumental in feature extraction from the input image. This process can be mathematically represented as:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{1}$$

Where K is the kernel and I is the image. It uses a set of learnable filters (also called kernels), which when convolved with the input image, produce feature maps. Each filter is responsible for extracting a particular feature from an input data. In the context of kernels, the parameter 'stride' refers to the number of pixels by which the kernel is moved or shifted across the input image, while 'padding' refers to the addition of extra pixels (usually zeros) around the edge of the input image to allow the kernel to be applied to the bordering elements of an input image. (Goodfellow et al., 2016, p. 328)

The pooling layer reduces the spatial size of the representation to reduce the amount of parameters and computation in the network. The most common form is max pooling, where the maximum value is taken from a particular window of values, and this window is slid over the feature map. After several convolutional and pooling layers, in typical CNNs the high-level reasoning is done via fully connected layers. (Goodfellow et al., 2016, p. 335 - 338)

Neural Networks Deeper networks with additional layers can explore a larger parameter space, but it has been noted their performance can decline beyond a certain depth. A phenomenon attributed to issues like the vanishing gradient problem. The ResNet, developed by Microsoft Research Asia, addresses this challenge by introducing a residual learning framework. This approach involves residual blocks visualized in Figure 5 (He et al., 2015)

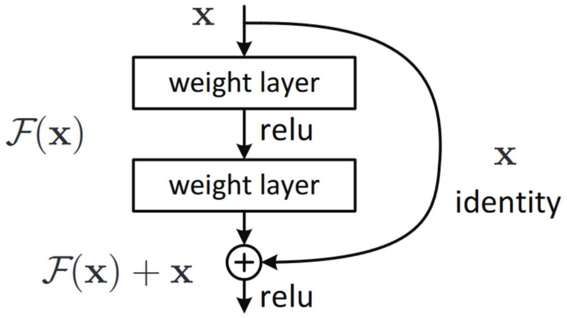

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 5: Residual learning: a building block (He et al., 2015)

A residual block consists of two main paths: a shortcut or identity path and a main path. The identity path simply passes the input tensor (identity) directly to the next layer, while the main path applies a series of convolutional and activation functions to the input tensor. The key innovation of ResNets lies in the addition of the output from the main path to the identity path. This addition operation creates a residual connection, which allows the model to learn the residual or the difference between the desired output and the input during backpropagation (He et al., 2015).

$$y = F(x, \{W_i\}) + x \tag{2}$$

Here x and y are the input and output vectors of the layers considered. The function $F(x, W_i)$ represents the residual mapping to be learned. This formulation enables the gradient to flow directly through the identity path, ensuring that even very deep networks can be trained effectively.. (He et al., 2015)

In the paper He et al. introduced different architectures of ResNets varying in depth and parameter count. In the following the smallest ResNet architectures is presented relevant for this work. ResNet-18, the most lightweight model with 11 million parameters, consists of 18 layers and is known for its efficiency, making it suitable for scenarios where computational resources are limited. (He et al., 2015)

2.2.3 Advanced Techniques in Machine Learning

Functions n the following subsection, mean squared error(MSE) and binary cross-entropy(BCE) loss are briefly introduced, as they will be utilized in later parts of this work.

The MSE is a popular loss function predominantly utilized in regression problems. It is calculated as the average of the squared differences between predicted and true values(Goodfellow et al., 2016, p.177)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

Where: NN is the total number of samples. y_i represents the true value of the i- th sample. $\hat{y}_i$ represents the predicted value of the i-th sample.

The primary advantage of MSE is its ease of computation and its clear focus on minimizing large errors, as these errors are squared and thus have a more substantial impact on the overall score. However, a potential drawback is that it can be sensitive to outliers, since the squaring of the differences magnifies the effect of large errors. (Bermejo and Cabestany, 2001)

BCE Loss, often just called log loss, is commonly used for binary classification problems. It quantifies the difference between two probability distributions: the true distribution and the predicted distribution. For binary classification, the true distribution is usually represented as 0 (negative class) or 1 (positive class), and the predicted distribution is the model's estimated probability of the sample belonging to the positive class.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{4}$$

Where: N is the total number of samples, y_i is the true label (0 or 1) of the i-th sample, $\hat{y}_i$ is the predicted probability of the i-th sample being in class 1. [p. 204](Rothman, 2018)

Optimizers In the field of deep learning, optimizers have a crucial role in fine-tuning the model by adjusting the weights of the neural network to minimize the loss function. Their importance lies in steering the training process towards convergence, ensuring that the model learns effectively from the data. In the following the two relevant optimizers for this work are briefly introduced.

The Stochastic Gradient Descent (SGD) with Momentum is a variant of the basic SGD optimization algorithm used in training neural networks. It incorporates a momentum term, significantly enhancing the convergence speed of standard SGD. This momentum term accumulates a fraction of past gradients and adds it to the current gradient, effectively smoothing out the updates. The inclusion of momentum is particularly useful in scenarios with high curvature, small but consistent gradients, or noisy gradients.[p. 294] (Goodfellow et al., 2016)

The ADAM optimizer, standing for Adaptive Moment Estimation, is a prominent tool in deep learning for optimizing neural networks. Introduced by Kingma and Ba in 2014, it combines the strengths of two other algorithms: AdaGrad and RMSProp. ADAM sets itself apart by calculating adaptive learning rates for each parameter. It does this by keeping track of the first (mean) and second (uncentered variance) moments of the gradients. (Kingma and Ba, 2017) ADAM adapts the learning rate for each parameter individually, based on the estimates of the first (mean) and second (uncentered variance) moments of the gradients.(Kingma and Ba, 2017) ADAM adjusts weights iteratively, enhancing convergence speed with minimal tuning required. This adaptability makes it effective and widely used in various deep learning applications. (Kingma and Ba, 2017)

3 Related Work

In the following section, related work in the field of CPath research is presented. This overview aims to summarize the current state of the art in AI models used in CPath, specifically focusing on gastric tissue classification. As this thesis focuses on implementing a deep learning algorithm for gastric mucosa classification, the subsequent section provides a selective review of successful approaches in this field, addressing the challenges previously outlined in digital pathology.

According to Niazi et al., the large file size of digital biopsies and the computational demands and complexities of deep learning, research has mostly focused on physically smaller tasks which look at small portions of the image like mitosis detection, anatomical region identification , and cancer identification. (Niazi et al., 2019) In their literature review, Rodriguez et al. also demonstrated that the majority of AI research in digital pathology has been focused on cancer detection, as opposed to other medical fields in pathology, which deals with a wider range of diseases including gastritis, appendicitis, and cholecystitis. (Rodriguez et al., 2022, S. 3)

Komura and Ishikawa classified the usecases of CPath into three categories: Computer-assisted Diagnosis, Content Based Image Retrieval, Discovering New Clinicopathological Relationships. (Komura and Ishikawa, 2018, S. 35) Due to the goal of this work, the author focus in following section on algorithms, papers and studies with the goal of computer-assisted diagnosis.

3.1 Clinical Problem

A fair amount of research has been conducted in the location and classification of Helicobacter pylori. Most of this research uses hystological images from gastric tissue , e.g. Klein et al., who built a deep-learning based decision support algorithm that can detect Helicobacter Pylori in gastric biopsies using a VGG style architecture and a tiling approach. (Klein et al., 2020) Their decision support algorithm demonstrated increased sensitivity, especially in giemsa stained sections, achieving a 100% sensitivity rate compared to the 68.4% sensitivity from traditional microscopic diagnosis. (Klein et al., 2020) This research indicates that such an algorithm can serve as a sensitive screening tool to aid pathologists in accurately diagnosing H. pylori presence in gastric biopsies. (Klein et al., 2020)

On the other hand helicobacter pylori infections and gastritis overall can be diagnosed based on gastrointestinal endoscopy. Here for example Itoh et al. trained a CNN based on endoscopy images to classify Helicobacter pylori infections. (Itoh et al., 2018) and Guimaraes et al. used a CNN to diagnose atrophic gastritis on endoscopy images. (Guimarães et al., 2020, S. 4) Futher Papers like (Li et al., 2020) predicted chronic gastritis using X-ray images.

Regarding the classification of gastric mucosal biopsies into different gastritis subytypes only few studies have been found. Franklin et al. utilzed Halo AI, a CNN based deep learning tool designed for image analysis in pathology, to classify 187 slides of type C and type B gastritis, with comparable results to that of specialized gastrointestinal pathologist. (Franklin et al., 2021) Steinbuss et al. employed a Xception, a light weight adaptation from the Inception architecture, to categorize gastritis into its three common subtypes: autoimmune, bacterial, and chemical. Utilizing a dataset of 1230 tile images from 135 antrum or corpus slides, the study achieved an 84% overall accuracy in subtype classification, particularly excelling in detecting type B gastritis with a sensitivity of 100% (Steinbuss et al., 2020)

In the study titled "A Deep Learning Convolutional Neural Network Can Recognize Common Patterns of Injury in Gastric Pathology," researchers used a CNN similar to VGG to identify over 400 slides from common gastric pathologies, including normal tissues, Helicobacter pylori infection, and reactive gastropathy.(Martin et al., 2020). The CNN demonstrated near-perfect agreement with the gold standard diagnoses, achieving an area under the curve (AUC) of up to 100% for H. pylori (Martin et al., 2020). The findings of (Martin et al., 2020) and (Steinbuss et al., 2020) suggest that the developed algorithms, utilizing a tiling method to manage the large size of WSIs and employing CNNs, can effectively differentiate common inflammatory patterns of gastric injury. While (Martin et al., 2020) showed that it may also be possible to distinguish between normal and inflamed tissue.

Although the author identified studies on tissue segmentation and classification in other tissue regions, such as in cardiac biopsies (Nirschl et al., 2017), no known research has been found on the classification of gastric tissue based on its anatomical zone.

3.2 Technical Problem

In addition to the pathological challenges associated with gastric tissue classification, the development of an AI-based algorithm for CPath involves addressing various technical problems, which are based on the unique challenges for digital pathology (see section 2.1.3). Although only a limited number of studies share the same objective as this work, a fair amount of research has been overall conducted in the field of WSI classification and segmentation. In the following abstract the author will present methods, papers and trends to deal with the previous stated challenges in WSI classification.

Komura and Ishikawa described the typically steps for histopathological image analysis as following:(see Figure 6) before applying machine learning algorithms preprocessing in form of dividing the WSI in local mini batches is necessary. In the following step feature extraction and classification of the local minibatches often with the help of deep learning is performed. After Preprocessing various machine learning techniques can be applied depending on the goal of the final application. (Komura and Ishikawa, 2018)

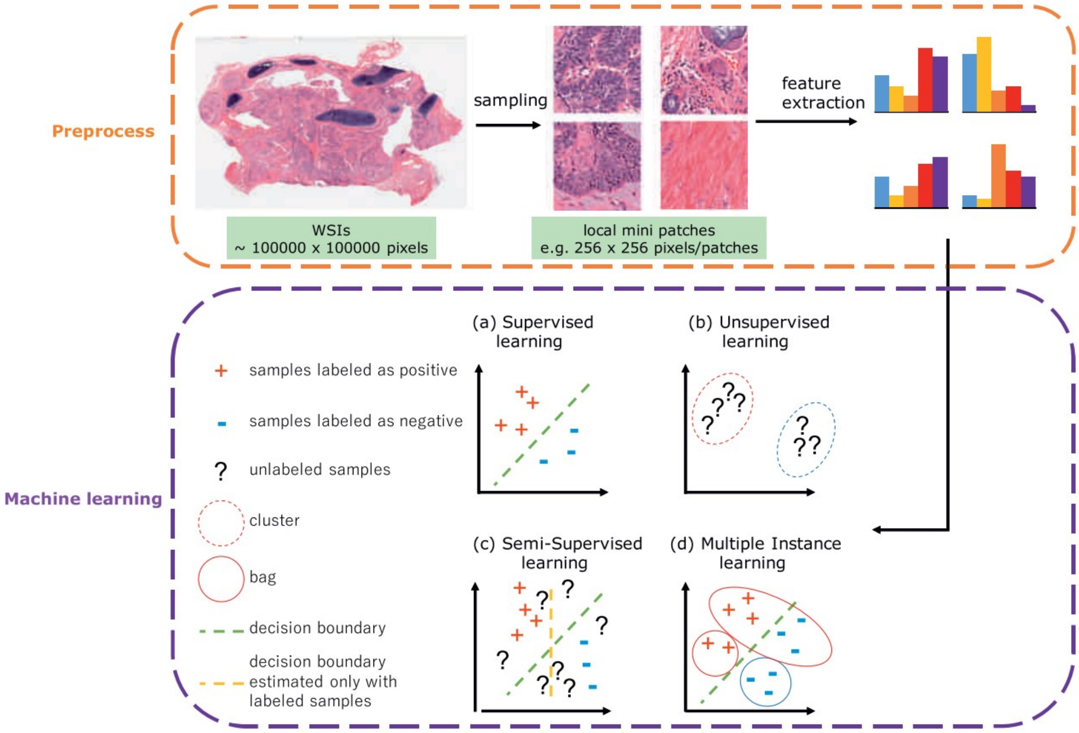

Preprocess
sampling
feature extraction
WSIs
~ 100000 x 100000 pixels
local mini patches
e.g. 256 x 256 pixels/patches
Machine learning
+ samples labeled as positive
- samples labeled as negative
? unlabeled samples
cluster
bag
decision boundary
decision boundary estimated only with labeled samples
(a) Supervised learning
(b) Unsupervised learning
(c) Semi-Supervised learning
(d) Multiple Instance learning

Figure 6: Typical steps for machine learning in digital pathological image analysis (Komura and Ishikawa, 2018, S. 35)

In the paper by Hou et al. (Hou et al., 2016), the challenges of using CNNs directly for whole-slide image (WSI) classification are addressed. One issue is the need for extensive downsampling of the WSI to fit within the memory constraints of the CNN, which can result in the loss of important discriminative details. Another challenge

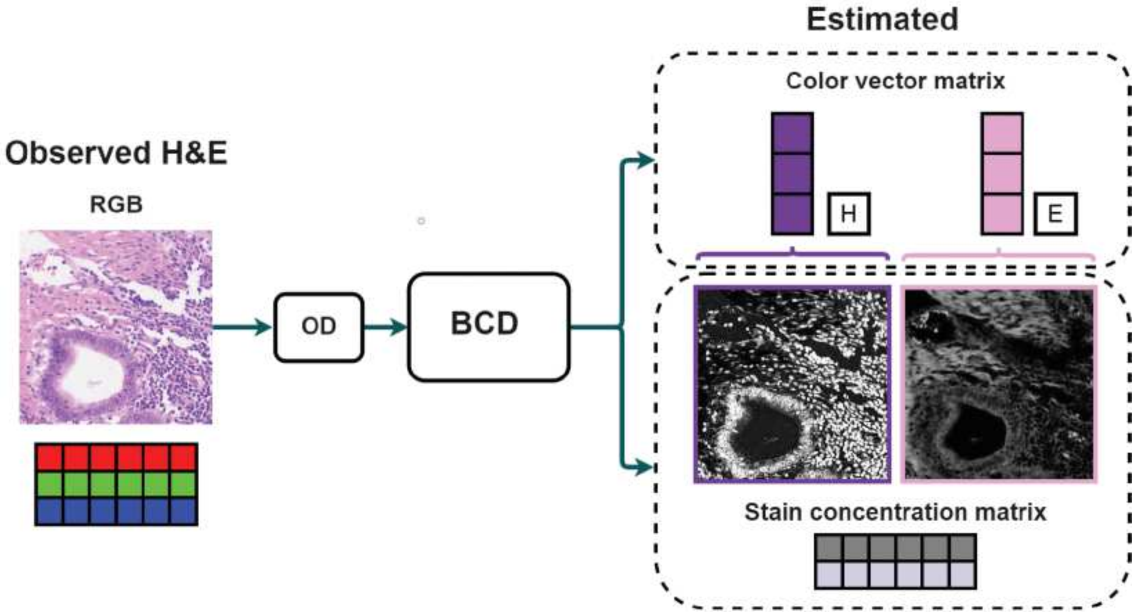


Figure 7: (Kanwal et al., 2022, S. 35)

is that a CNN may only learn from a single discriminative pattern within an image, leading to inefficient use of the available data.. (Hou et al., 2016) Most successful approaches to training deep learning models on WSIs do not use the whole image as input and instead extract and use only a small number of patches (Dimitriou et al., 2019, S. 2). This method of dividing a WSI in smaller patches is called tiling. All of the papers and studies presented here that use high-resolution WSI have used a tiling approach.

To address the issue of colour variation, authors used gray scale, colour normalization and color augmentation. (Komura and Ishikawa, 2018) While gray scale is an easy to implement method, the following CNN loses possible useful color information. Colour normalization aims to modify the colour values of an image pixel-by-pixel, ensuring the colour distribution of the source image aligns with that of a reference image.(Komura and Ishikawa, 2018) In their literature review Kanwal et. al had a more in depth analysis of techniques dealing with color and stain normalization. For example blind color deconvolution is a technique with the goal to seperate the observed multi-stained image into single-stain images.(Kanwal et al., 2022) The process involves estimating the color and concentration of stains in the image for each pixel, by converting the image from RGB into a optical density space and further using this density space to determine the color vectors of various strains (see Figure 7)(Ruifrok et al., 2003). Instead of directly using the RGB channels, the separated quantity of dye absorbed by the tissue can be input into CPath systems(Kanwal et al., 2022) A newer approach of dealing with color variation is the use of generative adversarial networks to generate a new color normalized copy of the source image (Zanjani et al., 2018)

A different strategy of dealing with color variation in WSIs is color augmentation. Color augmentation is a kind of data augmentation performed by applying random hue, saturation, brightness, and contrast.(Komura and Ishikawa, 2018) As stated in (Ai et al., 2021) machine learning requires large datasets and augmentation is a central data expansion approach. (Ai et al., 2021)

Often found data augmentation methods in the field of WSI classification other then the already mentioned include (Kanwal et al., 2022):

- 90-degree rotations: Rotating images in intervals of 90 degrees to introduce different orientations.
- Vertical and horizontal mirroring: Flipping images vertically or horizontally to diversify orientations.
- Gaussian noise: Introducing random pixel variations to mimic real-world imaging noise.
- Gaussian blurring: Smoothing images to simulate out-of-focus scenarios or instrument quality variations."

