Developing a comprehensive dataset and baseline model for classification and generalizability testing of gastric whole slide images in computational pathology

Master Thesis

Master of Science in Information Systems

Dominic Harold Liebel

March 27, 2025

Supervisor:

1st: Prof. Dr. Christian Ledig

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

Abstract

Despite advances in computational pathology, a critical limitation persists: the lack of comprehensive, multi-scanner datasets for gastric histopathology that enable robust evaluation of model generalizability. This master's thesis addresses this gap by developing the Gastric Slide Database (GSDB), a meticulously annotated dataset comprising 360 whole slide images from 274 patients from two scanner generations (Pannoramic MIDI I/II). Expert pathologists provided annotations for anatomical regions (corpus/antrum/intermediate) and inflammation status (non-inflamed/inflamed), with rigorous quality control protocols leading to 252 patients with validly annotated slides. Our systematic preprocessing pipeline converts proprietary MRXS files into 59,612 analysis-ready 256×256 pixel tiles within a hierarchical organization (slides → clusters → particles → tiles), yielding 3,803 particles (44.8% corpus, 46.8% antrum, 8.4% other) from 338 slides after validation (58.6% inflamed, 38.2% non-inflamed, 3.2% other). We implement and evaluate twelve deep learning models, demonstrating that they achieve within-scanner AUC test performances ranging from 91.15% to 93.61% particle-level AUC for tissue classification and 95.18% to 99.56% slide-level AUC for inflammation classification. Our comprehensive analyses demonstrate that combining pathology-specific corruptions with stain color jittering achieves the most balanced performance across classification tasks and scanner environments while revealing that tissue morphology features transfer effectively across scanners. In contrast, inflammation detection suffers significant degradation on new scanners. A key methodological contribution is our systematic evaluation of aggregation strategies, demonstrating that optimizing tile-to-particle and tile-to-slide prediction aggregation can substantially improve performance and cross-scanner robustness. The GSDB's multi-scanner design enables critical investigations into generalizability challenges, enabling scanner-agnostic diagnostics for clinical deployment.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AUC	Area Under the Receiver Operating Characteristic Curve
GSDB	Gastric Slide Database
HE	Hematoxylin and Eosin
MG	Modified Giemsa
MTA	Medical Technical Assistants
PAS	Periodic acid-Schiff
Slide	Whole Slide Image
WSI	Whole Slide Image

1 Introduction

1.1 Motivation of the Thesis

In recent years, computational pathology has gained significant traction, especially with the advent of AI-based models for medical image classification (Cui and Zhang, 2021). While much progress has been made in cancer diagnostics, a gap in the automated classification of other diseases such as gastric conditions, remains. Gastritis, potentially affecting over half of the world's population (Sipponen and Maaroos, 2015), poses a considerable challenge in pathology, in terms of diagnosis and treatment decisions. The growing workload on pathologists requires more efficient solutions. As WSIs (whole slide images) become integral to pathology workflows, developing reliable AI models to classify these images can enhance diagnostic precision and reduce time spent on repetitive tasks.

Building upon the foundation laid by previous works, particularly Hempel's (2023) Bachelor's thesis on dataset development and initial classification models and Höfling's (2023) Master's thesis on improving classification algorithms, this thesis seeks to push forward the classification of gastric tissue, focusing on distinguishing anatomical regions and detecting inflammation in WSIs. This work also aims to refine data preprocessing, tile-based classification approaches, and model generalization issues, as identified in previous research.

1.2 Objectives of the Thesis

The primary objective of this thesis is to develop a comprehensive dataset and baseline model for the classification of gastric tissue in WSIs. The focus is on two key classifications: the anatomical regions of the stomach (antrum and corpus) and the presence of inflammation. Leveraging deep learning models aims to build upon the prior proof-of-concept models and achieve greater generalizability and aggregated test accuracies. This thesis will explore:

- Creation of a robust dataset comprising high-resolution WSI tiles annotated for both anatomical regions and inflammation status.
- Development and tuning of deep learning models, including experimentation with architectures, hyperparameters, and threshold strategies.
- Improvement of tile-based classification.
- Addressing cross-scanner generalizability.
- Ensuring the reproducibility of experiments and developing a baseline that can be extended in future research.

2 Background

2.1 Prior Research Foundation

The methodologies and approaches presented in this work are built upon a foundation of prior research in gastric histopathology analysis. Significant contributions to this field have been made through systematic investigations at the Otto-Friedrich-University Bamberg in collaboration with clinical experts at Klinikum Nürnberg. The initial digitization protocols, annotation frameworks, and quality control measures established by Hempel (2023) provided the essential infrastructure for whole-slide image processing in gastric tissue analysis. Subsequently, Höfling (2023) advanced these methodologies with refined neural network architectures, targeted data augmentation strategies, and systematic hyperparameter optimization techniques specific to gastric tissue classification.

The technical details of these previous investigations, including their architectural innovations and methodological contributions, are thoroughly examined in Section 3.1. The present thesis systematically addresses specific limitations identified in prior works, particularly focusing on looking into cross-scanner generalizability, improving dataset quality, and increasing the accuracy of deep learning models across two different scanners.