In recent years, traditional image recognition methods in the field of CPath have largely been replaced by deep learning approaches (Komura and Ishikawa, 2018) Goncalves et al. has shown in their literature review that a large percentage of the proposed applications in literature use convolutional neural networks (CNNs) for the diagnosis of gastric cancer or the detection of lesions in gastric tissue. (e. Gonçalves et al., 2020)

In the study titled "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning," the authors address the limitation of Vision Transformers being traditionally studied for low-resolution images. (Chen et al., 2022) They introduce the Hierarchical Image Pyramid Transformer (HIPT) designed to process WSIs in CPath, exploiting the inherent hierarchical structure of these images through two levels of self-supervised learning. (Chen et al., 2022) Benchmarked on multiple slide-level tasks, there introduced architecture demonstrates superior performance in cancer subtyping and survival prediction compared to existing methods, underscoring the potential of self-supervised Vision Transformers in this field. (Chen et al., 2022)

Real-world medical applications often face a shortage of well-annotated training data, impacting model training and reconstruction. Many studies utilize methods like transfer learning and fine-tuning to compensate for this data insufficiency. (e. Gonçalves et al., 2020) Most of the found studies utilize transfer learning on models trained on the Imagenet dataset. ImageNet is a large-scale image database built on the WordNet structure, aiming to provide hundreds to thousands of annotated images for each of the 80,000 synsets in WordNet, facilitating advanced image indexing, retrieval, and computer vision research.(Deng et al., 2009)

Studies such as (Kang et al., 2023) assume that histopathology is unique in that the orientation of the image is inconsequential. This contrasts with typical image classification tasks where orientation and colour variations in real-world examples can greatly affect accuracy(Kang et al., 2023). As a result, Kang et al. showed in some examples that models pretrained on large histopathology datasets outperform those pretrained on ImageNet

While the introduced deep learning methods mainly are used at tile level, the next step in a typically CPath pipeline involves aggregating these individual tile-based results to a comprehensive slide-level classification. This transition from localized tile-based inferences to a broader slide-level diagnosis is vital for the accurate interpretation of WSI data.

While most fully supervised approaches use pixelwise or tilewise annotation of the data Campanella et al. built a multiple instance learning-based deep learning system which requires only a slide level annotation (reported diagnosis).(Campanella et al., 2019) Pixelwise or tilewise annotation is a labor- and time- intensive task, while slide level annotations safe tedious annotation work for pathologists (Campanella et al., 2019). This allowed the authors of (Campanella et al., 2019) to gather a respectively large dataset of 44,732 WSI. In their work (Campanella et al., 2019) showed that with a large dataset a weakly supervised deep learning approach can provide clinical-grade diagnosis.

Campanella et al. used a multiple instance learning approach where each WSI can be considered a bag consisting of a multitude of instances (in this case tiles of size 224 × 224 pixels). The diagnosis for each WSI is used as label in this. (Campanella et al., 2019) Given this tiling strategy the authors produce bags $B = \{B_{si} : i = 1, 2, \ldots, n\}$, where $B_{si} = \{b_{i,1}, b_{i,2}, \ldots, b_{i,mi}\}$, is the bag for slide s_i containing m_i total tiles. Campanella et al. (2019)

The most basic and common approach in aggregating the instance labels to a bag label in Multiple Instance Learning is the max-pooling approach (Wang et al., 2018).

$$y = \max_i y_i \tag{5}$$

In case of WSI classification, each tile is passed through a model or classifier providing a score or probability how positive or negative the tile is. Next the tile with the highest maximum score is taken as the representative score of the WSI. If the score surpasses a certain threshold the WSI is classified as positive.

Iizuka et al. compared max pooling, where a WSI is assigned the label with the maximum probability from all of its tiles to a RNN model, which was trained to combine the information from all of the tiles using deep CNN features as input. (Iizuka et al., 2020) They found that the RNN model has a overall better AUC than the max-pooling approach, which is more prone to errors as a single WSI can contain up to thousands of tiles, and all it would take is one false positive tile to result in an incorrect classification. (Iizuka et al., 2020) Similar findings were also made by Campenella et al., who compared max pooling to a random forest model on manually engineered features extracted from the heat map generated by their MIL-based tile classifier. (Campanella et al., 2019) Futher Li et al. introduced a MIL aggregatior that models the relations of the instances in a dual-stream architecture with trainable distance measurement. (Li et al., 2021, S. 14318)

While most presented papers and studies use tiles from a single magnification, Vanrijthoven et al.'s HookNet is a novel semantic segmentation model designed for histopathology whole-slide images that integrates both context and detailed information through multiple branches of encoder-decoder convolutional neural networks, using concentric patches at various resolutions with different fields of view.(van Rijthoven et al., 2021)

Goncalves et. al found that in the gastric CPath field accuracy is a commonly employed metric. Accuracy alone is insufficient for comprehensive model evaluation due to potential biases in training or test data leading to non-generalizable models. To address this limitation, a variety of additional metrics are employed alongside accuracy. These metrics include recall, precision, sensitivity, F1 score, and assessments using the receiver operating characteristic (ROC) and area under the curve (AUC) analyses. (e. Gonçalves et al., 2020)

4 Project Overview

This project aims to create a dataset of gastric biopsies and develop an AI-based algorithm for classifying gastric tissue, in collaboration between the Chair of Explainable AI at the University of Bamberg and the Department of Pathology at the Paracelsus Medical Private University Nuremberg. Project participants included developers and students Tom Hempel and Philipp Höfing and two senior level pathologists from Klinikum Nuremberg.

At the start of the project, the developers were given a guided tour of the pathology department's workflow, which provided valuable insights into slide preparation. The developers' understanding of the pathology workflow, particularly with regard to the processing of gastric tissue, played an important role in the development of the AI algorithm. A summarised version of the typical gastric biopsy workflow is presented below. When the gastric needle biopsy arrives at the pathology laboratory, the tissue sample is carefully processed and embedded in paraffin. Thin sections of the sample are then cut and placed on glass slides, followed by the application of various stains to highlight different cell types and structures. Under a microscope, a pathologist carefully examines the stained tissue sections, looking for abnormalities in cell size, shape and arrangement. They look for signs of inflammation, infection or malignancy.

Since core needle biopsies do not always extract tissue from the intended anatomical zones, pathologists are required to classify the tissue according to its original anatomical location. This additional classification is essential for further diagnosing various abnormalities or diseases, as some earlystage gastritis subtypes can only observed in specific anotomic regions of the stomach. For example gastritis type c can be found most commonly in antrum. (Thomas and Adler, 2001).

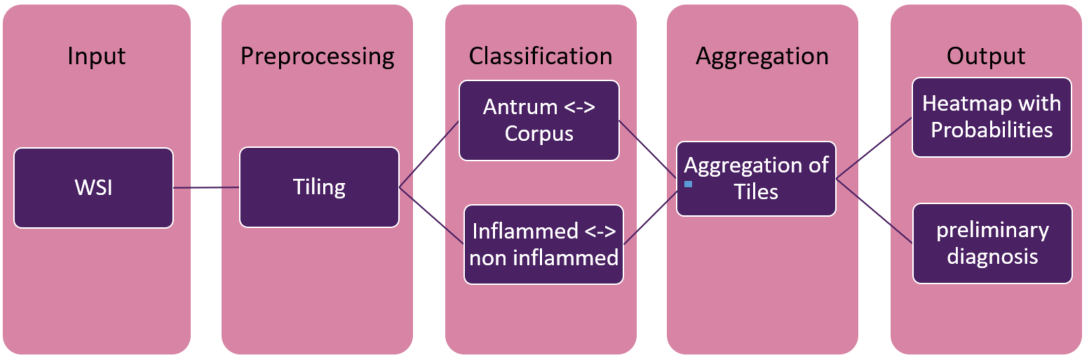


Figure 8: Visualisation of the planned algorithm at the beginning of the project

Based on the microscopic findings, the pathologist formulates a detailed report, providing insights into the condition of the gastric tissue. This report may include a diagnosis or a suggestion for further testing if needed, ultimately guiding the treating physician in the patient's subsequent care and treatment planning.

In subsequent planning meetings, led by the senior level pathologists, the project's scope was formulated and the key topics of antrum, corpus, and gastritis, were explained to the project participants.

As the project progressed, detailed discussions were held to develop a system plan defining the exact nature of inputs and outputs (see Figure 8). The developers were trained in the classification of gastric regions, using medical textbooks and practical experience with microscopes. Training in gastric region classification and understanding the typically classification challenges of pathologists were valuable insights for the developers in their task of creating an effective AI model.

5 Data

This section provides an overview of the dataset used in the study aimed at developing an AI-based algorithm for gastric tissue classification in computational pathology. It is crucial that the data is representative of the real-world scenario, including different types of gastric tissue, disease stages and potential artifacts that may be present in pathology slides. The following subsections present the specifics of the dataset, including its source, composition, pre-processing steps and annotation process. The dataset in this paper was created in cooperation with the Klinikum Nuremberg. The author attempted to locate other datasets used in research and to request their use, but none of the dataset owners responded to the request. As the collection of the dataset was the focus of Tom Hempel's project and the topic of his bachelor thesis 'Development of a dataset and an AI-based proof-of-concept algorithm', this work uses Tom Hempels dataset, with minor modifications (Hempel, 2023).

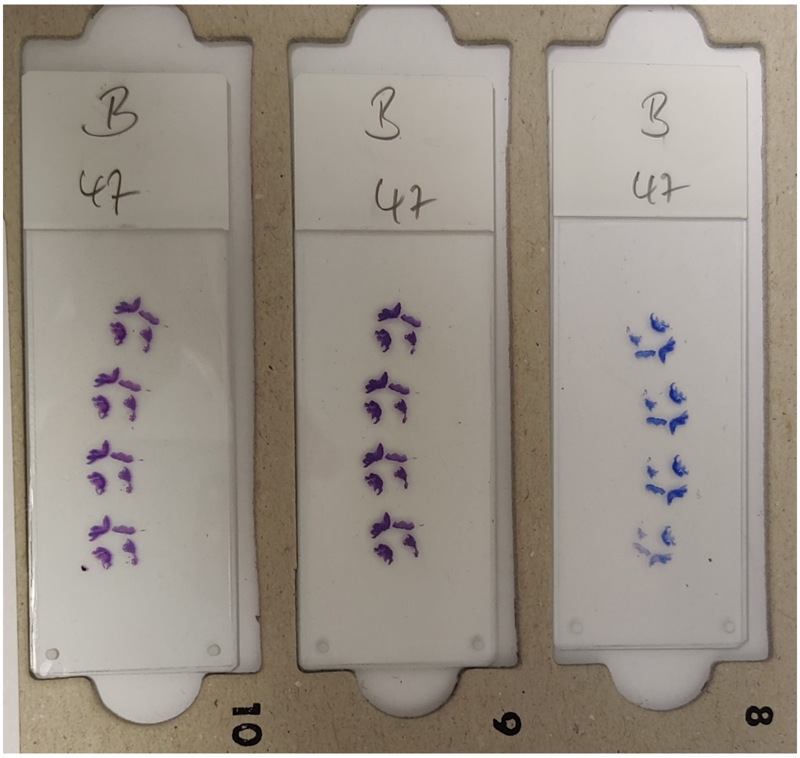


Figure 9: anonymized 47th Type B Gastritis case stained in order H& E, PAS, modified Giemsa (Hempel, 2023)

5.1 Data Collection

The project participants were handed selected cases of stomach biopsies from the archives of Klinkum Nuremberg. The cases where selected by two experienced pathologists of the Klinikum Nuremberg. The cases were stained using H&E, PAS (Periodic Acid-Schiff), and modified Giemsa techniques. Slides that were not scanned in H&E are not discussed in this paper, but are useful in the clinical work of pathologists in detecting stomach biopsy anomalies.

To ensure the protection of patient privacy and maintain confidentiality, the cases underwent anonymisation prior to digitisation. As a result, all identifiable information was removed, satisfying the mandatory requirements for data transfer beyond the hospital's network, and allowing the cases to be stored on the universities file server. The anonymization process involved removing all personal identifiers, labeling slides with incrementing numbers, and appending specific letters for different types of gastritis or non-classifiable cases, ensuring complete anonymity and compliance with privacy standards. For an example of the anonymized slides see Figure 9.

The initial strategy for the project was to start with a dataset consisting of roughly 100 samples each for the antrum, corpus, and intermediate region and later expand the scope to include inflammatory samples. The WSIs were scanned at 20x magnification by the project participants on a Pannoramic MIDI I slide scanner by the

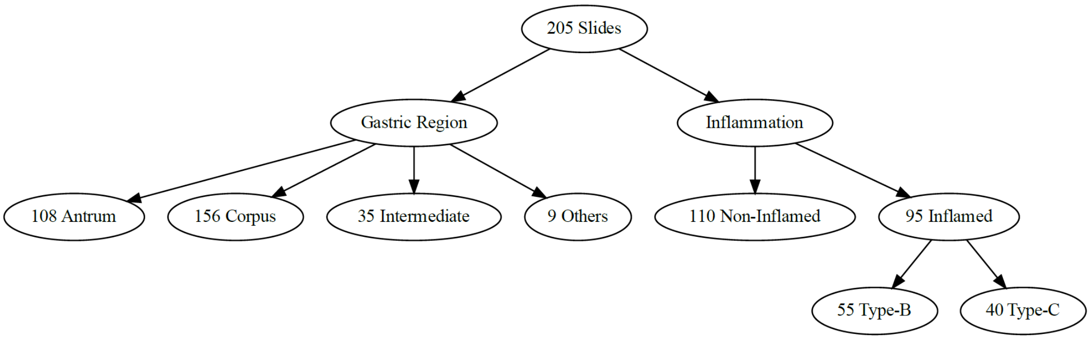


Figure 10: Updated class distribution of WSIs, adapted from (Hempel, 2023)

company 3DHistotech Ltd. The slide scanner is located at the southern location of 'Klinikum Nuremberg' (see Figure 2. The slide scanner saved the data as MIRAX (.mxrs) data, a proprietary file format used by digital slide scanners produced by 3DHistotech Ltd.

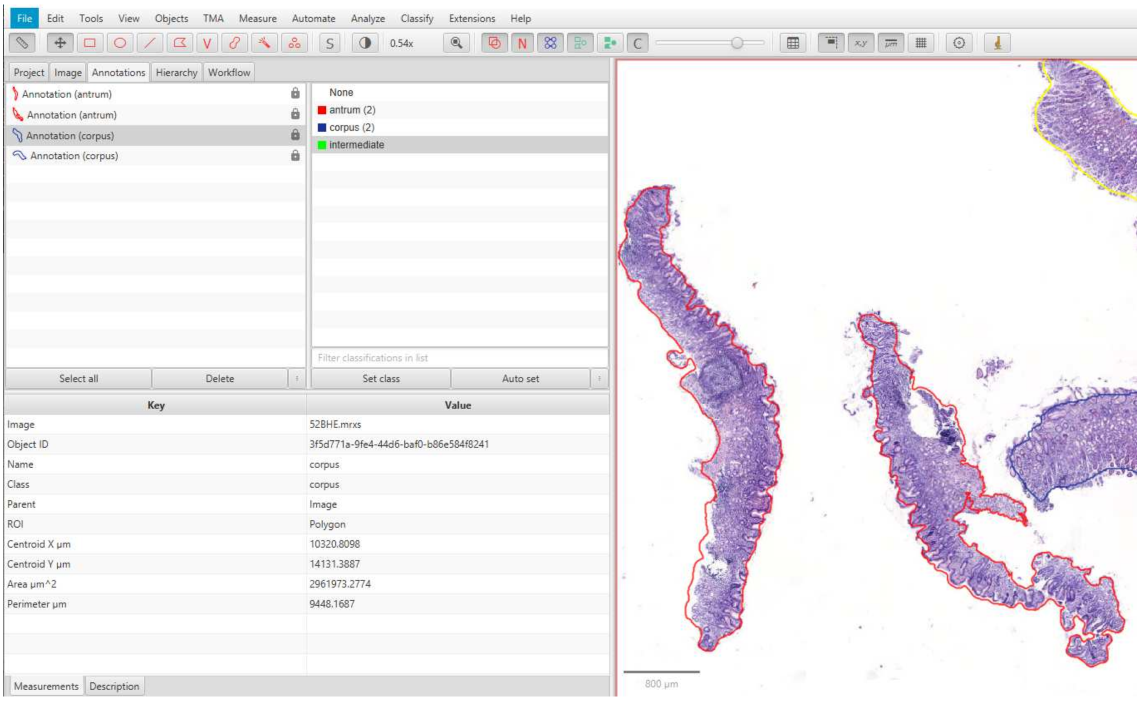
Tom Hempel's project consists of a digitalized collection of 205 H&E WSIS. In total, there are 156 samples containing corpus particles, 108 samples with antrum, 35 with intermediate and 9 samples that could not be classified as antrum/corpus/intermediate. For the inflammatory classification, there are 97 non-inflamed and 108 inflamed samples (55 Type-B, 40 Type-C). 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. An update to Tom Hempel's dataset involves reclassifying the 'Others' category as 'non-inflamed,' based on feedback received from pathologists.

The author planned to create a second small test set of 15 slides. However, due to technical issues at the Klinikum Nuremberg, the slide scanner was unavailable for the final two months of the project.

5.2 Data Annotation

Data annotation, in the realm of deep learning, refers to the process of labeling or tagging data in a way that the algorithm can understand and learn from. The quality and accuracy of data annotation directly influence the performance of the resulting machine learning model, making it a vital component of the algorithm development process.

The annotations of the cases were conducted by Tom Hempel and Philipp Höfling after the presented annotation protocol in (Hempel, 2023) and later reviewed by two senior level pathologists. The annotation protocol created by (Hempel, 2023) can be found in the appendix A.2. Care was taken to exclude artifacts and non-classifiable tissue during annotation, ensuring these were not present in the tile exports.



Key	Value
Image	52BHE.mrxs
Object ID	3f5d771a-9fe4-44d6-baf0-b86e584f8241
Name	corpus
Class	corpus
Parent	Image
ROI	Polygon
Centroid X µm	10320.8098
Centroid Y µm	14131.3887
Area µm^2	2961973.2774
Perimeter µm	9448.1687

Figure 11: Screenshot of annotation process

The annotations were transferred to QuPath, an open-source bioimage analysis software designed for digital pathology and WSI analysis, providing tools for tumor identification, biomarker evaluation, batch-processing, and script functionalities, alongside a flexible platform to develop and share new algorithms across various biomedical research domains. (Bankhead et al., 2017) Here the corresponding tissue parts were marked and annotated by the project participants with the goal to later export the WSI in a format suitable for deep learning architectures to learn from.

Further the author introduced a custom export script for tiles in Qupath (see listing A.3). This script enables users to adjust not only the output pixel size of the tiles but also their resolution and overlap. Additionally, it automatically appends the class name, assigned earlier in QuPath, to the filename of each exported tile, thereby simplifying the process of tile identification and organization for later use. The slide level annotations later used to evaluate the models on WSI level, were created using a python script, which iterates over all tiles and aggregates their label. The results were stored in a CSV file as python dictionary, these files are also available with the data.

It is a common occurrence for errors to arise during the data annotation process, as it is a demanding task that requires great attention to detail. In the later part of the project the project participants took the initiative to review the annotations, focusing particularly on the instances of the baselines tile errors per slide. Slides with a high ratio of tile errors, have been reviewed manually by the author. With assistance from an experienced pathologist, 10 slides in the dataset were found to have been either misclassified by the pathologists or incorrectly classified in Qupath. Fur-

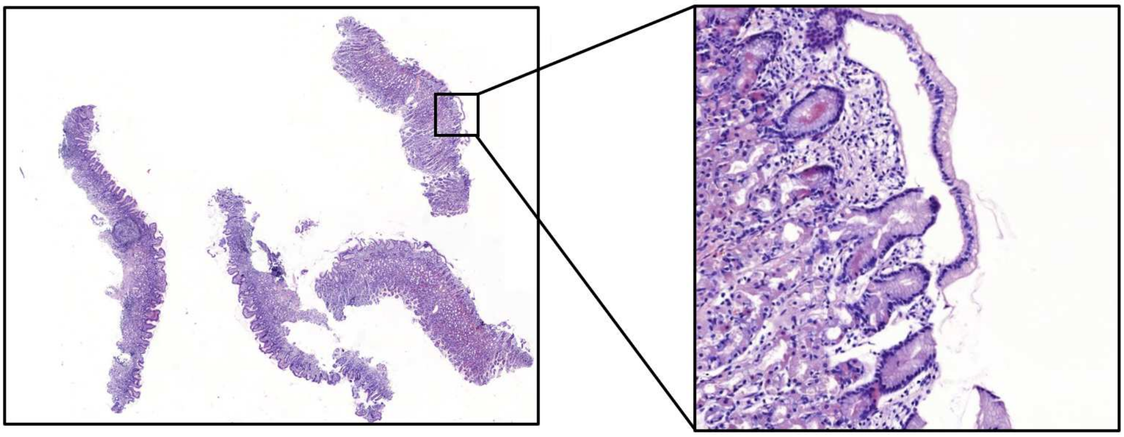


Figure 12: Left side: excerpt of a WSI approx. 8 mm wide ; right side: extracted tile approx. 600 μm

thermore, the author corrected an error in the original dataset's annotation, where instances belonging to other classes were inaccurately attributed to the inflamed class.

5.3 Data Cleaning

Different data cleaning techniques were employed to address the challenges associated with training a machine learning model at the tile level and subsequently applying this model for WSI classification.

The process of extracting the tile dataset, which encompasses various annotations, was carried out using QuPath. During this extraction, a downsampling factor of 10 was utilized to strike a balance between resolution and computational efficiency. This resulted in tiles of 256 by 256 pixels, featuring an overlap of 64 pixels to ensure thorough coverage and continuity of the tissue sections. A visualization of the resulting tiles can be seen in Figure 12.

A notable challenge was the presence of annotated blank or nearly blank tiles, which provide minimal valuable information for a CNN. A useful and straightforward approach to solve this issue is the use of a RGB Pixel threshold to detect white tiles. The tiles had to contain less then 90% of white or grey pixels to be included into the training dataset. Therefore the author built a function which converts an image to grayscale, giving pixel values from 0 (black) to 255 (white). Every pixel is then examined using a nested loop. Pixels with grayscale values above 200 are counted as white or light grey. After evaluating all pixels, the function determines the white/-grey pixel percentage of the image and compares it to a set threshold in this case 90% . This information was then used to discard blank and nearly blank tiles in the tile dataset.

While Interference on WSI-level the author encountered similar issues as during prepossessing the tile data. In this case the WSIs contained beside the already

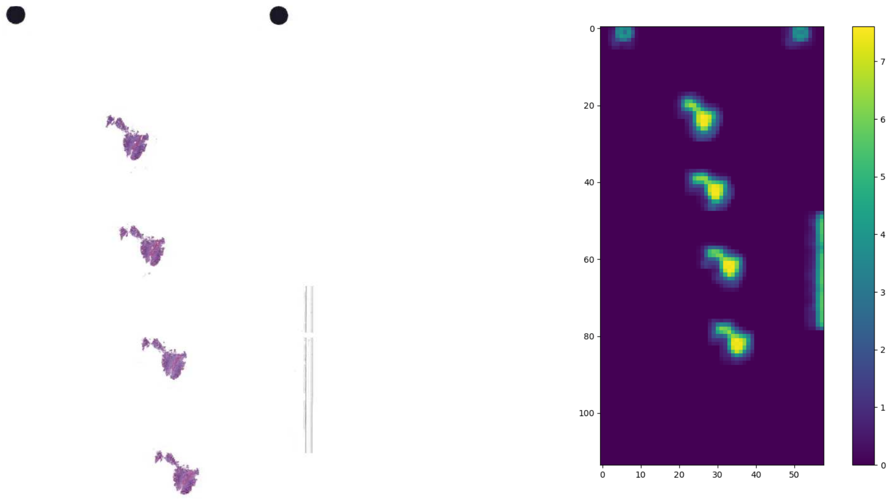


Figure 13: Example of a slide where entropy was not able to detect artifacts

mentioned white or nearly white tiles, artifacts (see Figure 3). To tackle this issue the author experimented with pixel entropy to determine which tiles contain no information and filter them out. As show in Figure 13 some artifacts have a entropy value similar to target tissue. A entropy threshold which reliable discards all artifact tiles, would also discard tiles containing tissue.

To solve this issue the author introduced a function which uses color information to find tiles of interest. First the given image is converted into a HSV (Hue, Sataration and Value) color space and then the percentage of pixels lying in a given color threshold is computed. The function returns true if the percentage reaches a certain threshold in our case 5%. With this approach and a color boundary between pink and purple, the author was able to filter out artifacts shown in Figure 13 and also blank tiles, while simultaneously detecting tissue.

5.4 Data Splitting

The data splitting process was implemented at slide level to prevent potential data leakage. This involved randomly allocating WSIs into training, validation, and test sets with an 80%, 10%, and 10% distribution, respectively. This selection was carried out using Numpy's random function, and the outcomes of this split were additionally documented and made available as a CSV file alongside the dataset. For the inflamed Dataset all tiles of a WSI with a inflamed label, where also labeled as inflamed, regardless of the degree of the inflammation. On the other hand, for the Tissue Dataset, a WSI may comprise various tissue types, as detailed in the section 'Data Collection'. Here, each tile is labeled according to its annotation in QuPath. In total, the tile datasets encompasses approximately 34,000 tiles each.

Because Tom Hempel's work also includes a baseline model, this approach of a fixed train, validation and test split allows the author and others to compare the results of theses. In the next step data cleaning as prior described was conducted

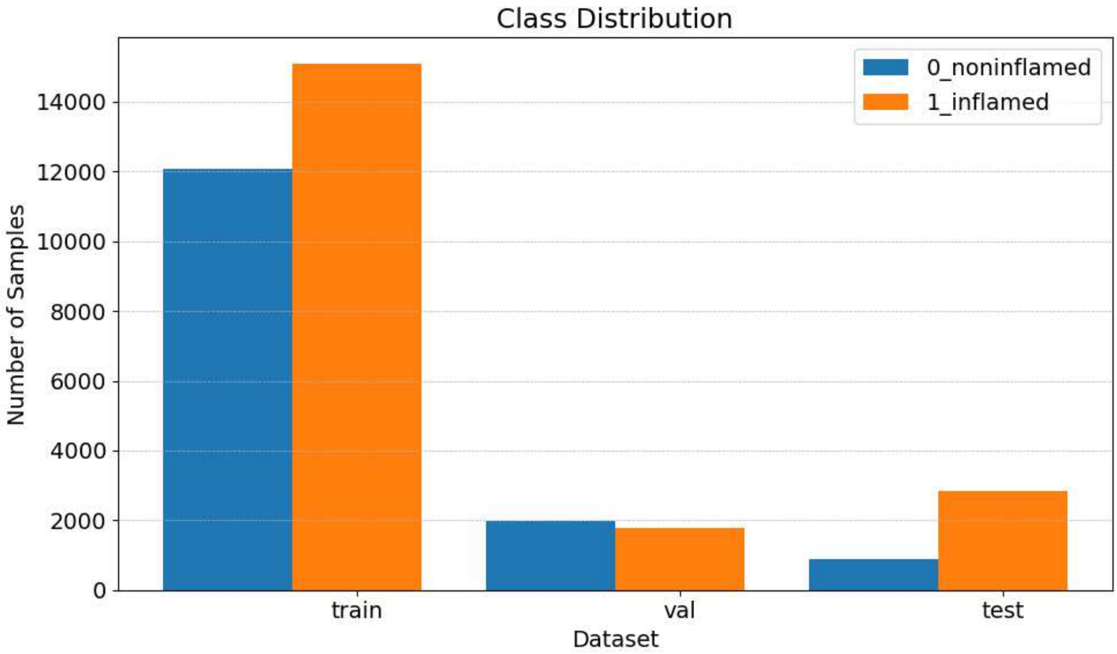


Figure 14: Distribution of the classes in train, validation and test for the inflammation dataset

on tile level while splitting the annotated tiles in the corresponding datasets. The final distribution for training, validation and test can be seen for the inflammation dataset in Figure 14 and for the tissue type dataset in Figure 15.

The distribution of gastritis cases at the tile level within the inflammation dataset is presented in Figure 16. This illustration shows a variation in the frequency of gastritis subtypes across different datasets, most notably between the test and validation sets.

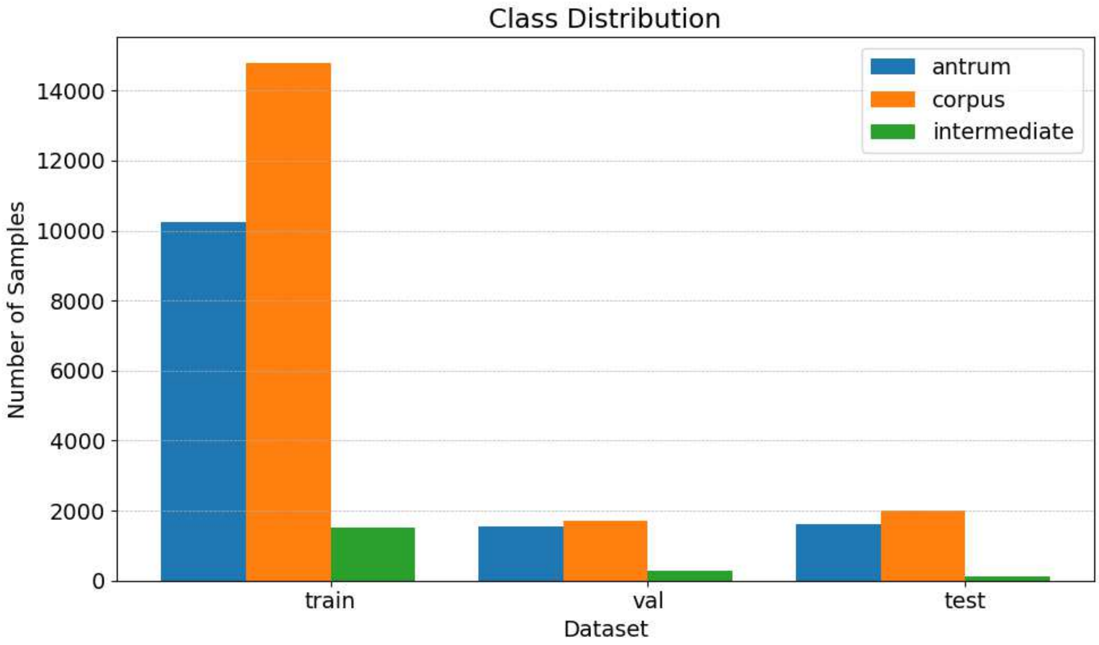


Figure 15: Distribution of the classes in train, validation and test for the tissue type dataset

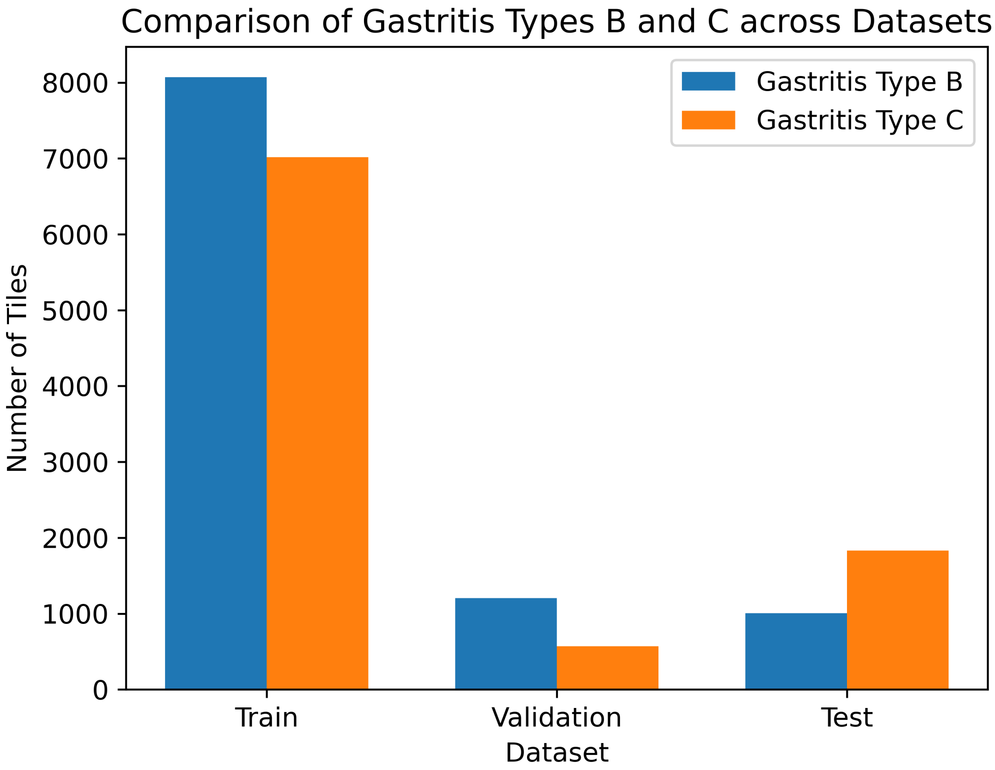


Figure 16: Distribution of the gastritis subtypes in train, validation and test

6 Methods

This part of the thesis describes the methods used to develop gastric tissue classification algorithms. The focus is on the use of the ResNet18 architecture, with an iterative process of testing and improvement. The steps of training, validation and evaluation of these models are explained in detail.

6.1 Technical Framework

This subsection outlines the technical infrastructure and tools employed in the development and training of the AI models for the classification of gastric tissue in CPath.

The development of the AI models was primarily conducted using the Dataspell Integrated Development Environment (IDE). This development environment was chosen due to its features, user-friendly interface, and integrated support for Git, which streamlined version control and code management. Its compatibility with a wide range of plugins and tools made it an ideal choice for this project, enabling efficient workflow management and code development.

The initial training of the baseline models was performed on a local machine equipped with a NVIDIA GeForce RTX 3070 Ti graphics card. This setup provided the necessary computational power to process the dataset development and perform preliminary model training and testing. As the project advanced, further model training and refinement were conducted on a server provided by the Otto-Friedrich university Bamberg, which held a NVIDIA A100 80GB PCIe graphics card. The A100's enhanced computational capabilities enables faster processing times, crucial for computationally intensive tasks such as grid search.