2.2 Whole Slide Imaging in Computational Pathology

Whole slide imaging has transformed the field of pathology by enabling the digitization of entire tissue slides at high resolution, mirroring the impact digital imaging had on radiology over the past three decades (Griffin and Treanor, 2017). This digitization has led to significant advancements in computational pathology, where researchers apply machine learning models to analyze and classify histopathological images for tasks such as tumor detection, tissue morphology analysis, and disease diagnosis (Aeffner et al., 2019). The integration of WSI into clinical workflows promises to enhance diagnostic accuracy, reproducibility, and efficiency across multiple disease domains.

However, whole slide imaging presents unique computational challenges due to the extreme resolution of digitized slides, which often reach dimensions of hundreds of thousands of pixels. These WSI require specialized approaches for efficient processing (Koohbanani et al., 2021). Analyzing such large images typically requires breaking them down into smaller tiles, enabling efficient memory management during both the training of deep learning models and inference. Furthermore, the size of WSI datasets leads to substantial demands for data storage infrastructure and computational resources, creating barriers for implementation in resource-constrained settings.

In gastric pathology specifically, whole slide imaging facilitates critical diagnostic tasks such as distinguishing between different anatomical regions (corpus and antrum) and identifying inflammatory conditions. These distinctions are essential for accurate diagnosis of conditions including gastritis, gastric cancer, and other gastroenterological diseases (Rugge et al., 2011; Banks et al., 2019). The ability to digitally analyze tissue characteristics enhances consistency in classification and enables more quantitative assessment of disease features, potentially improving diagnostic accuracy and treatment planning (Obuchowicz et al., 2024).

The development of robust computational models for WSI analysis depends fundamentally on comprehensive, well-annotated datasets. Creating such datasets involves multiple preprocessing steps, including converting proprietary formats (such as MRXS) into standard formats, generating tiles at appropriate magnifications, and implementing downsampling strategies for classification tasks. These preprocessing pipelines must address several critical challenges, including ensuring dataset consistency across multiple scanners, managing class imbalance in pathology data, and developing models that generalize across institutions and patient populations. The quality and breadth of training data ultimately determine the clinical utility of computational pathology systems in gastric tissue analysis and beyond (Hu et al., 2022).

3 Related Work

This section examines relevant literature and previous research efforts in computational pathology, particularly emphasizing methodologies applicable to gastric tissue analysis. We begin by analyzing foundational work that preceded this study, followed by an examination of public benchmark datasets and state-of-the-art approaches in histopathological image analysis.

3.1 Previous Work in the Research Group

This work builds upon two previous theses (Hempel, 2023; Höfling, 2023) conducted at the Chair of Explainable Machine Learning at Otto-Friedrich-University Bamberg in collaboration with the Klinikum Nürnberg.

3.1.1 Initial Dataset Development and Proof-of-Concept Models

The foundational work by Hempel (2023) established one of the first comprehensive frameworks for gastric tissue classification using digital pathology, addressing a significant gap in computational pathology research. Prior to this work, most digital pathology research focused on cancer detection (Hu et al., 2021; Sakamoto et al., 2020; Bera et al., 2019; Ehteshami Bejnordi et al., 2017; Nero et al., 2022), with limited attention to gastric tissue classification despite its clinical relevance. In collaboration with expert pathologists at the Klinikum Nürnberg, Hempel (2023) developed a carefully curated dataset of digitized slides specifically designed for gastric tissue analysis. This dataset comprised 270 slides digitized at 20x magnification using a Pannoramic MIDI I scanner (3DHISTECH Ltd., Budapest, Hungary), representing a diverse spectrum of tissue types and pathological conditions essential for robust model development.

A significant contribution of this work was the systematic annotation protocol developed through iterative refinement with senior pathologists. This protocol delineated anatomical regions (antrum, corpus, intermediate) and inflammatory states (inflamed/non-inflamed), with clearly defined criteria for each classification category. The annotation methodology incorporated robust quality control measures, including multi-reviewer validation and standardized classification criteria (Hempel, 2023). This meticulous approach to annotation provided a reproducible foundation for subsequent machine learning applications and addressed a key challenge in developing reliable ground truth for algorithm training.

The preprocessing pipeline introduced several methodological innovations for efficient slide analysis. Raw slides were systematically processed using a tile-based approach, extracting 256×256 pixel segments with 64-pixel overlap to maintain spatial context while managing the computational challenges posed by gigapixel-sized WSIs. Unlike many studies that use non-overlapping tiles, this approach preserved important contextual information at tile boundaries. A comprehensive quality assessment framework filtered artifacts and non-informative regions through automated tissue content validation (threshold >10% tissue content per tile). The final preprocessed dataset contained approximately 34,000 annotated tiles across both classification tasks, representing one of the larger annotated datasets for gastric tissue analysis.

Initial deep learning models demonstrated promising results in both classification tasks. The tissue classification model achieved a tile-level area under the receiver operating characteristic curve (AUC) of 0.95 on the test set, while inflammation detection reached a tile-level AUC of 0.98. Perfect classification accuracy was observed at the slide-level for inflammation detection, though as the authors noted, these results warrant careful interpretation given the limited test set size (n=20 slides). These initial results, while preliminary, suggested the viability of automated gastric tissue classification and laid the groundwork for more sophisticated approaches. However, the study also identified several limitations requiring further research, including larger test sets, more diverse training data, and improved handling of intermediate tissue types and subtle inflammation patterns.