Both systems utilized Python and PyTorch, an open-source machine learning framework that provides a comprehensive platform for building, training, and deploying deep neural networks with a strong focus on research and experimentation in the field of artificial intelligence. (Paszke et al., 2019)

Throughout the project, Git was employed as the version control system, with the repository hosted on GitHub. This approach facilitated code tracking, collaboration, and provided a secure backup for the project codebase. The link to the GitHub repository can be found in the Appendix A.1. Towards reproducibility all procedures involving NumPy random number generation, such as data splitting, were controlled using a fixed random seed (42). This approach ensures that these operations can be replicated with the same outcomes.

Because the proprietary and multi-resolution MIRAX file format is not supported by standard Python image libraries, the author utilized OpenSlide's Python implementation to access the WSIs for inference. Openslide is an open-source C library that provides a simple interface to read whole-slide images and supports a range of proprietary formats. (Goode et al., 2013)

6.2 Evaluation Methods

Assessing models on a test set in deep learning is essential as it offers an assessment of the model's performance on new, unseen data. This evaluation helps in understanding the model's generalization ability and ensures that it performs well beyond the specific examples it was trained on. This section presents the methods used to evaluate the models in this thesis and provides the reasoning behind the choice of these methods.

Tile Level In the tile-level analysis, the performance of the model on individual image tiles is examined, which is essential for understanding its effectiveness on whole-slide images. The test dataset's ground truth was obtained using PyTorch's ImageFolder. For the model evaluation on the test dataset, a range of metrics, as detailed in subsection 3.2, was employed. Accuracy was used to measure overall performance, while Recall and Precision were assessed to determine the model's ability to correctly identify true positive cases and its precision in classification, respectively. The F1 Score, a combination of Precision and Recall, was utilized. Additionally, the ROC Curve and AUC were generated to evaluate the model's diagnostic ability across different threshold levels, crucial for understanding the balance between true positive and false positive rates. Confusion matrices were also used, providing a visual representation of the model's correct and incorrect predictions and highlighting potential biases towards a certain class. All these metrics were computed using the scikit-learn Python package, ensuring robust and reliable statistical analysis. (Pedregosa et al., 2011)

To compare different deep learning models without overfitting to the test dataset, BCE Loss and Accuracy on the validation set were prioritized. These metrics not only guided the model selection process but also offer insights into the model's learning progress during training, such as variations in validation loss.

Each of these methods contribute to a comprehensive understanding of the model's predictive capabilities and limitations, ensuring a robust evaluation framework for the developed algorithms.

WSI Level The evaluation at the WSI level adopted an unique approach compared to the tile level analysis, emphasizing the model's practical application in a clinical setting and therefore how well the model can perform on WSIs.

As detailed in Figure 23, the evaluation involved a specialized WSI-test pipeline. The initial step in this pipeline was the division of the WSI into individual tiles. This was achieved using OpenSlide's DeepZoomGenerator, which segmented the WSIs into smaller and manageable sections of 224 times 224 pixels with a similar resolution as the tile dataset. Each tile underwent a two-step evaluation process. Firstly, the presence of tissue within the tile was verified using the color threshold method described in 5.3. If tissue was detected, the tile was further analyzed using either the inflammation classification model and the tissue classification model. The inflammation model assessed whether the tile contained inflamed tissue, while the tissue type model categorized the tile as either antrum or corpus tissue.

Upon completion of these assessments for each tile, the results (ranging from 0 to 1) were aggregated to form a comprehensive overview of the entire WSI. The slide level label, which was obtained by a script from tile level, was then compared with the aggregated result. A short experiment was conducted to evaluate different tile aggregation strategies on the test dataset in section 7.4. Further the probabilities of the tiles were visualized in a heatmap, to help the pathologists and developers find regions of interest.

6.3 Training and Validation

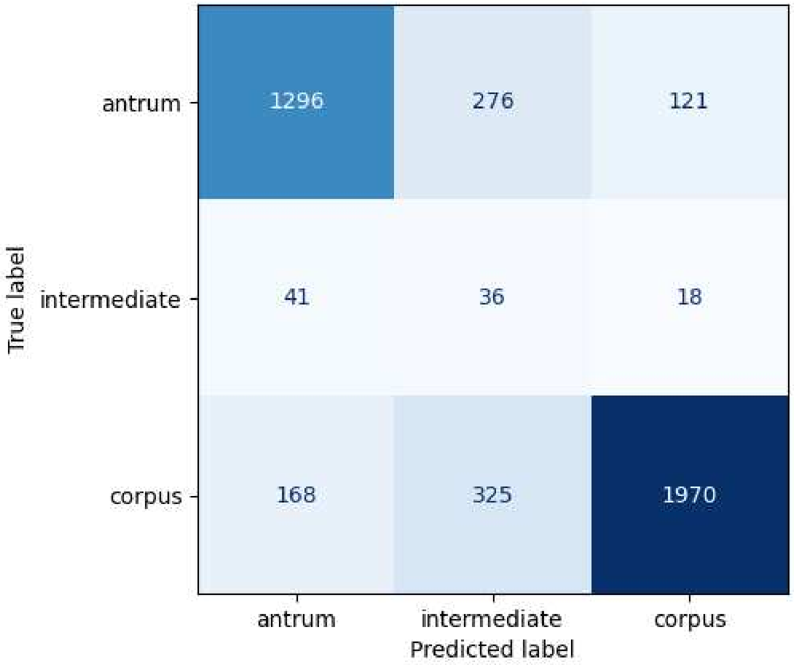
The following chapter presents methods used for training and validation of the ResNet18 models presented in this work. Because of the novelty of the research area, an iterative approach of prototype models was used at the beginning.

The author started his search for a suitable deep learning architecture with ResNet18 due to the previously presented trade-off between the simplicity, short computation time and high accuracy. The models were built using pretrained weights from ImageNet.

In an early stage of the project the author tried to formulate the tissue classification problem as regression task to determine if a deep learning model is capable to learn the intermediate class. The author trained a Regression ResNet18 with MSE Loss, a learning rate of 0.0001 and a batchsize of 32, the intermediate class (1) was between the antrum(0) and corpus(2) class, so that the classification problem represents the pathological transition between corpus and antrum. Despite this approach, the model continued to show limitations in correctly identifying intermediate tissues, as seen in Figure 17 the model struggled to classify the intermediate and further the overall accuracy of the model on the test set worsen. This can be attributed to two main factors: the complexity of classifying intermediate tissues and the limited amount of data available for these tissue types (see Figure 15. Due to this results the author has chosen with the project participants to narrow the scope of the classification down to a antrum and corpus problem.

For the model's final layer, a single output fully connected layer was used, with Binary Cross-Entropy (BCE) Loss as the loss function. In detail PyTorch's BCEWithLogitsLoss was used, as it combines BCE Loss with a sigmoid layer, which makes it more numerical stable. Both SGD with momentum (momentum of 0.9) and the ADAM optimizer were tried as optimizer to determine their effectiveness.

Training and validation data, organized using PyTorch's ImageFolder, were used as inputs. During each training epoch, the model's accuracy and loss for each batch were calculated, followed by the assessment of validation loss and accuracy at the end of the epoch. To prevent the model from just memorizing the training data, a method called early stopping was used. If the model's performance didn't improve after 15 training cycles, called epochs, training was stopped. This not only helps



True label \ Predicted label	antrum	intermediate	corpus
antrum	1296	276	121
intermediate	41	36	18
corpus	168	325	1970

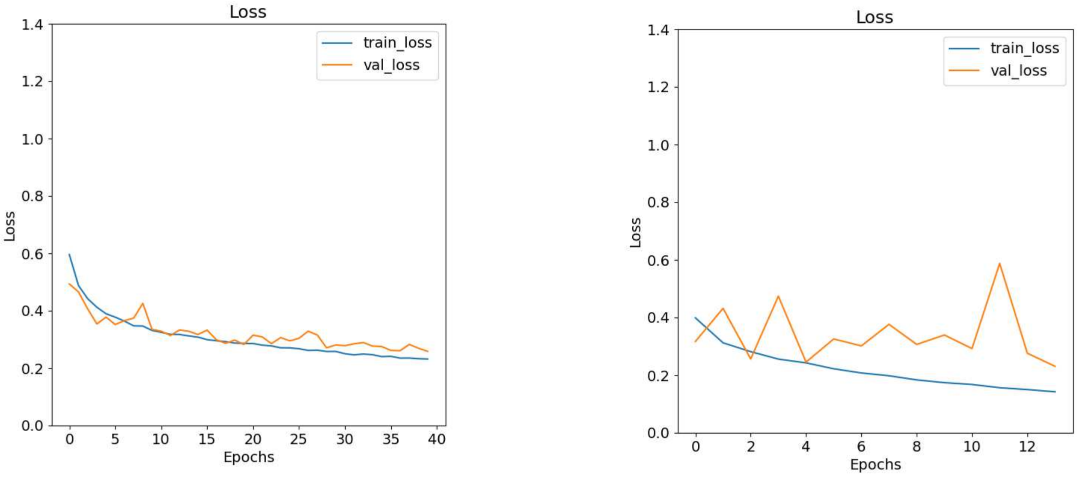
Figure 17: Confusion matrix of ResNet18 Regression model with intermediate class

to avoid overfitting but also cuts down on unnecessary training time. Each model was allowed to train for up to 60 epochs. The best-performing model, determined by the lowest validation loss at the stopping point, was saved for later review. This was done using Python's pickle operation, which turns the model into a byte stream for future analysis.

6.4 Analysis Loss Functions

During initial iterations of developing the inflammation baseline model, random spikes in validation loss were observed. After evaluating different error sources which could cause the irregular learning behaviour of the ResNet, the author found that using SGD with momentum as optimizer had a positive effect on the smoothness of the validation loss curve. The author therefore trained two ResNet18 on the inflammed dataset using the following setup, data augmentation random flipping and rotating, no learning rate scheduler, BCE-Loss, a learning rate of 0.0001 and a batch size of 32. One model used the ADAM optimizer while the other employed SGD with a momentum of 0.9.

The comparison of the training behaviors of these models, as illustrated in Figure 18, highlights the impact of the optimizer choice. Despite a relatively low learning rate, the model trained with the ADAM optimizer showed spikes in validation loss. This could be attributed to potentially improper ADAM settings, such as the learning



(a) Training and validation loss curves for SGD optimizer trained model

(b) Training and validation loss curves for ADAM optimizer trained model

Figure 18: Comparison of SGD and ADAM optimizer trained models

rate and momentum calculation parameters (betas), or due to outliers and noise in the training and validation datasets. Given the more consistent training behavior observed with the SGD with momentum model, it was chosen as the optimizer for all subsequent models in this study.

7 Experimental Setup

The following section details the experimental setups explored in the development of the AI models for classifying gastric tissue. This setup forms the foundation for the model development and evaluation discussed in subsequent sections of the thesis.

7.1 Data Augmentations

In order to enhance the generalization capabilities of the Baseline model and mitigate the risk of overfitting due to limited training data, various data augmentation techniques have been explored in this study. Data augmentation is an important method in deep learning-based image classification problems, as it artificially increases the size of the training dataset by applying a series of transformations to the original images, resulting in varied and diverse examples for the model to learn from.

In the following subsection the augmentation pipelines used in the experiment are introduced and examples of the transformations are shown. Utilizing the built-in functions provided by PyTorch's torchvision.transforms module ensures a fast and efficient integration into the training pipeline. The random application of these aug-

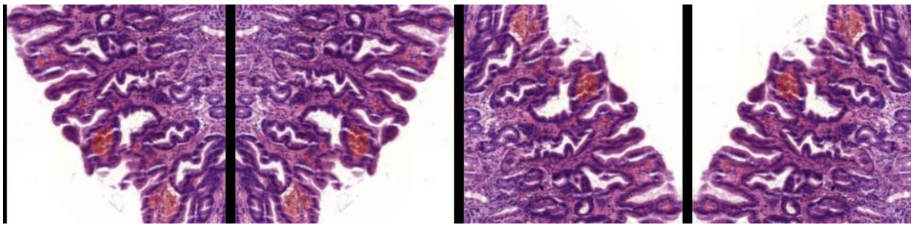


Figure 19: Examples of images generated by random flips

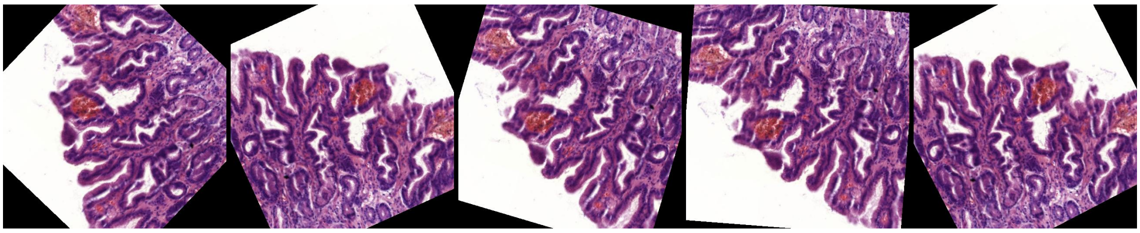


Figure 20: Examples of images generated by random rotations

mentation techniques in the training dataloader ensures that the model is exposed to a unique set of training data in each epoch.

Random Flips: Random Flips are designed to address the variability in orientations encountered in WSIs. As highlighted in Section 3.2, the orientation of WSIs does not significantly impact the diagnostic features, which distinguishes this application from other image classification tasks. To enhance the model's ability to generalize across different orientations, the pipeline integrates random horizontal and vertical flips. These flips are applied with a 50% probability along the respective axis. The augmentation is visualized at an example tile in Figure 19.

Random Rotations: A random rotation transform was applied to images, allowing them to be rotated by 180 degrees. This adjustment may aid the model in becoming insensitive to the angle of the image, ensuring that rotation of a tissue in a WSI does not affect the models capability to correctly classify the tissue. The augmentation is visualized in Figure 20.

Color Jitter: Consistent with findings noted in previous sections 3.2, there was variability in the color intensity of H&E stains on the WSIs. This observation was validated by the senior pathologists. To tackle this issue, PyTorch's transforms.ColorJitter was used with parameters that have been fine-tuned by the author for possible color differences in H& E WSIs. The transformation adjusts brightness, contrast and saturation by 20% and 2% for hue, introducing a controlled spectrum of color diversity. This ensures that the model is exposed to a wide range of color conditions, simulating the variability encountered in practical diagnostic settings. The author expects that the pipeline will increase the resistance of the model to

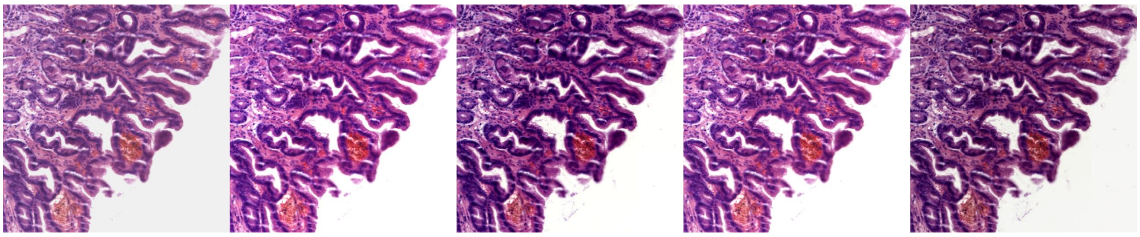


Figure 21: Examples of images generated by Color Jitter

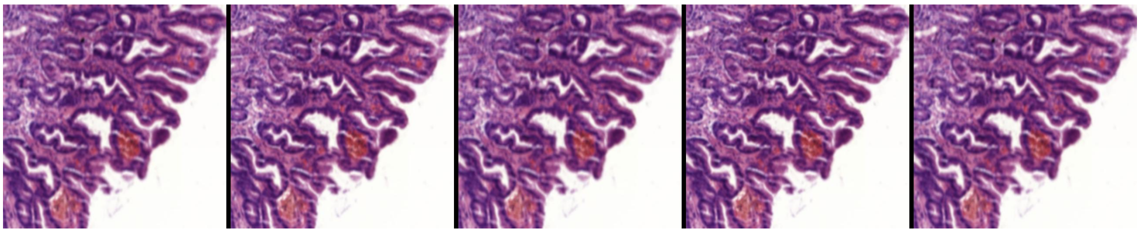


Figure 22: Examples of images generated by gaussian blur

color variation to ensure consistent performance across WSI data from different laboratories. Possible variations of data generated by this pipeline are visualized in Figure 21

Gaussian Blur: Given that some WSIs may appear blurry or out of focus to a certain degree due to differences caused by scanner quality and the digitization process, Gaussian Blur aims to train the model to be less sensitive to such artifacts. Gaussian Blur is a filtering technique used to smooth images and reduce noise and detail by averaging the pixels. The ‘transforms.GaussianBlur‘ is applied with a kernel size of 5 times 5 and a sigma between 0.1 and 5.0 to introduce a degree of blurring that is representative of potential fluctuations during image acquisition. The author expects that by smoothing out the images, mimicking the effect of a slight blur, the model learns to recognize important features without being dependent on the image’s sharpness.

Each augmentation technique was initially tested in isolation to assess its singular effect on the model’s performance. Here, the validation accuracy and validation loss were used as indicators of the augmented model’s performance. Further to test the possible synergistic impacts, the author tested combinations of these augmentations. The best performing combination was chosen to be used in further experimental setups.

7.2 Hyper Parameter Tuning

In deep learning, the grid search is an exhaustive search method used to optimize hyperparameters, the parameters that govern the training process of a neural network. During a grid search, the algorithm evaluates a model for each combination of hyperparameter values specified within a predefined grid, allowing for the systematic assessment of model performance metrics, such as accuracy or loss. This process assists in identifying the optimal set of hyperparameters that lead to the best generalization performance of the model on unseen data.

The author decided to test learning rate and batch size to find the best combination of these parameters to train future models and to find a basis for following experiments (see code snippet 1)

Listing 1: Hyperparameter Candidates for Grid Search Exploration

```
'learning_rate': [0.01, 0.001, 0.0001, 0.00001],
'batch_size': [16, 32, 64, 128]
```

The grid search was selected as the hyperparameter optimization algorithm due to its straightforward implementation and the comparable small hyperparameter space.

The implementation of learning rate schedulers was also tested. Learning rate schedulers adjust the learning rate during training, typically reducing it as the number of epochs increases. This approach can lead to more precise model tuning, as it adapts the learning process based on the training progress. Additionally, the role of dropout was considered as a technique to prevent overfitting. By randomly deactivating a subset of neurons during training, dropout adds a layer of complexity to the model's learning process, encouraging the development of more robust features.

7.3 Human Performance Evaluation

The objective of this experimental study was to evaluate how accurately the senior pathologists from Klinikum Nuremberg, can classify gastric tissue slides, and to compare their diagnostic performance with that of an AI-based algorithm developed in this thesis. Both pathologists bring extensive diagnostic experience with stomach biopsies from their clinical practice.

The testset consists of 15 anonymzied H&E stained WSIs of gastric tissue, sourced by the testset described by (Hempel, 2023) and 60 tiles (256 times 256 pixel). Additionally, the study utilized 60 tiles, each measuring 256x256 pixels. Four tiles were randomly selected from each test set slide, encompassing various tissue types, gastritis types, and pathologically irrelevant cases.

In the experimental procedure, the pathologists were first presented with the 15 H&E stained digital WSIs. Their task was to classify whether each WSI depicted chronic inflammation and to determine the gastric tissue type, categorizing them as antrum, corpus, or intermediate tissue. This classification task was then replicated with the 60 individual tiles. Unlike the previous step, each pathologist performed this

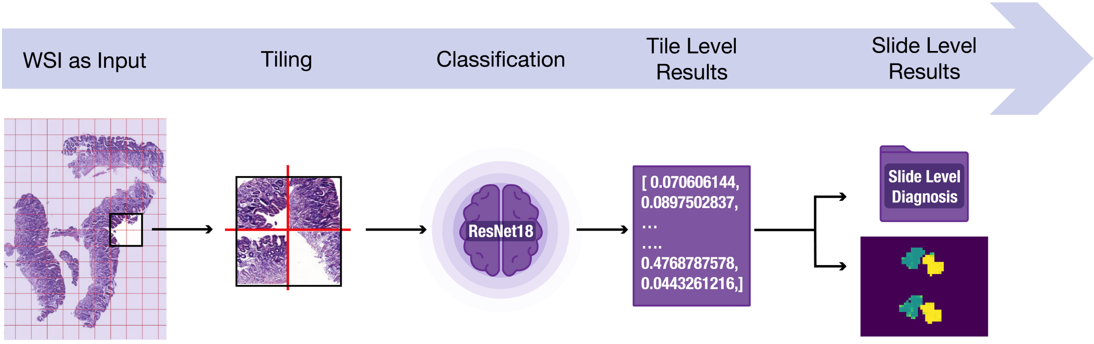


Figure 23: Simplified visualization of the framework for evaluating on WSI

classification independently. The experiment's results were benchmarked against the annotations provided as ground truth.

Notably, the author decided not to measure time efficiency in the pathologists' performance. This decision was made considering that digital slides and tiles are not commonly used as diagnostic bases in routine pathology practice. It was anticipated that introducing time as a factor might adversely affect the pathologists' performance, given their unfamiliarity with digital mediums in diagnostic contexts. Standard metrics such as confusion matrices, F1 scores, and the area under the Receiver Operating Characteristic curve were utilized for this comparison. Additionally the results of the human experts were systematically compared to the outputs of the AI algorithm using the same test set. The inter-rater agreement was measured using Cohen's Kappa and evaluated after Landis et al. (Landis and Koch, 1977) All mentioned metrics were calculated in Python with SKlearn, the code for the evaluation of the results can be found in the GitHub repository.

Further statistical analysis was conducted to assess performance differences between the pathologists and the AI system, as well as to examine inter-observer variability among the human experts to further contextualize the performance of the AI model.

7.4 Slide Level aggregation methods

This section of the experimental setup outlines the techniques used to combine predictions at the tile level to generate classifications on the slide level.

The test pipeline for models on WSIs is depicted in Figure 23 Initially, the WSI is loaded and segmented into tiles. Tiles devoid of tissue are then discarded. Subsequently, each tile undergoes classification using the developed model. Finally, the model's outputs are aggregated to provide a slide-level result and generate a corresponding heatmap. Two approaches for aggregation were tested: a majority voting method and a thresholding approach based on confidence scores. Although max pooling is a common choice in CPath cancer research (Campanella et al., 2019), there is a high risk of misclassifying random tiles in gastric tissue. Moreover, there exist tiles in a slide that could be categorized as non-inflamed even though they exhibit signs of inflammation, owing to the nature of the classification task.

The majority voting approach is based on determining the most prevalent class among individual tile predictions in a WSI. Each tile's predicted class represents a vote, and the overall slide classification is designated based on the class receiving the majority of votes. This method assumes equal weight for all tiles, irrespective of their confidence score, and concentrates only on the frequency of predicted classes. The implementation defines that a slide is classified as non-inflamed if over 50% of the relevant tiles are predicted as such. Similarly, a slide is classified as inflamed if the percentage of tiles predicted as inflamed meets or exceeds the 50% threshold.

The probabilistic averaging method differs from the majority voting approach in that it determines the overall classification of a WSI through aggregating the individual predictions of each tile. The method takes the sum of probabilities for each tile's prediction and calculates their average. The slide's final classification is determined by comparing this average probability to a predetermined threshold. The probabilistic averaging approach emphasizes the cumulative confidence level across all tiles, unlike the majority voting method, which equally weighs all tiles regardless of their prediction confidence. This approach seeks to enhance the accuracy of slide-level classification. If the average probability of the tiles being inflamed is equal to or greater than the set threshold, the slide is classified as inflamed. Otherwise, it is classified as 'non-inflamed'. The classification threshold is established using experimental analysis.

Both methods were tested and compared to determine their effectiveness in aggregating at the slide level. The performance of each method was measured based on their accuracy to correctly classify slides from the test set.

8 Results

In the following section, the results of the experiment and this work are presented. Central to these results are the two ResNet18 models developed, along with their performance on the tile and slide datasets. The performance of these models is then compared to that of human pathologists and discussed in relation to related work in the field.

8.1 Data Augmentation Results

This section presents the outcomes of the data augmentation experiments, as outlined in 7.1. The findings are detailed for two distinct models: the inflammation dataset model, and the tissue dataset model. These results offer insights into the

Augmentation Combination	Epochs	Train Loss	Train Accuracy	Val Loss	Val Accuracy
Gaussian Blur	11	0.2372	0.9003	0.2677	0.8878
Flips	9	0.2805	0.8776	0.2786	0.8829
Rotation	55	0.1885	0.9218	0.2468	0.9062
Color Jitter	15	0.3419	0.8430	0.4003	0.8137
All Combined	45	0.4550	0.7769	0.3442	0.8469
Blur + Rotation	52	0.2401	0.8966	0.2002	0.9174
Flips + Rotation	10	0.3262	0.8558	0.2776	0.8787
Blur + Flips + Rotation	47	0.2487	0.8935	0.2032	0.9177

Table 1: Performance metrics for each augmentation combination on the inflammation ResNet18 model

impact of various augmentation techniques on model performance and generalizability.

8.1.1 Inflammation Classification Model

The experiments were carried out utilizing a pre-trained ResNet18 model from the ImageNet dataset. All models underwent training with a fixed learning rate of 0.0001 and a batch size of 32, using the inflammation dataset. The training process continued for a total of 60 epochs, with early stopping implemented after 15 epochs. The outcomes of the various augmentation methods applied are presented in table 1.

Analysing the individual augmentation techniques, random rotation was found to have the best influence on the models performance to learn. Achieving the lowest training loss and lowest validation accuracy after 55 epochs of training. The color jitter augmentation on the other hand seems to have a negative impact on the models capability to learn from the training data.

Furthermore, the author decided to test various combinations of the mentioned augmentation methods. The most effective combination, consisting of random blur and random rotations, yielded a training loss of 0.2401 and a training accuracy of 89.66%, as well as a validation loss of 0.2002 and a validation accuracy of 91.74%. In contrast, applying all augmentations simultaneously did not yield similarly strong results. This outcome suggests that excessive variation in the data might negatively affect the training process. For comparison, a baseline model without any data augmentation trained under the same hyperparameters, achieved a training loss of 0.2783, a training accuracy of 88.14%, a validation loss of 0.3160, and a validation accuracy of 85.20%.

In summary, the experiment indicates that the data augmentation random rotation and gaussian blur had a positive effect on the model's capability to generalize on the validation dataset, improving the validation loss by 0.11 and validation accuracy by nearly 6%.

Table 2: Performance metrics for each augmentation combination on the tissue ResNet18 model

Augmentation Combination	Epochs	Train Loss	Train Accuracy	Val Loss	Val Accuracy
Gaussian Blur	30	0.3729	0.8460	0.4756	0.7992
Flips	30	0.3544	0.8561	0.4675	0.7995
Rotation	31	0.3880	0.8319	0.4608	0.7964
Color Jitter	37	0.3996	0.8240	0.5041	0.7522
All Combined	57	0.4720	0.7706	0.5225	0.7467
Blur + Rotation	52	0.3976	0.8250	0.4671	0.8004
Flips + Rotation	28	0.3918	0.8297	0.4601	0.7955
Blur + Flips + Rotation	57	0.3879	0.8304	0.4832	0.7936

8.1.2 Tissue classification model

The experiment was repeated with the tissue dataset. The results of the augmentation methods applied to the tissue dataset are summarized in Table 2. The models were trained with a learning rate of $1 * 10^{-5}$ and a batch size of 32, as this combination had shown positive results in previous experiments. SGD with a momentum value set to 0.9 was chosen as optimizer and BCE loss was utilized. The training used an early stopping criterion after 15 epochs, with a total of 60 epochs planned. Similar to the experiment on the inflamed dataset, color jitter and all combinations of augmentation combined had a negative impact on the models performance to generalize on the validation set. The most effective combination, consisting of random flips and rotation, was selected for all subsequent experiments involving the tissue classification model. This combination achieved the lowest validation loss, indicating its potential for further model enhancements.

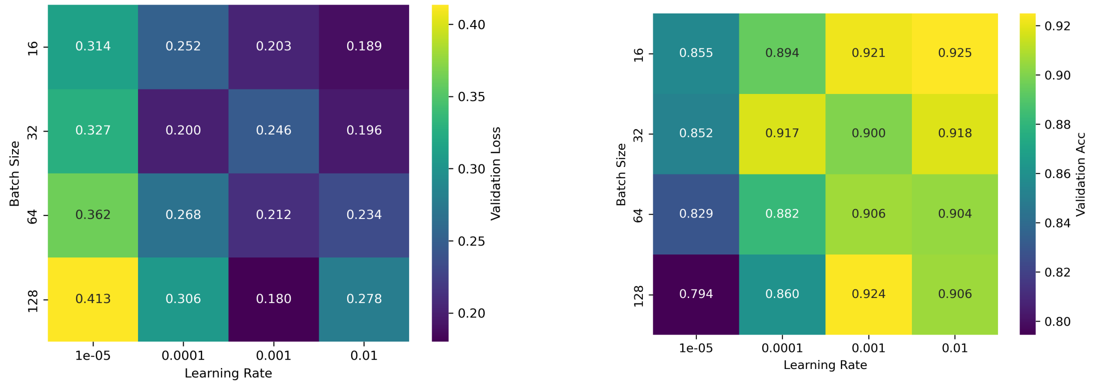
8.2 Hyperparameter Tuning Results

In the following the results of the hyper parameter tuning experiment of section 7.2 are presented. The experiments were designed to optimize the performance of the deep learning models for both inflammation and tissue classification.

8.2.1 Inflammation Classification Model

The grid search was undertaken to identify the optimal combination of batch size and learning rate for the inflammation classification model. All models were trained for a duration of 60 epochs, with early stopping enforced after 15 epochs. A learning rate scheduler was not employed, and the best combination of data augmentation from the previous experiment was utilized consisting of random rotation and random blur. The results of the gridsearch can be seen in Figure 24.

The Resnet18 model was chosen with the lowest validation loss. With a learning rate of 0.001 and a batch size of 128, it was trained for 25 epochs. The training loss is reported at 0.208 with a training accuracy of 91.01%. For validation, the model archived the lowest loss of 0.1801 and a slightly higher accuracy of 92.41%. The



(a) Heatmap of validation loss for different learning rate and batch Size combinations

(b) Heatmap of validation accuracy for different learning rate and batch Size combinations

Figure 24: Results of ResNet18 models grid search

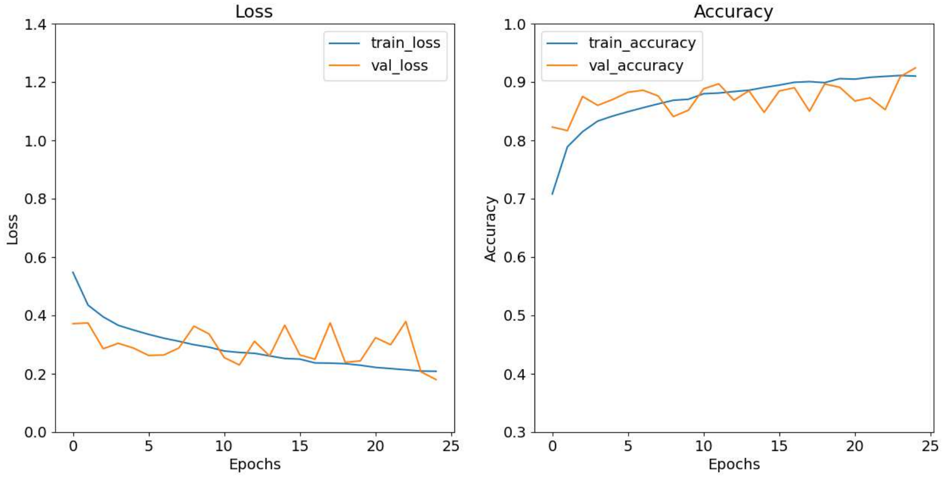
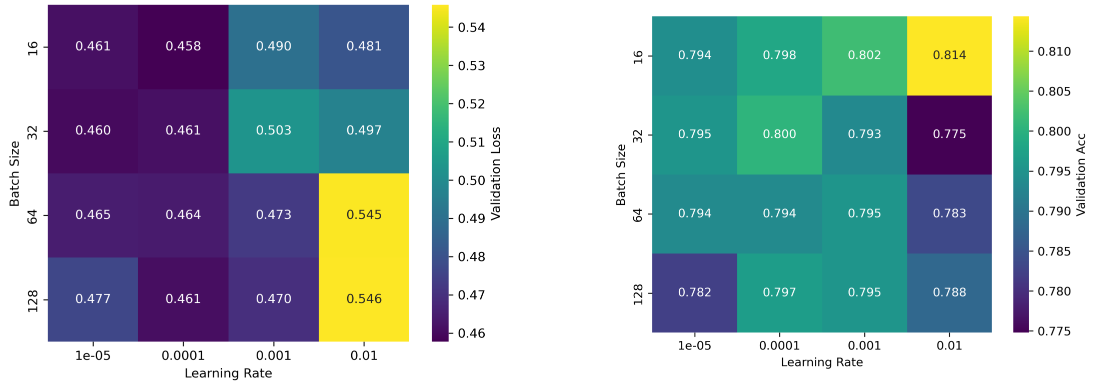


Figure 25: Training behaviour of the final inflamed ResNet18 model

training behaviour of the model is visualized in Figure 25. These figures indicate that the model is performing a little better on the validation dataset. Overall the training behaviour of the ResNet18 appears smooth with smaller amplitudes in validation loss and validation accuracy. Due to the already higher validation loss than the training loss, the author decided not to test other methods like dropout or weight decay. Further experiments and metrics are built upon this model.

8.2.2 Tissue Classification Model

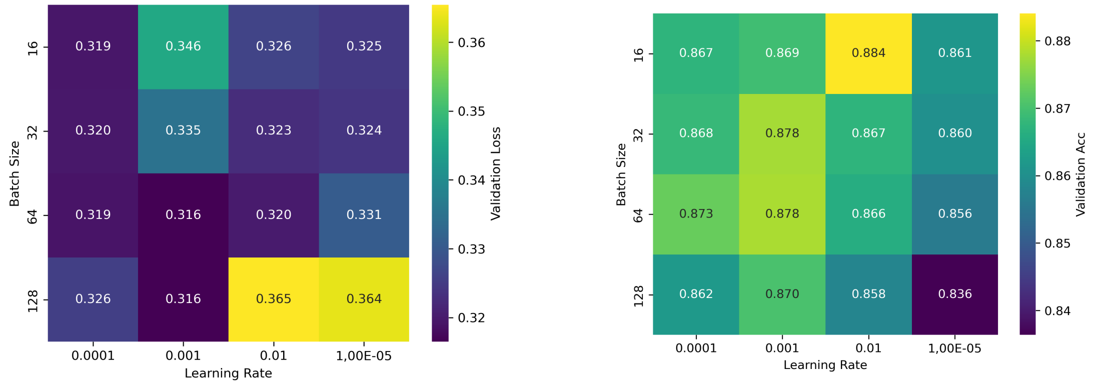
The grid search was conducted again, this time focusing on gastric tissue. For this iteration, the best combination of augmentation methods from section 8.1 was employed. Each model underwent training for a total of 60 epochs, with early



(a) Heatmap of validation loss for different learning rate and batch Size combinations

(b) Heatmap of validation accuracy for different learning rate and batch Size combinations

Figure 26: Results of the tissue classification models grid search



(a) Heatmap of validation loss for different learning rate and batch size combinations

(b) Heatmap of validation accuracy for different learning rate and batch size combinations

Figure 27: Results of the new tissue classification models grid search

stopping implemented after 15 epochs. The outcomes of this grid search are depicted in Figure 26.