3.1.2 Advanced Model Development and Dataset Refinement

Building upon the initial dataset framework, Höfling (2023) introduced substantial methodological improvements in both model architecture and evaluation approaches. This work focused on enhancing the robustness and clinical applicability of automated gastric tissue classification systems.

The study applied a systematic approach to address class imbalance, which was particularly pronounced in the inflammation detection task (62.67% inflamed vs. 37.33% non-inflamed in the training set). A combination of weighted loss functions and strategic data augmentation techniques was employed, including rotation

(90°, 180°, 270°), random flips, and controlled color perturbations to improve model generalization.

Architectural innovations were introduced by implementing two distinct ResNet18-based models, each optimized for their respective classification tasks. The tissue classification model incorporated a novel probabilistic averaging approach for tile aggregation, achieving a validation accuracy of 87.80% while maintaining computational efficiency. The inflammation classification model demonstrated improved performance by integrating cross-entropy loss with class-specific weights, reaching a validation accuracy of 92.41%.

Performance metrics demonstrated that the tissue classification model achieved consistent performance across different anatomical regions, with a tile-level F1 score of 90.11% on the test-set. For inflammation status classification, the model reached an accuracy of 72.73% at the tile-level on the test-set and maintained high consistency in slide-level predictions (94.74% accuracy) on the test-set.

The study introduced a systematic approach to hyperparameter optimization, employing grid search across learning rates (10^{-4} to 10^{-2}) and batch sizes (16 to 128). This methodical optimization resulted in more stable training dynamics and improved convergence characteristics compared to previous implementations.

Despite advancements laying crucial groundwork for clinical deployment considerations, several challenges remained unaddressed. In particular, the need for multi-scanner validation and larger-scale clinical validation were identified as critical areas for future investigation. The work also highlighted the importance of developing more sophisticated approaches to handling ambiguous cases and intermediate tissue types.

3.2 Public Pathology Benchmark Datasets

The availability of public benchmark datasets has significantly advanced the development of robust deep learning models for histopathology analysis. These datasets facilitate fair comparisons between algorithms and accelerate progress in computational pathology by enabling researchers to build upon prior work. This Section reviews key benchmark datasets that have influenced the field, particularly those relevant to gastric histopathology.

3.2.1 CAMELYON Challenges

The CAMELYON challenges represent pivotal efforts in standardizing the evaluation of deep learning algorithms for computational pathology. These challenges have established rigorous benchmarks for metastasis detection in histological images and have significantly advanced the field.

CAMELYON16 The CAMELYON16 challenge (CAncer MEtastases in LYmph nOdes challeNge) focused on the automated detection of breast cancer metastases in WSIs of lymph nodes (Ehteshami Bejnordi et al., 2017). This landmark challenge provided 399 WSIs (270 training, 129 testing) from two medical centers in the Netherlands, with exhaustive pixel-level annotations of metastatic regions.

The challenge demonstrated that deep learning algorithms could achieve performance comparable to or exceeding that of expert pathologists in this specific task. The winning algorithm achieved an AUC of 0.994 for slide-level classification, surpassing the average performance of a panel of 11 pathologists operating under time constraints (mean AUC of 0.810) and approaching that of a pathologist with unlimited time (AUC of 0.966) (Ehteshami Bejnordi et al., 2017).

Key methodological insights from CAMELYON16 included the importance of patch-based analysis for processing gigapixel whole-slide images, effective handling of color variation between scanning centers, and strategies for reducing false positives. The challenge established that convolutional neural networks, particularly those with residual connections, were particularly effective for histopathological image analysis.

CAMELYON17 CAMELYON17 extended the scope of its predecessor by addressing more clinically relevant tasks (Bandi et al., 2019; Litjens et al., 2018). While CAMELYON16 focused on the binary classification of individual slides, CAMELYON17 aimed to classify the pathologic lymph node status (pN-stage) at the patient level, which more closely mirrors clinical practice. The challenge introduced several important advances:

- **Multi-slide analysis:** Each patient case included multiple WSIs, requiring algorithms to integrate information across slides.
- **Broader classification scope:** The challenge included isolated tumor cells (ITCs), which were excluded from CAMELYON16, creating a more complete clinical scenario.
- **Increased diversity:** Data from five medical centers (versus two in CAMELYON16) better represented real-world slide preparation and scanning protocols variation.
- **Expanded dataset:** The dataset size increased to 1,000 WSIs, enabling more robust model development and evaluation.

The results of CAMELYON17 revealed both progress and persistent challenges. The best-performing algorithm achieved a quadratic-weighted kappa score of 0.8993, indicating substantial agreement with pathologist assessments. However, all algorithms struggled with reliably identifying isolated tumor cells, with detection rates below 40%. Simple combinations of the top algorithms resulted in better performance than any individual system (Bandi et al., 2019), suggesting the value of ensemble approaches in this domain.

The CAMELYON challenges have established benchmark methodologies and datasets that continue to influence computational pathology research, particularly in tumor detection, slide-level classification, and integration of multi-slide information.

3.2.2 Prov-GigaPath: A Foundation Model for Digital Pathology

A significant advancement in computational pathology was recently introduced by Xu et al. (2024) with Prov-GigaPath, a foundation model pre-trained on an unprecedented scale of histopathological data. The model was trained on 1.3 billion 256×256 pixel image tiles derived from 171,189 WSIs, spanning 31 major tissue types from over 30,000 patients. This extensive dataset, Prov-Path, is one of the largest histopathology datasets to date.

The model consists of two main components:

- A tile encoder that processes individual 256×256 pixel tiles to extract local features.
- A slide encoder based on the LongNet architecture that aggregates information across the entire slide, capable of handling up to 70,121 tiles per slide.

Prov-GigaPath represents a breakthrough in foundation model development for computational pathology, with several key innovations:

- **Scale and coverage:** The unprecedented scale of training data offers comprehensive coverage of histopathological variations across multiple cancer types and tissue sites.
- **Multi-scale integration:** The model effectively integrates information from local cellular patterns to global tissue architecture.
- **Transfer learning capabilities:** Pre-trained representations demonstrate strong transfer learning potential for downstream tasks with limited labeled data.
- **Cross-domain generalization:** Strong performance across diverse tissue types and scanning protocols suggests robust generalization capabilities.

Notably, the model achieved state-of-the-art performance across 25 out of 26 evaluation tasks, including cancer subtyping and mutation prediction. For pathology tasks similar to our gastric tissue classification, Prov-GigaPath demonstrated significant improvements over previous methods, particularly in capturing complex tissue patterns and their spatial relationships.

The introduction of Prov-GigaPath represents a paradigm shift similar to what occurred in natural language processing with large language models, suggesting that foundation models trained on massive histopathology datasets can provide powerful representations that generalize across a wide range of downstream tasks. This approach is particularly promising for areas such as gastric histopathology, where labeled data is limited but the patterns have commonalities with other epithelial tissues Xu et al. (2024).

3.2.3 GasHisSDB: A Public Benchmark Dataset

GasHisSDB (Hu et al., 2021), a pioneering public dataset for gastric histopathology image classification, influenced the development of our dataset and methodology. GasHisSDB demonstrated several key principles that informed our approach:

Dataset Organization: GasHisSDB introduced a hierarchical approach to organizing histopathological data, structuring 245,196 sub-size images into multiple resolutions (160×160, 120×120, and 80×80 pixels). While we adopted a similar hierarchical organization, our work extends this concept by using a fixed tile size of 256×256 pixels to better balance resolution and computational efficiency. We further enhanced this approach by implementing a 64-pixel overlap between adjacent tiles to maintain spatial context and developing a more stringent tissue content validation process requiring >90% tissue content compared to GasHisSDB's less strict criteria.

Classification Tasks: GasHisSDB focused on binary classification between normal and abnormal tissue samples. Our work builds upon this foundation by expanding to specific anatomical classification (antrum vs. corpus), adding dedicated inflammation detection capabilities, and implementing a more granular annotation system at the particle-level that enables hierarchical prediction aggregation.

Model Evaluation: The evaluation framework in GasHisSDB demonstrated the importance of comprehensive model assessment. Their work showed performance variations across different architectures, with traditional machine learning methods achieving 86.08% accuracy, deep learning approaches reaching 96.47% accuracy, and Vision Transformers showing promise despite requiring longer training periods. Our work extends this evaluation framework by systematically comparing multiple modern architectures and assessing cross-scanner generalization capabilities.

These findings influenced our model selection and evaluation strategy, particularly our focus on state-of-the-art architectures like ConvNeXt, implementation of robust cross-validation procedures, and scanner-specific evaluations to assess real-world deployment challenges.

3.2.4 HMU-GC-HE-30K: A Histological Dataset for Gastric Cancer

Another significant recent contribution to the field is the HMU-GC-HE-30K dataset by Lou et al. (2025), which provides a comprehensive collection of gastric cancer histological images with detailed tumor microenvironment (TME) annotations. This dataset addresses a critical gap in gastric cancer research, as the TME plays a crucial role in disease progression and treatment response.

The dataset comprises nearly 31,000 histological image patches extracted from 300 WSIs of gastric cancer patients. A distinguishing feature of this dataset is its comprehensive annotation of eight distinct tissue classes within the tumor microenvironment:

- Adipose tissue (ADI)
- Debris (DEB)
- Lymphocyte aggregates (LYM)
- Mucus (MUC)
- Muscle (MUS)
- Normal mucosa (NOR)
- Stroma (STR)
- Tumor epithelium (TUM)

This granular annotation approach provides significant advantages over previous datasets like GasHisSDB, which offered only binary classification (normal/abnormal). The detailed TME classification enables a more sophisticated analysis of histological patterns and their relationship to clinical outcomes.

An additional strength of HMU-GC-HE-30K is the inclusion of comprehensive clinical data for each patient, including demographic information, TNM staging, histological type, Lauren classification, and various invasion parameters. This histological and clinical data integration facilitates research into clinically relevant biomarkers and prognostic factors.

The dataset has been validated through the development of two deep learning models: a Transformer-based architecture (ViT) and a CNN-based model (EfficientNet). Both achieving high classification performance (AUC of 0.94 and 0.96, respectively) across the eight tissue classes. This demonstrates the utility of the dataset for developing and evaluating computational pathology algorithms.

3.3 Machine Learning Methods in Medical Imaging

While benchmark datasets provide essential resources for model development and evaluation, selecting appropriate machine learning methodologies is equally crucial for advancing computational pathology applications. The evolution of these methods has significantly influenced how histopathological images are analyzed and interpreted.