The author noticed a trend of models plateauing at a certain validation loss. To further investigate this issue all missclassified validation dataset tiles from the best performing model were compiled into a CSV file for analysis. This led to the identification of certain slides with a high ratio of missclassified tiles. After another review of the dataset the author could determine certain miss annotation in Qupath with the help of the senior pathologists. The author corrected the annotations and repeated the gridsearch.

This updated grid search showed a general enhancement in validation loss, improving by over 0.13 and a rise in validation accuracy by approximately 7%. The model with the lowest validation loss was selected, which had a batch size of 64 and a learning

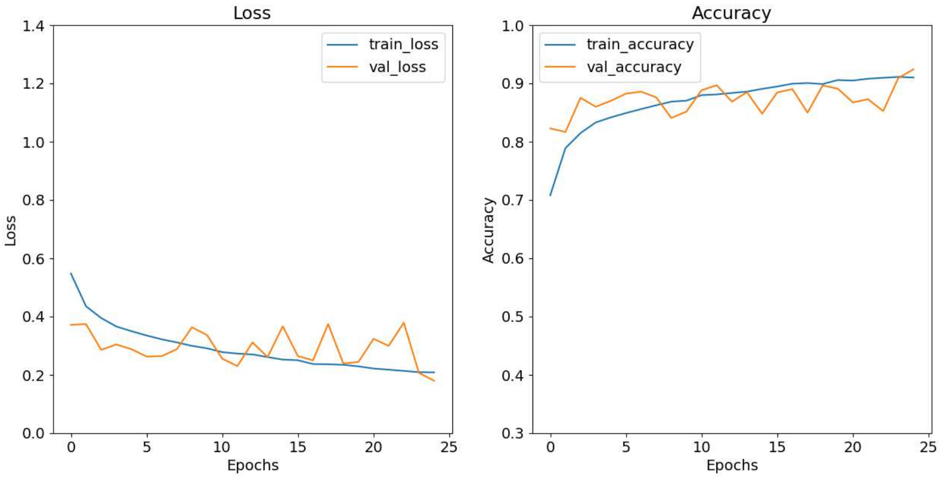


Figure 28: Training behaviour of the final gastric ResNet18 model

rate of 0.001. Implementing dropout, learning rate schedulers or weight decay did not further reduce the validation loss of this batchsize learningrate combination.

With a learning rate of 0.001 and a batch size of 64, the model was trained for 11 epochs. The training loss is reported at 0.2167 with a training accuracy of 91.77%. For validation, the model achieved a loss of 0.3165 and an accuracy of 87.80%. The train loss and accuracy, as well as the train accuracy and train validation are depicted in Figure 28.These figures suggest that the model iTrs performing consistently, though there is a noticeable gap between training and validation performance.

8.3 Tile level results

This subsection details the evaluation of the final models on the tile test datasets. Here the models' capability to classify on unseen data is evaluated, indicating the expected performance of the model on new data.

8.3.1 Inflammation Classification Model

The accuracy achieved was 72.73%, indicating a low but present ability of the model to correctly classify tiles. Precision, a measure of the models' ability to correctly identify positive cases, was notably high at 94.74%. The ROC curve of the model on the test dataset is visualized in Figure 30. Sensitivity, or the true positive rate, was relatively low with 67.86%. The models tendency for false negatives is also visualized in 29. The F1 Score, which combines precision and sensitivity, was 79.08%, suggesting an

Considering the context of the test results and the discrepancy between training accuracy and validation loss, it becomes evident that the model's performance is moderate to suboptimal. While the training accuracy was relatively high at 91.01%,

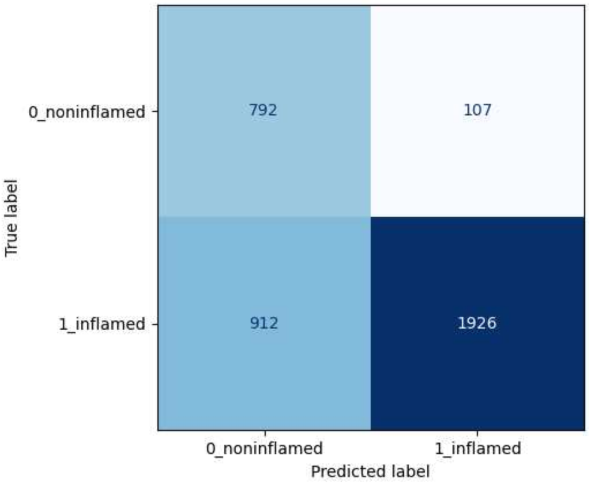


Figure 29: Confusion matrix on the inflamed test dataset

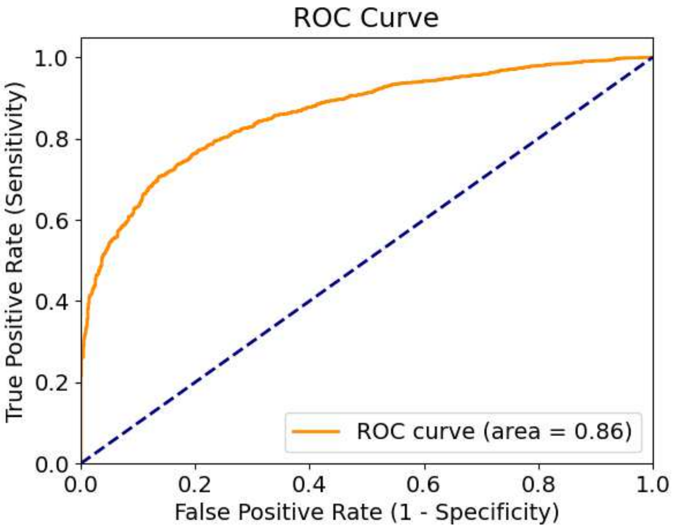


Figure 30: ROC curve over the inflamed test dataset

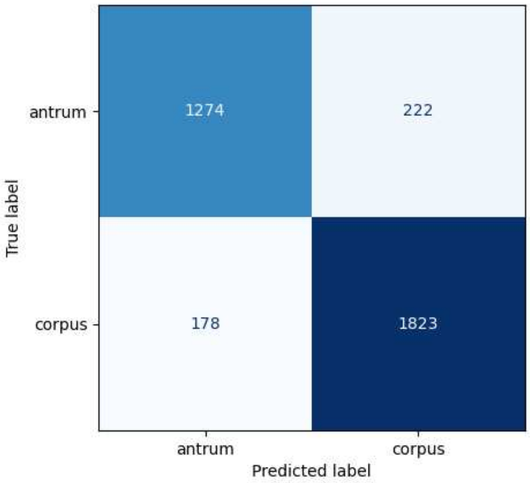


Figure 31: Confusion matrix over the tissue test set

indicating the model's effectiveness on the training dataset, the validation accuracy stood at a high 92.41%. This significant loss in accuracy while inference on the test dataset suggests issues such as overfitting on the validation dataset, lack of generalizability to new data or issues concerning the distribution and size of the datasets.

After analysing the missclassified tiles the author noticed that 761 of 912 false negatives belong to type C gastritis slides. This suggests a better generalizability of the inflammation classification model on gastritis type B cases than on type c cases. As shown in section 5.4, the testset and validation set have a imbalance in gastritis subtype frequency, may explaining this behaviour.

8.3.2 Tissue Classification Model

The Tissue classification model demonstrated a good performance in the classification of test set tiles. The model achieved an accuracy of 88.56%, reflecting its strong capability in correct tile classification. The accuracy on the test set is similar to the validation accuracy with approx. 87% indicating a good generalizability between the two datasets. The precision metric, which evaluates the model's ability to accurately identify positive cases, was high at 89.14%. Similarly, the sensitivity or the true positive rate was 91.10% The ROC curve of the model on the test dataset is depicted in Figure 32. The F1 Score of 90.11% suggests a balanced and effective model performance, indicating the model's adeptness in handling both positive and negative classifications accurately, as also visualized in the confusion matrix 31

These results highlight the tissue classification model's reliable and consistent performance on the test set, demonstrating its effectiveness in analyzing unseen data. Such performance not only indicates the model's strong generalizability but also suggests its potential in accurately detecting anatomical regions of gastric tissue at the tile level.

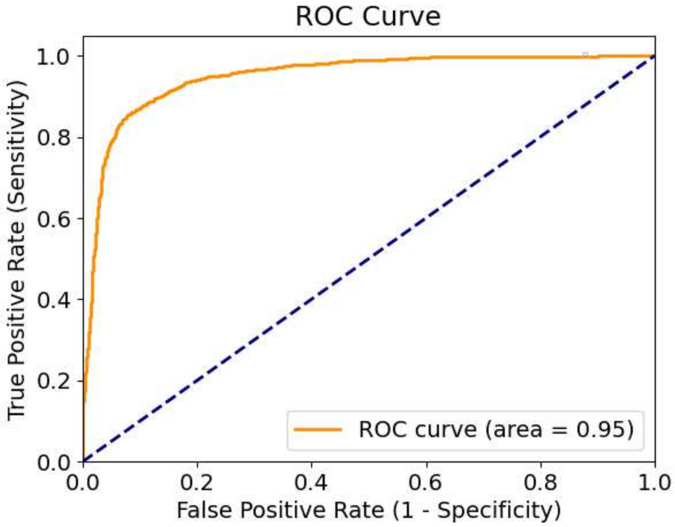


Figure 32: ROC over the gastric test set

8.4 Slide Level Results

The ResNet18 models, optimized with the best augmentation techniques and hyperparameters, was employed to evaluate on the test set WSIs. Overall the model evaluation pipeline is, while not fully optimized, relatively fast with a mean of 20 seconds for each WSI classification. In the following the results of the evaluation on slide level are presented.

8.4.1 Inflammation Classification Model

In the conducted experiment utilizing the majority voting approach on WSI level classification, the method demonstrated a limited effectiveness, misclassifying four slides of the total 19 test slides. This resulted in a relatively low overall accuracy of 73.68%, indicating potential areas for improvement in the approach.

In the probabilistic averaging method experiment, the results showed a notable improvement in the model's capability to accurately classify WSIs. By evaluating all predictions a threshold of 0.55 has been chosen and the model successfully classified 18 out of 19 test WSIs, with the only misclassification occurring in a gastritis type c slide. This method achieved an accuracy rate of 94.74%, demonstrating the benefits of using cumulative confidence levels across tiles over the majority voting method that gives equal weight to all tiles. Further the model archived a ROC AUC of 0.97. This approach is chosen for further evaluation on WSI level. This experiment's results suggest the potential of a probabilistic averaging approach in improving diagnostic precision at the slide level.

However, further evaluation with a larger set of test WSIs is necessary for a comprehensive understanding of the model's overall performance, and a more concise and better tuned threshold.

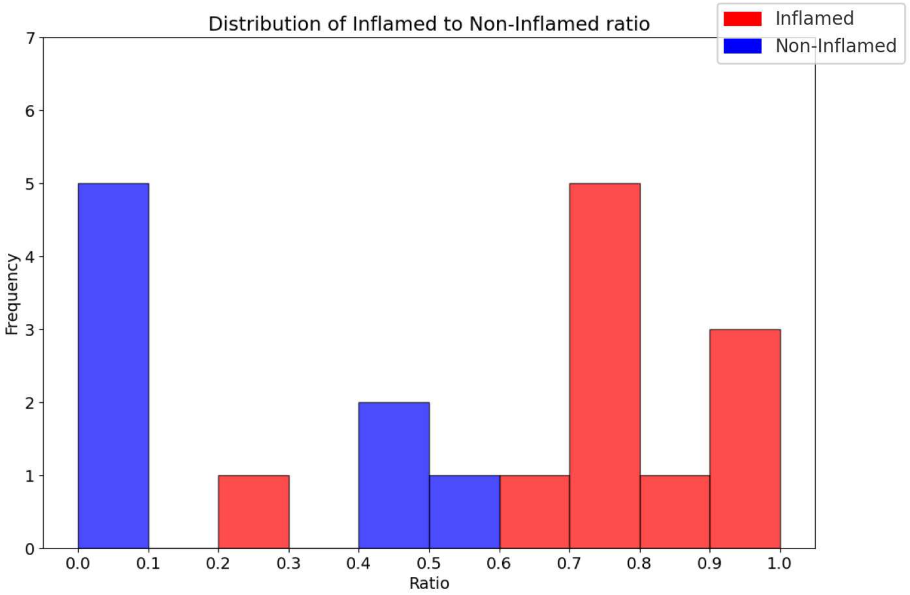


Figure 33: Distribution of the inflamed to non inflamed tile-ratio over the test set

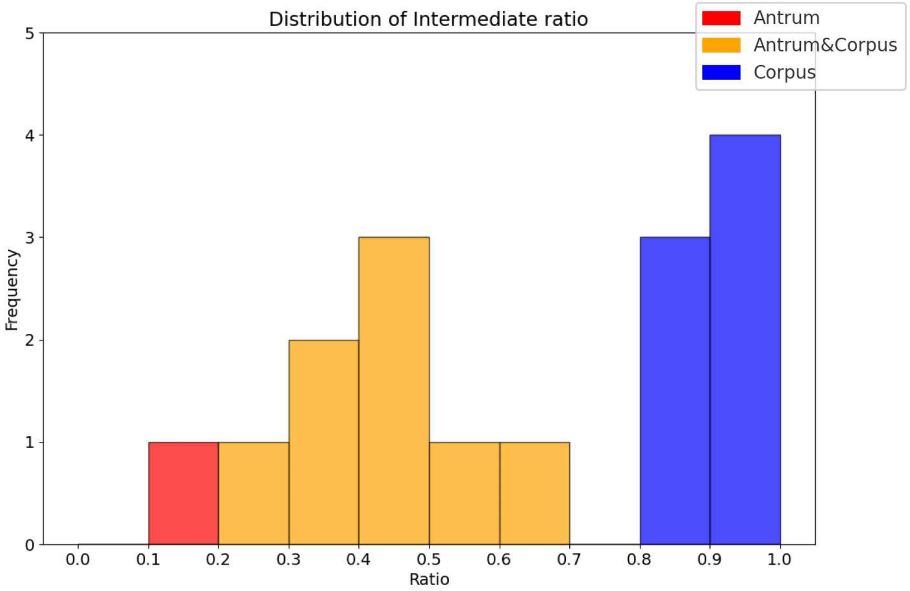


Figure 34: Distribution of the aggregated tile-ratio over the test set

8.4.2 Tissue Classification Model

Due to the more complex nature of WSI aggregation for the anatomical region classification and the moderate performance on the inflammation dataset the author decided not to test majority voting for gastric tissue and instead directly used the probabilistic averaging approach. Further all 3 slides containing intermediate tissue were discarded first from the test set, to make the aggregation problem, a 3 class

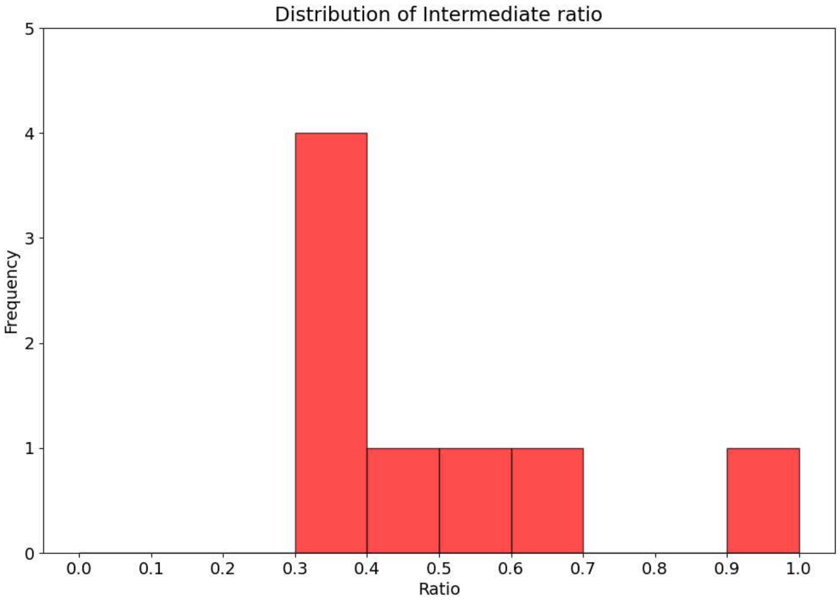


Figure 35: Distribution of the intermediate slides aggregated tile-ratio over the test set

problem of antrum, corpus and antrum&corpus. The antrum&corpus class holds all slides which contain tissue from both anatomical regions. The results are visualized in Figure 34, which demonstrates the model's capability to accurately classify slides. By setting the threshold ranges from 0 to 0.25 for antrum and 0.75 to 1 for corpus, the model achieved an perfect accuracy of 100%.