A comprehensive review by Husain et al. (2024) systematically analyzed the application of different machine learning algorithms across diverse medical imaging contexts, highlighting several key findings directly relevant to computational pathology. The review emphasized the importance of comprehensive evaluation using multiple metrics rather than relying on accuracy alone, which can be particularly misleading in scenarios with imbalanced datasets, a common challenge in pathology where certain tissue types or pathological conditions may be underrepresented. The combination of accuracy with AUC was found to provide a more complete assessment of model performance, allowing for better evaluation of sensitivity-specificity tradeoffs that are critical in medical diagnostic contexts.

The effectiveness of different ML architectures was found to vary significantly based on the imaging modality and the specific medical condition being analyzed. For histopathology specifically, convolutional neural networks (CNNs) consistently demonstrated superior performance, achieving 95-99% accuracy in various tissue classification tasks. Among CNN variants, residual networks (ResNets) and densely connected networks (DenseNets) showed particular promise due to their ability to learn hierarchical features while mitigating the vanishing gradient problem in deep architectures.

Beyond architecture selection, the review highlighted the importance of appropriate preprocessing steps and data augmentation techniques specific to histopathological images. Color normalization, in particular, was identified as a critical preprocessing step for handling stain variation across slides from different laboratories or scanning equipment. Similarly, carefully selected data augmentation strategies were shown to effectively address the challenge of limited training data, with rotation and color perturbation proving especially valuable for histopathology applications.

These insights into machine learning methods provide a robust theoretical foundation for our work in gastric tissue classification, informing our choices of model architecture, evaluation metrics, and preprocessing approaches.

3.4 WSI Preprocessing Methods

Effective preprocessing of WSIs constitutes a critical foundation for computational pathology applications. These images' large dimensions necessitate specialized approaches to transform raw data into analyzable units while preserving diagnostic information.

3.4.1 The Tiling Approach for WSI Processing

The tiling approach has emerged as the predominant method for managing computational challenges posed by WSIs (Xu et al., 2024). Nahhas et al. (2023) established a systematic protocol for preprocessing WSIs into smaller image patches optimized for deep-learning workflows. Their methodology first converts WSIs from proprietary formats (SVS, MRXS, NDPI) to standardized formats, followed by segmentation into tiles of consistent dimensions. This crucial step enables the analysis of computationally prohibitive gigapixel images while preserving diagnostically relevant information.

The selection of optimal tile dimensions represents a critical balance between contextual information preservation and computational feasibility. While common dimensions range from 256×256 to 512×512 pixels, the choice depends on the specific diagnostic task, available computational resources, and required analytical resolution. Standardizing the tiling process across datasets ensures consistency in downstream analysis.

In our implementation, as detailed in Section 4.4, we extract 256×256 pixel patches using a protocol that includes 64-pixel overlap to preserve contextual continuity at tile boundaries. This dimension choice builds upon established methodologies from previous research Höfling (2023); Hempel (2023).

3.4.2 Color Normalization and Artifact Management

Beyond tiling, stain normalization has been recognized as a fundamental preprocessing step. Mahbod et al. (2024) demonstrated that structure-preserving color normalization methods substantially improve cross-dataset generalization by transforming color distributions of source images to match reference standards while preserving underlying tissue structure. Their research showed that non-deterministic training with normalization followed by deterministic testing significantly enhanced model generalization across diverse datasets.

Tellez et al. (2018) developed a novel approach for stain normalization by directly modifying H&E color channels to create diverse and realistic stain variations. Their data augmentation strategy and network ensembling resulted in stain-invariant convolutional networks for mitosis detection in breast histology. This work demonstrated that specialized augmentation strategies can significantly reduce generalization error when transferring models across different centers without requiring multicenter training cohorts or conventional stain standardization algorithms.

Di Salvo et al. (2024) established MedMNIST-C, a comprehensive benchmark spanning 12 datasets and nine imaging modalities, to evaluate model robustness against realistic image corruptions. Their research confirmed that domain-specific data augmentation targeting modality-specific artifacts (such as stain variations, blur, and noise) consistently outperforms generic augmentation methods. This work highlights the importance of designing targeted preprocessing strategies that account for the unique characteristics of histopathological images. We explore similar domain-specific augmentation approaches in our research, as detailed in Section 10.4.

Artifact detection and filtering constitute another critical preprocessing component. Our preprocessing pipeline employs quality assessment through tissue content validation, excluding regions with insufficient diagnostic content. The effectiveness of such quality control measures is supported by findings from Tellez et al. (2018), who emphasized the importance of filtering artifacts when building robust histopathological image analysis systems.

The careful implementation of these preprocessing methods transforms raw histopathological data into a standardized format suitable for computational analysis while addressing the unique technical challenges inherent in whole slide imaging. Establishing a robust preprocessing framework ultimately enhances the reliability and reproducibility of downstream analytical results.

3.5 Research Gap and Contributions

This review of related work highlights the significant progress made in computational pathology and the remaining opportunities. While public datasets like CAMELYON, GasHisSDB, and HMU-GC-HE-30K have established important benchmarks, and foundation models like Prov-GigaPath have demonstrated the potential of large-scale pretraining, the domain of gastric tissue classification, particularly the distinction of anatomical regions and inflammation states, remains underexplored.

Similarly, while advanced machine learning methods and preprocessing techniques have shown promise in various histopathological applications, their application to gastric tissue analysis has been limited. Most existing work focuses on cancer detection rather than the more routine but clinically important tasks of anatomical region classification and inflammation assessment, which constitute a significant portion of pathologists' daily workload.

Our work addresses these gaps by building upon the initial dataset development and proof-of-concept models from Hempel (2023) and the methodological advancements from Höfling (2023), while incorporating insights from broader computational pathology research. We specifically focus on improving gastric tissue classification through enhanced dataset curation, architectural innovations, and evaluation methodologies designed to address the unique challenges of this domain.

The primary contributions of our work include: (1) an expanded and refined dataset of annotated gastric tissue WSIs, (2) improved deep learning models for anatomical region and inflammation classification, (3) a systematic evaluation of model performance at both tile and slide-levels, and (4) insights into the potential clinical applicability of automated gastric tissue classification systems. These contributions advance the field toward the broader goal of developing practical computational pathology tools to assist pathologists in routine diagnostic tasks, potentially improving efficiency and consistency in gastric tissue assessment.

4 Definitions

This Section introduces the key terminology and concepts used throughout this work, establishing a clear hierarchy from WSIs down to individual tiles.

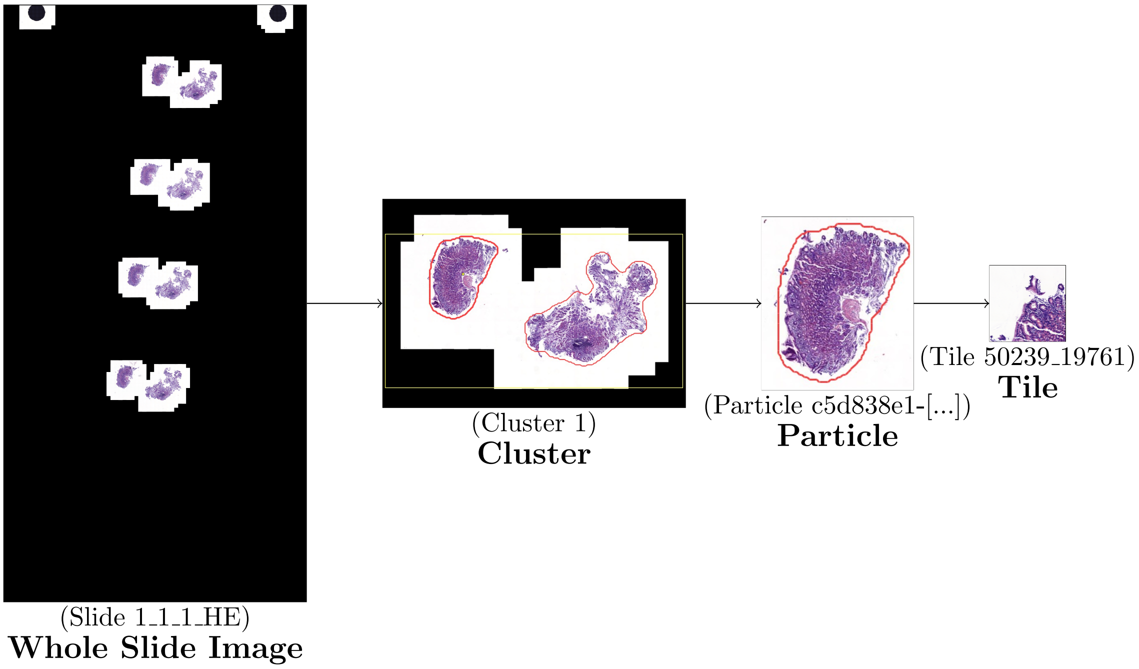


Figure 1: Progression from slide to individual tiles

4.1 Whole Slide Images

WSIs/Slides are high-resolution digital scans of complete histological glass slides. The slides are stored in the proprietary MRXS format with dimensions of 103356 x 203931 pixels. Each slide may contain multiple tissue sections, typically requiring 0.5-4GB of storage space. In previous works (Höfling, 2023; Hempel, 2023), 270 slides were digitized at the Südklinikum Nürnberg using a Pannoramic MIDI I scanner at 20x magnification. In this work, 90 additional slides, digitized at the Südklinikum Nürnberg using a Pannoramic MIDI II scanner at 20x magnification were added to the existing dataset.

Three different staining techniques (HE,PAS,MG) were used in the preparation of the slides, and we will exclusively work with HE-stained slides during this thesis. More information about staining selection is available in Section 6.2.3.

Slides follow a consistent naming scheme:
`[slideID]_[patientID]_[scanner]_[stain].mrxs`, for example `70_50_1_HE.mrxs`. Each MRXS file is accompanied by a directory containing auxiliary data files:

- Multiple `Data[xxxx].dat` files containing image data

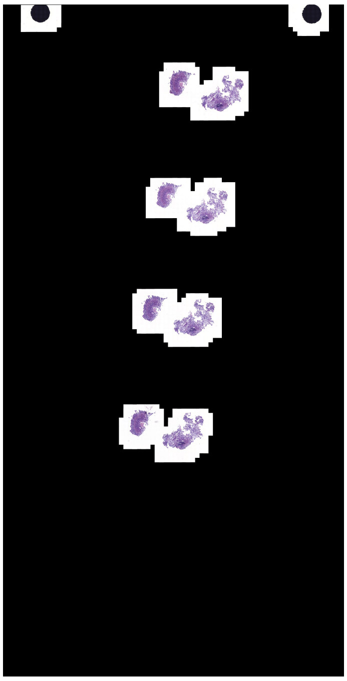


Figure 2: Example slide 1_1_1_HE

- `Index.dat` for indexing information
- `Slidedat.ini` containing slide metadata

This naming scheme represents a significant improvement over the initial non-standardized naming conventions used with the two scanners (detailed in Appendix A.2). The original naming schemes posed challenges for data organization and cross-scanner analysis, necessitating the development of the current unified system.