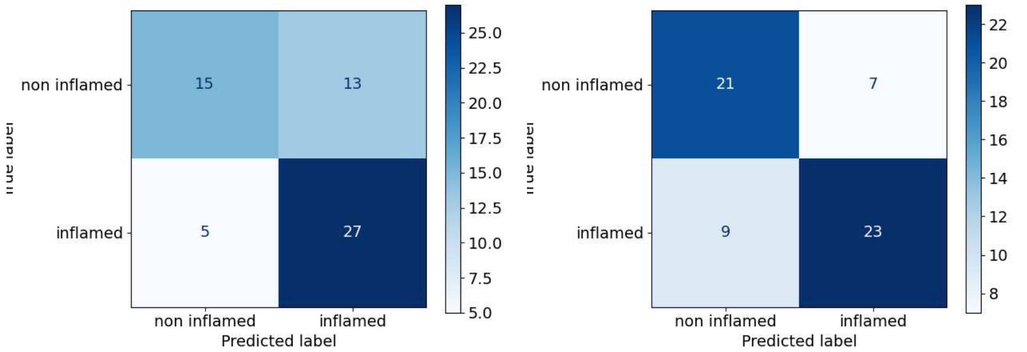
The author tested further the performance of the algorithm on slides, which only contain intermediate. These slides were novel to model, which was trained only on antrum and corpus tiles. The distribution of the averaged probabilities can be seen in Figure 35. The results suggest that while on tile level the model was not able to detect intermediate tissue between corpus and antrum, on slide level the model is mostly capable in identifying intermediate tissue as between corpus and antrum.

8.5 Human Performance Evaluation

In the evaluation on slide level, there was perfect agreement between the pathologists and the ground truth annotations. The accuracy for both pathologists at this level was 100%. However, it is important to note the relatively small sample size of 15 slides, which could influence the generalizability of these results. No additional metrics were evaluated at this level due to the pathologists' perfect performance.

At the tile level for inflamed tissue classification, the performance of the pathologists showed notable variation, as depicted in Figure 36.

The resulting metrices are presented in table 3. The Cohen's Kappa coefficient, measuring inter-rater agreement, was 0.6, indicating moderate agreement between

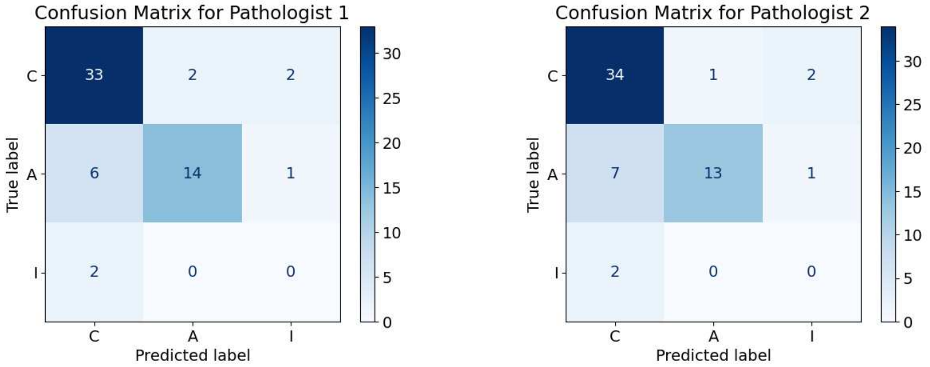


(a) Confusion matrix of Pathologist 1 (b) Confusion matrix of Pathologist 2

Figure 36: Confusion matrices for inflamed tiles

Table 3: Comparison of diagnostic metrics between the two pathologists on the inflamed test tiles

Metric	**Pathologist 1**	**Pathologist 2**
Accuracy	70%	73.33%
Precision	67.5%	76.67%
Recall	84.375%	71.875%
F1 Score	75%	74.19%
AUC ROC	0.69	0.73



(a) Confusion matrix of Pathologist 1 (b) Confusion matrix of Pathologist 2

Figure 37: Confusion matrices for Tissue Tiles

the pathologists. Given the small sample size of 60 tiles, these metrics may be subject to variability. On tile level the pathologists performed compared to the groundtrouth in classifying tissues like visualized in Figure 37. The resulting metrices are presented in table 4. To calculate the AUC ROC the author ignored the two tiles where intermediate is groundtrouth and let the three tiles where the pathologists guessed intermediate pass as correct.

Table 4: Comparison of diagnostic metrics between the two pathologists on the tissue test tiles

Metric	Pathologist 1 (%)	Pathologist 2 (%)
Accuracy	86.44	80.00
Precision	86.61	84.05
Recall	86.44	80.00
F1 Score	86.07	80.73
AUC ROC	83.08	82.00

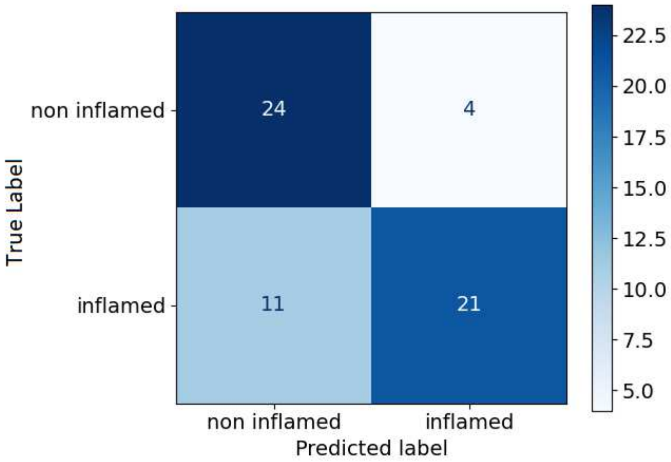


Figure 38: Confusion matrix of the inflammation classification model over the experiments dataset

8.6 Comparative Analysis

This section provides a comparative analysis between the AI models developed in this thesis and the performance of human pathologists, as well as a comparative review of related studies in the existing literature.

8.6.1 Comparative Analysis with Pathologists

Slide level At slide level, the AI model achieved an accuracy of 93.33% in the inflammation WSI experiment dataset. This result is noteworthy, but it is marginally lower than the flawless 100% accuracy attained by the human pathologists. This difference underlines the complex comprehension and interpretation skills of pathologists in slide-level diagnosis. In contrast, the AI model achieved a 100% accuracy rate in tissue classification when the 'intermediate' class was excluded. This underscores the AI model's capability in accurately identifying unique tissue types at the slide level, closely aligned with the pathologists' results.

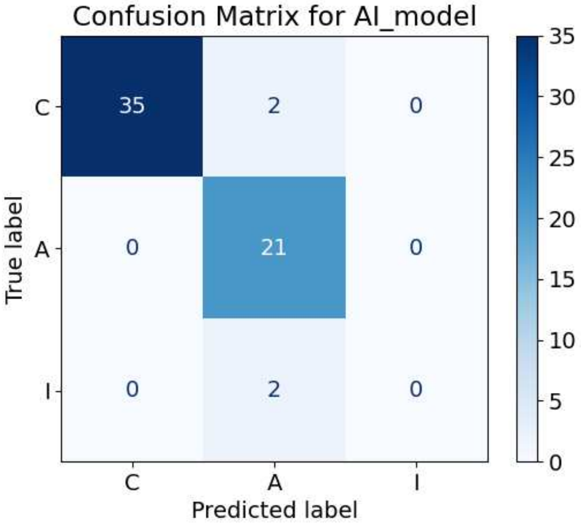


Figure 39: Confusion matrix of the inflammation classification model over the experiments dataset

Tile level The AI model, while showing clear flaws on the test tile dataset, slightly outperforms the pathologists on inflammation tile level classification as visualized in Figure 38. The model demonstrates higher accuracy 75% compared to both Pathologist 1 70% and Pathologist 2 73.33%, suggesting it is more consistent in correctly identifying both positive and negative cases. While the AI model has a higher precision 84% than the Pathologists 1, indicating it is better at minimizing false positives, in terms of recall, the AI model 65.62% is outperformed by both pathologists. The ROC AUC of the inflammation classification model is slightly higher with 0.76 then 0.69 and 0.73 for the pathologists respectively.

The Cohen's Kappa scores indicate a slight agreement between Pathologist 1 and the AI model 0.15 and a fair agreement between Pathologist 2 and the AI model 0.30. These results suggest that the AI model's classification criteria significantly differ from those of the pathologists, especially evident in the minimal agreement with Pathologist 1.

For the classification of the anatomical region the tissue classification model outperformed the pathologists in every metric employed:

- Accuracy: 93.33%
- Precision: 91.07%
- Recall: 93.33%
- F1 score: 91.91%

While the Cohen's Kappa coefficient of 0.85 between the pathologists, reflects a strong agreement, Pathologist 1's and Pathologist 2's agreement with the AI model yielded kappa values of 0.56 and 0.48, respectively. The resulting confusion matrix can be seen in Figure 39. These values suggests a moderate correspondence between the pathologists with the AI's classifications, maybe indicating different criteria for classification.

For the AUC ROC calculation, the author applied the same criteria as used in the evaluation of human performance, disregarding the two tiles where 'intermediate' was the ground truth. The AI model achieved a ROC AUC of 0.97, underscoring its robustness in distinguishing between the anatomical regions on the test tiles.

8.6.2 Comparative Review with Existing Literature

This section undertakes a comparative analysis between the outcomes of this study and related prior research within the same project and wider academic literature.

Tom Hempels Work In the context of the same project, Tom Hempel's study (Hempel, 2023) presents notable differences and outcomes. An analysis of Hempel's training and validation loss curves suggests a tendency towards overfitting on the train dataset particularly noticeable in the inflammation classification model. In contrast in this thesis early stopping mechanisms and model choice on lowest validation loss have been employed. Hempel's models, which operated without early stopping and were trained for only 7 epochs, employed a two-class output for the ResNet and utilized cross-entropy loss. (Hempel, 2023)

In terms of tissue classification, Hempel's model achieved a ROC AUC of 0.95, mirroring the performance of the model developed in this thesis. On slide level a direct comparison of slide level performance is not feasible due to differing classification objectives. Interestingly, Hempel's inflammation classification model outperforms this study's model with an AUC ROC of 0.98 in comparision towards the author's 0.86. Indicating better generalizability of Hempel's model in classifiying unseen inflamed tissue tiles. At the slide level, Hempel's model correctly classified all test slides, versus the author's model, which misclassified an inflammation slide.(Hempel, 2023)

Literature Reflecting on the discussion in subsection 3.1, results of studies regarding gastric inflammation classification are compared to this work. Steinbus et al. worked with a smaller set of tiles and achieved an overall accuracy of 84%. Their model was particularly effective in identifying gastritis type B tiles, similar indices are found in this study.(Steinbuss et al., 2020)

In another study, Martin et al. used a dataset of 400 slides and achieved impressive results. Their model reached an AUC of over 99% for gastritis types B, C, and normal tissues. This performance significantly exceeds that of the models in this thesis, highlighting the high level of accuracy that can be achieved with larger datasets. (Martin et al., 2020)

9 Discussion

In this discussion section, a examination of the thesis results is conducted, with a particular focus on evaluating the performance of the developed AI models in the classification of gastric tissue. Implications of these findings in CPath are explored, addressing challenges such as dataset limitations and annotation accuracy.. The section also highlights the potential for future research, emphasizing the importance of this study in advancing the field of digital pathology. The aim of this analysis is to offer a thorough comprehension of the study's influence and the possibilities it offers for additional investigation.

9.1 Interpretation of results

The results achieved, considering the relatively small size of the dataset, can be regarded as positive. However, the inflammation model, in particular, exhibits significant potential for improvement.

The outcome of the human performance experiment underscores the complexity of classifying inflamed tissue at the tile level. Pathologists typically diagnose gastritis type C based on a range of characteristics observed at the slide level. This disparity in diagnostic approach highlights the challenges faced by the AI model in mimicking human-level classification accuracy in tile-based analysis. Regarding the results from the data augmentations, the attempt to improve model generalizability through color variation augmentations did not yield the expected enhancements. This could be attributed to the dataset's limited color variability. The generalization of the models was only tested on WSI coming from the same laboratory and slide scanner, so no conclusion about the generalization of the models on new unseen data from other sources can be made.

The tissue classification model showed a plateau in performance beyond a certain threshold, mirroring Tom Hempel's findings. This plateau may be linked to the characteristics of the training and validation data. Nevertheless, the promising results at both the slide and tile level provide encouraging indications for the development of future clinical-grade models.

The inflammation model's poor performance on the test set points to overfitting on the validation dataset. A likely cause is that the validation dataset does not accurately represent the overall data distribution, particularly with an underrepresentation of gastritis type C cases. Additionally, type C gastritis often exhibits localized features, leading to many tiles without distinct type C gastritis characteristics. The annotation of corpus tissue from type C gastritis slides as inflamed, despite it typically exhibiting minimal characteristics of type C gastritis, further complicates the model's learning process.

Another result of this work is the development of an evaluation pipeline for models on WSI. The WSI testpipeline allows direct classification on WSI images and provides features like heatmaps for the user. These heatmaps offer intuitive visual representations, allowing users to easily interpret the model's classification outputs on WSIs. Enhanced annotation accuracy for both the inflamed and gastric tissue dataset stands out as another outcome of this study, which is important for future work in this project, enabling the training of more accurate AI models.

9.2 Clinical Implications

The initial step towards integrating AI-based diagnostic tools, such as the one developed in this study, into clinical practice at Klinikum Nuremberg involves transitioning to a digital workflow, particularly for gastric biopsies. Currently, the absence of a digital pathology workflow in Klinikum Nuremberg presents the biggest challenge in introducing an algorithm for tissue classification. The integration of digital pathology systems would not only facilitate the adoption of AI tools but also enhance overall efficiency and accuracy in diagnostic processes. This transition would require investment in digital scanners, storage solutions, and training personnel to adapt to the digital system.

The journey towards the practical application of AI in clinical pathology encompasses several challenges, particularly in obtaining legal clearance. The regulatory landscape for AI in healthcare is complex and varies by region, requiring more validation studies, ethical considerations, and adherence to privacy laws.

The potential of deep learning models, as evidenced by this study, to support pathologists in routine gastric biopsies is significant. These AI models can assist in screening and preliminary diagnosis, thereby reducing the workload on pathologists and improving diagnostic speed and accuracy. The indication is that deep learning could be particularly useful in identifying key pathological features in gastric biopsy slides, allowing pathologists to focus on more complex cases.

Drawing on the insights from the study by Campanella et al., the developed system could, with additional fine-tuning and testing, be employed to mark regions of interest in WSI, such as those with a high level of inflammation, in gastric biopsy examinations. This approach could streamline the workflow, allowing pathologists to concentrate on slides that require more detailed examination. Such systems not only optimize the use of human expertise but also minimize the risk of oversight, contributing to more reliable and efficient diagnostic processes. (Campanella et al., 2019)

9.3 Limitations

In the course of this project, limitations related to WSI scanners and image quality posed significant challenges. Technical issues rendered the WSI scanner unavailable for two months, which substantially constrained the time available for data collection. Additionally, the digitization process for a single slide was time-intensive, taking up to 10 minutes per slide, restricting the dataset's expansion. Further the image quality of the digitized WSIs was often compromised due to out-of-focus tissue, a limitation attributed to the scanner's autofocusing capabilities.

The annotation process presented a set of challenges and limitations. First inter-observer variability was a huge challenge in finding annotations for tissue. Especially in intermediate tissue classification the opinions between senior level pathologists vary. Further while the labeling accuracy and consistency was reassessed in this work, the ground truth of the dataset leaves room for improvement, due to the difficulty of the task of annotation.

A notable limitation was the mislabeling of tiles within inflamed tissue. All tissue in a slide which contains inflamed tissue was annotated as inflamed, regardless of whether individual tiles exhibited inflammation characteristics. This method was especially problematic for type C gastritis. This blanket annotation approach resulted in tiles being incorrectly labeled as inflamed, despite lacking visible signs of inflammation. Such inconsistencies in the dataset presented obstacles in accurately training and evaluating the model, impacting its ability to detect inflamed tissue at tile level.

Time constraints significantly impacted the model training phase, primarily due to delays in finalizing the dataset. The dataset was only completed in the final two months of the project, a factor that considerably limited the time available for training and fine-tuning the model and other experiments focusing on building a robust deep learning model.

10 Future Work

There is considerable scope for future development and improvement as this is the first project aimed at classifying gastric tissue.

One key area of focus is the expansion and diversification of the dataset. By incorporating data from various laboratories, the model's ability to generalize across different H&E staining protocols can be assessed more effectively. Enriching the dataset with a broader range of classes, for example intermediate tissue types and different stages of gastritis, is also crucial. This expansion would not only improve the model's accuracy but also pave the way for developing sophisticated regression models that can handle a wider variety of tissue states.

The problem of the distribution of gastritis type c in the test and validation set could be addressed by using a different data split or even cross validation. Although a larger test dataset and validation dataset could give the model more room to generalize. Reevaluating the annotations for the inflamed class stands as a key area for future research. Consistent and dependable annotations are essential for the effective training of models, and any enhancements in this field could result in notable advancements in model performance. Additionally, exploring different resolutions for extracting tiles and experimenting with varying degrees of overlap and tile sizes presents another promising field for future research.

Another Consideration in future work is the possibility to explore a deep learning model's capability to differentiate between various types of gastritis using the current dataset. This exploration could provide valuable insights into the amount of data needed for such a classification. Building a system that can analyze slides stained with different agents (H&E, PAS, Giemsa Modified) would be a interesting step. Such a system could offer a more accurate diagnosis by utilizing the distinct advantages of each staining method. For the aggregation of the tile results more sophisticated aggregation models could be implemented like, a recurrent neural network or random forest, which also use spatial features of the tile classification for their aggregation and therefore built more precise aggregation results.

Creating a user-friendly interface would be important for a clinical use of the introduced algorithms, so that pathologists can easily visualize the model's output. This system will make the results more accessible and interpretable for clinical professionals, facilitating their decision-making process. Implementing batch processing for the classification of WSIs is another important objective. This approach is expected to significantly accelerate the classification process, making it more efficient and scalable for practical use.