For brevity and clarity in subsequent sections, the complete identifier string `[slideID]_[patientID]_[scanner]_[stain]` will be denoted as `[Slide_name]` in all references.

4.2 Clusters

Clusters represent rectangular regions (colored yellow in Figure 3) within a slide that group together related regions of interest. They serve multiple purposes:

1. Organization of similar tissue sections within a slide
2. Enabling researchers to use one or multiple clusters per slide in future experiments

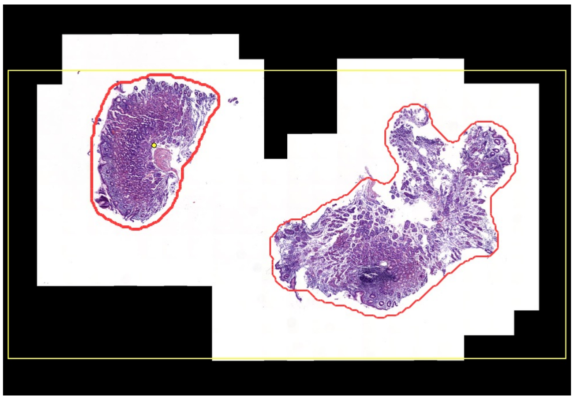


Figure 3: Example cluster 1 of slide 1_1_1_HE

Each cluster is defined by its identifier and bounding box coordinates that specify its spatial extent within the slide. Clusters are stored in JSON format, with files following the naming convention `[slide_name]_clusters.json`. The following example demonstrates the structure of a cluster file:

```
[
{
        "id": 1,
        "bounds": {
                "left": 58477,
                "right": 78791,
                "top": 41783,
                "bottom": 70210
        }
},
{
        "id": 2,
        "bounds": {
                "left": 52704,
                "right": 70740,
                "top": 77511,
                "bottom": 102600
        }
},
[...]
]
```

(Example from 270_244_1_PAS_clusters.json)

4.3 Particles

A particle represents a distinct slice of gastric tissue within a slide. Particles are stored as annotations in GeoJSON format using polygonal regions drawn around areas of interest. They are created either manually or semi-automatically using

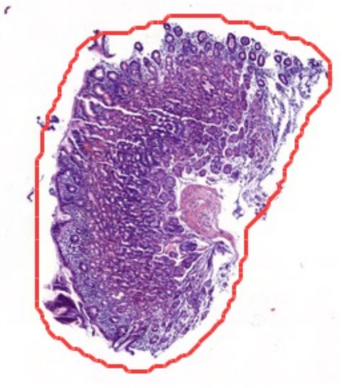


Figure 4: Example particle 5d838e1-8449-4d08-a8f4-7b46a43664b8, located in cluster 1 of slide 1_1_1_HE

specialized annotation software (see Appendix A.4) and stored in files following the naming convention `[slide_name]_annotations.json`.

Each particle is represented as a GeoJSON feature object containing:

- A unique identifier
- Classification metadata including:
  - Anatomical region (antrum, corpus, intermediate, or other)
  - Inflammation status (inflamed, non-inflamed or other)
- An associated cluster ID
- Geometric boundaries defined by polygon coordinates

The following example illustrates the GeoJSON structure for a single particle:

```
 {
        "type": "Feature",
        "id": "c5d838e1-8449-4d08-a8f4-7b46a43664b8",
        "properties": {
                "objectType": "annotation",
                "classification": {
                        "tissue_type": "corpus",
                        "inflammation_status": "inflamed"
                },
                "cluster_id": 1
        },
        "geometry": {
                "type": "Polygon",
                "coordinates": [ [[[53347, 19761], ...]]
        }
}
```

(Example from 1_1_1_HE_annotations.json)

These particle annotations form the ground truth for model training and evaluation, bridging raw slides and the extracted tiles used for training. The geometric boundaries of each particle determine the regions from which tiles are generated.

4.3.1 Valid Annotations

Annotations were classified as valid based on two independent criteria:

Valid Inflammation Status:

- Valid categories: inflamed or noninflamed
  - Slides with any other inflammation category were excluded from inflammation status analysis
- Assessment performed at the slide-level
- The slide's inflammation status applies to all its components (clusters, particles, and tiles)

Valid Tissue Type:

- Valid categories: corpus or antrum
- Assessment performed at the particle-level
- A slide is considered valid if it contains at least one particle annotated as either "corpus" or "antrum"
- Particles annotated as other tissue types (e.g., "intermediate", "other") within a valid slide:
  - inherit the slide's inflammation status
  - are excluded from tissue type analysis

For more information, a detailed example is explained in Section A.5 of the Appendix.