Investigating more advanced and deeper neural network architectures for the classification challenges at hand is also a vital part of future work. Leveraging cutting-edge developments in deep learning could further enhance the accuracy and efficiency of the models.

11 Conclusion

The primary aim of this thesis was to design, develop, and validate an algorithm tailored for the classification of gastric tissue. This work introduced two ResNet18 models for the classification of gastric mucosa biopsies WSIs along with a test pipeline for evaluating trained models on WSIs and a revised dataset of gastric mucosa biopsies. Notably, the inflammation classification model achieved a remarkable accuracy of 94.7%, while the gastric tissue classification model has a perfect accuracy of 100% on the slide test dataset. The performance of the gastric tissue classification model is particularly noteworthy, as the author found no research in this area outside of this project.

The study's findings reveal that while AI models can achieve accuracy comparable to human pathologists at slide level, the limited size and homogeneity of the dataset pose obstacles to broader applicability. Data augmentation and hyperparameter tuning were explored, but their impact was constrained by these dataset limitations. A key finding of this study for future work is the need for a more sophisticated annotation process regarding type c gastritis. Future research should also focus on expanding dataset size and variety, improving annotation accuracy, and exploring advanced deep learning methods.

In summary, this thesis represents a step forward in the application of AI for gastric tissue classification, laying groundwork for further studies in this underrepresented area of computational pathology.

A Appendix

A.1 Code availability

The code developed in this master thesis is available under "https://github.com/PhilippHoefling/W Gastric-Tissue-Classification" The dataset, final models and annotations are hosted at the the Otto-Friedrich University Bamberg's chair of Explainable Machine Learning.

A.2 Annotation Protocol by Tom Hempel

The following subsection is cited from Tom Hempels work (Hempel, 2023) 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. 7 ANNOTATIONS 21
- 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.

For further information regarding the annotation process please see (Hempel, 2023)

A.3 Qupath Tile Export Script

Listing 2: Export Script of Tiles

```
// Get the current image (supports 'Run for project')
def imageData = getCurrentImageData()

// Define output path (here, relative to project)
def name = GeneralTools.getNameWithoutExtension(imageData.getServer().getMetadata().getName())
def pathOutput = buildFilePath(PROJECT_BASE_DIR, 'tiles', name)
mkdirs(pathOutput)

// To export at full resolution
double downsample = 5

// Define tile image format
extension = '.png'

// Get all annotations in the current image
def annotations = getAnnotationObjects()

// Loop over each annotation and export its tiles
annotations.eachWithIndex { annotation, index ->
    // Define a unique directory for this annotation's tiles
    def annotationOutputPath = buildFilePath(pathOutput, "annotation_${index+1}")
    mkdirs(annotationOutputPath)

    // Create a RegionRequest for the current annotation
    def request = RegionRequest.createInstance(imageData.getServerPath(), downsample, annotation
        .getROI())

    // Create an exporter for this annotation's tiles
    new TileExporter(imageData)
        .downsample(downsample)    // Define export resolution
        .imageExtension(extension)    // Define file extension
        .tileSize(512)              // Define size of each tile, in pixels
        .annotatedTilesOnly(true) // If true, only export tiles if they are annotated
        .overlap(128)                // Define overlap, in pixel units at the export resolution
        .region(request)           // Set the region to the current annotation
        .writeTiles(annotationOutputPath)
    // Write tiles to the specified directory


    def dirOutput = new File(annotationOutputPath)
    for (def file in dirOutput.listFiles()) {
        if (!file.isFile() || file.isHidden() || !file.getName().endsWith(extension))
            continue
        def newName = file.getName().replaceAll("=","-").replaceAll("\\[","").replaceAll("\\]","
            ").replaceAll(",","_").replaceAll(".png","_") + annotation.getPathClass() + '.png'
        if (file.getName() == newName)
            continue
        def fileUpdated = new File(file.getParent(), newName)
        //println("Renaming ${file.getName()} ---> ${fileUpdated.getName()}")
        file.renameTo(fileUpdated)
    }
}

println('Done!')
```

Bibliography

Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D. Zarella, Jeroen van der Laak, Marilyn M. Bui, Venkata Np Vemuri, Anil V. Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, Andrew H. Beck, and Cleopatra Kozlowski. Computational pathology definitions, best practices, and recommendations for regulatory guidance: A white paper from the Digital Pathology Association. *The Journal of pathology*, 249(3):286–294, 2019. doi: 10.1002/path.5331.

Shiliang Ai, Chen Li, Xiaoyan Li, Tao Jiang, Marcin Grzegorzek, Changhao Sun, Md Mamunur Rahaman, Jinghua Zhang, Yudong Yao, and Hong Li. A State-of-the-Art Review for Gastric Histopathology Image Analysis Approaches and Future Development. *BioMed Research International*, 2021:6671417, 2021. ISSN 2314-6133. doi: 10.1155/2021/6671417.

Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):16878, December 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-17204-5.

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*, 16(11):703–715, November 2019. ISSN 1759-4774. doi: 10.1038/s41571-019-0252-y.

Sergio Bermejo and Joan Cabestany. Oriented principal component analysis for large margin classifiers. *Neural Networks*, 14(10):1447–1461, December 2001. ISSN 0893-6080. doi: 10.1016/S0893-6080(01)00106-X.

Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer New York, New York, NY, softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009) edition, 2016. ISBN 978-1-4939-3843-8.

Alexander D. Borowsky, Eric F. Glassy, William Dean Wallace, Nathash S. Kallichanda, Cynthia A. Behling, Dylan V. Miller, Hemlata N. Oswal, Richard M. Feddersen, Omid R. Bakhtar, Arturo E. Mendoza, Daniel P. Molden, Helene L. Saffer, Christopher R. Wixom, James E. Albro, Melissa H. Cessna, Brian J. Hall, Isaac E. Lloyd, John W. Bishop, Morgan A. Darrow, Dorina Gui, Kuang-Yu Jen, Julie Ann S. Walby, Stephen M. Bauer, Daniel A. Cortez, Pranav Gandhi, Melissa M. Rodgers, Rafael A. Rodriguez, David R. Martin, Thomas G. McConnell, Samuel J. Reynolds, James H. Spigel, Shelly A. Stepenaskie, Elena Viktorova, Robert Magari, Keith A. Wharton, Jinsong Qiu, and Thomas W. Bauer. Digital Whole Slide Imaging Compared With Light Microscopy for Primary Diagnosis in Surgical Pathology. *Archives of pathology & laboratory medicine*, 144 (10):1245–1253, 2020. ISSN 1543-2165. doi: 10.5858/arpa.2019-0569-OA.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

Miao Cui and David Y. Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021. ISSN 0023-6837. doi: 10.1038/s41374-020-00514-0.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

Neofytos Dimitriou, Ognjen Arandjelović, and Peter D. Caie. Deep Learning for Whole Slide Image Analysis: An Overview. *Frontiers in Medicine*, 6, 2019. ISSN 2296-858X.

DPA. Digital Pathology Association - Glossary of Terms, July 2022.

Wanderson Gonçalves e. Gonçalves, Marcelo Henrique de Paula Dos Santos, Fábio Manoel França Lobato, Ândrea Ribeiro-dos-Santos, and Gilderlanio Santana de Araújo. Deep learning in gastric tissue diseases: A systematic review. *BMJ Open Gastroenterology*, 7(1):e000371, 2020. ISSN 2054-4774. doi: 10.1136/bmjgast-2019-000371.

Filippo Fraggetta, Vincenzo L'Imperio, David Ameisen, Rita Carvalho, Sabine Leh, Tim-Rasmus Kiehl, Mircea Serbanescu, Daniel Racoceanu, Vincenzo Della Mea, Antonio Polonia, Norman Zerbe, and Catarina Eloy. Best Practice Recommendations for the Implementation of a Digital Pathology Workflow in the Anatomic Pathology Laboratory by the European Society of Digital and Integrative Pathology (ESDIP). *Diagnostics*, 11(11):2167, 2021. ISSN 2075-4418. doi: 10.3390/diagnostics11112167.

Michael M. Franklin, Fred A. Schultz, Marissa A. Tafoya, Audra A. Kerwin, Cory J. Broehm, Edgar G. Fischer, Rama R. Gullapalli, Douglas P. Clark, Joshua A. Hanson, and David R. Martin. A Deep Learning Convolutional Neural Network Can Differentiate Between Helicobacter Pylori Gastritis and Autoimmune Gastritis With Results Comparable to Gastrointestinal Pathologists. *Archives of Pathology & Laboratory Medicine*, 146(1):117–122, April 2021. ISSN 0003-9985. doi: 10.5858/arpa.2020-0520-OA.

Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, 4(1):27, January 2013. ISSN 2153-3539. doi: 10.4103/2153-3539.119005.

Ian Goodfellow, Aaron Courville, and Yoshua Bengio. *Deep Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts, 2016. ISBN 0-262-33737-1.

Jon Griffin and Darren Treanor. Digital pathology in clinical use: Where are we now and what is holding us back? *Histopathology*, 70(1):134–145, 2017. doi: 10.1111/his.12993.

Pedro Guimarães, Andreas Keller, Tobias Fehlmann, Frank Lammert, and Markus Casper. Deep-learning based detection of gastric precancerous conditions. *Gut*, 69(1):4–6, 2020. ISSN 1468-3288. doi: 10.1136/gutjnl-2019-319347.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, December 2015. doi: Tech.

Tom Hempel. Development of a dataset and AI-based proof-of-concept algorithm for the classification of digitized whole slide images of gastric tissue, September 2023.

Andreas Holzinger, Randy Goebel, Michael Mengel, and Heimo Müller. *Artificial Intelligence and Machine Learning for Digital Pathology*, volume 12090. Springer International Publishing, Cham, 2020. ISBN 978-3-030-50401-4. doi: 10.1007/978-3-030-50402-1.

Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. (arXiv:1504.07947), March 2016. doi: 10.48550/arXiv.1504.07947.

Osamu Iizuka, Fahdi Kanavati, Kei Kato, Michael Rambeau, Koji Arihiro, and Masayuki Tsuneki. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Scientific reports*, 10(1):1504, 2020. doi: 10.1038/s41598-020-58467-9.

Takumi Itoh, Hiroshi Kawahira, Hirotaka Nakashima, and Noriko Yata. Deep learning analyzes Helicobacter pylori infection by upper gastrointestinal endoscopy images. *Endoscopy International Open*, 6(2):E139–E144, 2018. ISSN 2196-9736. doi: 10.1055/s-0043-120830.

Stephan W. Jahn, Markus Plass, and Farid Moinfar. Digital Pathology: Advantages, Limitations and Emerging Perspectives. *Journal of Clinical Medicine*, 9(11), 2020. doi: 10.3390/jcm9113697.

Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, September 2021. ISSN 1422-8890. doi: 10.1007/s12525-021-00475-2.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking Self-Supervised Learning on Diverse Pathology Datasets. (arXiv:2212.04690), April 2023. doi: 10.48550/arXiv.2212.04690.

Neel Kanwal, Fernando Pérez-Bueno, Arne Schmidt, Kjersti Engan, and Rafael Molina. The Devil is in the Details: Whole Slide Image Acquisition and Processing for Artifacts Detection, Color Variation, and Data Augmentation: A Review. *IEEE Access*, 10:58821–58844, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3176091.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. (arXiv:1412.6980), January 2017. doi: 10.48550/arXiv.1412.6980.

Sebastian Klein, Jacob Gildenblat, Michaele Angelika Ihle, Sabine Merkelbach-Bruse, Ka-Won Noh, Martin Peifer, Alexander Quaas, and Reinhard Büttner. Deep learning for sensitive detection of Helicobacter Pylori in gastric biopsies. *BMC gastroenterology*, 20(1):417, 2020. doi: 10.1186/s12876-020-01494-7.

Daisuke Komura and Shumpei Ishikawa. Machine Learning Methods for Histopathological Image Analysis. *Computational and structural biotechnology journal*, 16: 34–42, 2018. ISSN 2001-0370. doi: 10.1016/j.csbj.2018.01.001.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Vinay Kumar, Abul K. Abbas, and Jon C. Aster, editors. *Robbins Basic Pathology.* Elsevier, Philadelphia Pennsylvania, tenth edition edition, 2018. ISBN 978-0-323-35317-5.

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006-341X. doi: 10.2307/2529310.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning. pages 14318–14328, 2021.

Zongyao Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Chronic gastritis classification using gastric X-ray images with a semi-supervised learning method based on tri-training. *Medical & Biological Engineering & Computing*, 58(6): 1239–1250, 2020. ISSN 1741-0444. doi: 10.1007/s11517-020-02159-z.

Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, October 2016. ISSN 1361-8415. 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(3):370–378, 2020. ISSN 1543-2165. doi: 10.5858/arpa.2019-0004-OA.

Stacey E. Mills, editor. *Histology for Pathologists.* Wolters Kluwer/Lippincott Williams & Wilkins Health, Philadelphia, fourth edition edition, 2012. ISBN 978-1-4511-1303-7.

Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media, Inc, Sebastopol, CA, first edition edition, 2016. ISBN 978-1-4493-6941-5.

Muhammad Khalid Khan Niazi, Anil V. Parwani, and Metin N. Gurcan. Digital pathology and artificial intelligence. *The Lancet. Oncology*, 20(5):e253–e261, 2019. doi: 10.1016/S1470-2045(19)30154-8.

Jeffrey J. Nirschl, Andrew Janowczyk, Eliot G. Peyster, Renee Frank, Kenneth B. Margulies, Michael D. Feldman, and Anant Madabhushi. Chapter 8 - Deep Learning Tissue Segmentation in Cardiac Histopathology Images. In S. Kevin Zhou, Hayit Greenspan, and Dinggang Shen, editors, *Deep Learning for Medical Image Analysis*, pages 179–195. Academic Press, January 2017. ISBN 978-0-12-810408-8. doi: 10.1016/B978-0-12-810408-8.00011-0.

Liron Pantanowitz, Navid Farahani, and Anil Parwani. Whole slide imaging in pathology: Advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, page 23, 2015. doi: 10.2147/PLMI.S59826.

Liron Pantanowitz, Ashish Sharma, Alexis B. Carter, Tahsin Kurc, Alan Sussman, and Joel Saltz. Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives. *Journal of pathology informatics*, 9:40, 2018. ISSN 2229-5089. doi: 10.4103/jpi.jpi6918.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. (arXiv:1912.01703), December 2019. doi: 10.48550/arXiv.1912.01703.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu

Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928.

João Pedro Mazuco Rodriguez, Rubens Rodriguez, Vitor Werneck Krauss Silva, Felipe Campos Kitamura, Gustavo Cesar Antônio Corradi, Ana Carolina Bertoletti de Marchi, and Rafael Rieder. Artificial intelligence as a tool for diagnosis in digital pathology whole slide images: A systematic review. *Journal of pathology informatics*, 13:100138, 2022. ISSN 2229-5089. doi: 10.1016/j.jpi.2022.100138.

Denis Rothman. *Artificial Intelligence by Example: Develop Machine Intelligence from Scratch Using Real Artificial Intelligence Use Cases*. Packt Publishing, Birmingham, UK, 2018. ISBN 978-1-78899-002-8.

Arnout C. Ruifrok, Ruth L. Katz, and Dennis A. Johnston. Comparison of Quantification of Histochemical Staining By Hue-Saturation-Intensity (HSI) Transformation and Color-Deconvolution. *Applied Immunohistochemistry & Molecular Morphology*, 11(1):85, March 2003. ISSN 1541-2016.

Barik A. Salih. Helicobacter pylori infection in developing countries: The burden for how long? *Saudi Journal of Gastroenterology : Official Journal of the Saudi Gastroenterology Association*, 15(3):201–207, 2009. doi: 10.4103/1319-3767.54743.

Ahmed Serag, Adrian Ion-Margineanu, Hammad Qureshi, Ryan McMillan, Marie-Judith Saint Martin, Jim Diamond, Paul O'Reilly, and Peter Hamilton. Translational AI and Deep Learning in Diagnostic Pathology. *Frontiers in Medicine*, 6: 185, 2019. ISSN 2296-858X. doi: 10.3389/fmed.2019.00185.

Pentti Sipponen and Heidi-Ingrid Maaroos. Chronic gastritis. *Scandinavian Journal of Gastroenterology*, 50(6):657–667, 2015. ISSN 0036-5521. 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*, 21 (18):6652, September 2020. ISSN 1422-0067. doi: 10.3390/ijms21186652.

Carlos Thomas and C.-P. Adler, editors. *Histopathologie: Lehrbuch Und Atlas Zur Allgemeinen Und Speziellen Pathologie*. Pathologie / Hrsg. von C. Thomas. Schattauer, Stuttgart, 13., durchges. und korrigierte aufl. edition, 2001. ISBN 3-7945-2120-X.

Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial Intelligence and Digital Pathology: Challenges and Opportunities. *Journal of pathology informatics*, 9:38, 2018. ISSN 2229-5089. doi: 10.4103/jpi.jpi5318.

Mart van Rijthoven, Maschenka Balkenhol, Karina Siliņa, Jeroen van der Laak, and Francesco Ciompi. HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical Image Analysis*, 68:101890, February 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101890.

Yun Wang, Juncheng Li, and Florian Metze. A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35, October 2018.

Richard Whitehead, editor. *Gastrointestinal and Oesophageal Pathology.* Churchill Livingstone, Edinburgh, 2nd ed. edition, 1995. ISBN 978-0-443-04764-0.

Farhad Ghazvinian Zanjani, Svitlana Zinger, Babak Ehteshami Bejnordi, Jeroen A W M van der Laak, and Peter H. N. de With. Stain normalization of histopathology images using generative adversarial networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 573–577, April 2018. doi: 10.1109/ISBI.2018.8363641.

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, 14.12.2023

Place, Date

P.

Signature