4.4 Tiles

Tiles are fixed-size image patches (256x256 pixels) extracted from annotated particles. The tiling process:

1. Processes annotated particles using geometric containment checks

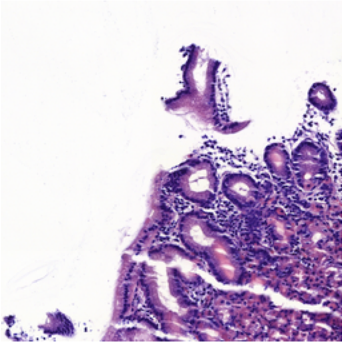


Figure 5: Example tile 50239_19761 within particle 5d838e1-8449-4d08-a8f4-7b46a43664b8, which is within cluster 1 of slide 1_1_1_HE

2. Extracts 2560×2560 pixel regions
3. Downsamples regions by 10× to 256×256 pixels using Lanczos resampling
4. Maintains 64-pixel overlap in downsampled space (640 pixels at native scale)
5. Uses parallel processing (32 workers) with batched coordinate handling

The selection of 10x downsampled 256×256 pixel tiles was determined through consideration of multiple technical and practical factors. This dimension builds upon established methodologies from previous research, specifically the student theses of Höfling (2023) and Hempel (2023). From a computational perspective, the power-of-two dimensions (2^8) provide optimal memory alignment and GPU operations, enhancing processing efficiency. The chosen dimensions also facilitate effective feature extraction through a 64-pixel overlap protocol, ensuring the continuity of histological features across tile boundaries without introducing excessive redundancy in the dataset. The effectiveness of 256×256 pixel tiles has been further validated in large-scale histopathological applications. The Prov-GigaPath project (Xu et al., 2024) successfully processed 1.3 billion tiles of this size, demonstrating the dimension's suitability for large-scale deep learning applications. This empirical validation reinforces our dimensional choice for the current study.

Tiles represent the fundamental input unit for the deep learning models used in this work. Tiles are saved as PNG files with a naming scheme that encodes their origin and location:
`[Slide_name]_particle_[particle_id]_tile_[x]_[y]_size_[width]x[height]`

For example:
`1_1_1_HE_particle_5d838e1-[...]_tile_50239_19761_size_2560x2560.png`

where:

- `[x]` and `[y]` represent the pixel coordinates in the original slide where the tile extraction begins

- These coordinates mark the top-left corner of the extracted region
- Values are in level 0 (highest resolution) coordinates of the WSI
- Example: `tile_50239_19761` indicates the tile was extracted starting at:
  - x = 50239 pixels from the left edge of the WSI
  - y = 19761 pixels from the top edge of the WSI

The implementation utilizes a Cartesian coordinate system where coordinates (x, y) represent pixel distances from the WSI origin (top-left corner). This standardized approach ensures consistency across different scanning platforms and facilitates integration with existing digital pathology workflows. Through this coordinate system, precise spatial relationships between extracted tiles and their source material can be maintained, enabling comprehensive validation and reproducibility of results.

5 Dataset Analysis and Cohort Characteristics

Developing robust deep learning models for computational pathology necessitates high-quality, well-annotated datasets. We present a comprehensive data processing pipeline that systematically transforms slides through multiple hierarchical levels: raw digital slides, annotated tissue regions, classified particles, and analysis-ready tiles.

5.1 Data Acquisition and Preprocessing

Digital slides were acquired from gastric tissue biopsies using Pannoramic MIDI I (Scanner1) and MIDI II (Scanner2) scanners (3DHISTECH Ltd., Budapest, Hungary) at 20x magnification at the Department of Pathology, Klinikum Nürnberg. Following standard pathological practice, all gastric biopsies were stained with a complete set of three staining methods: Hematoxylin and Eosin (HE), Periodic acid–Schiff (PAS), and Modified-Giemsa (MG). Our clinical collaborator (Dr. Braunecker), confirmed that this comprehensive staining protocol is standard practice for complete gastric pathological assessment, as each stain highlights different tissue structures critical for accurate diagnosis.

The two scanner cohorts followed distinct acquisition protocols:

- **Scanner1 Cohort**: Slides were specifically created for this research project from archived patient tissue blocks. These slides were consistently processed by the same medical technical assistant, assigned standardized case numbers, and remain physically available for potential reexamination.

- **Scanner2 Cohort**: Slides were obtained from routine clinical diagnostics during a limited scanner availability window. These slides were temporarily relabeled for anonymized scanning and then returned to clinical use, making them unavailable for rescanning. According to our clinical collaborator, this cohort was deliberately curated to include underrepresented cases (such as non-inflamed antrum particles and intermediate zone mucosa) and exemplary clear cases, potentially introducing a selection bias different from the Scanner1 cohort.

5.2 Expert Annotation Protocol

Expert pathologists provided annotations in two complementary formats to ensure accurate tissue classification. For straightforward cases, written documentation was provided detailing the classification of each tissue particle (e.g., "Slide 20HE: left particle corpus, center antrum, right particle corpus"). Pathologists provided additional visual annotations for more complex slides containing multiple tissue types or ambiguous regions, by marking tissue types directly on slide images, as shown in Figure 6. Image from slide 277_247_2_HE was selected as a representative example containing all possible expert tissue annotation types.

These visual annotations were crucial for capturing spatial relationships and precise tissue boundaries, particularly in slides containing intermediate regions or multiple tissue types close to each other. The annotations used standardized markings:

- C: Corpus
- A: Antrum
- IM: Intermediate
- NK: Non-classifiable regions (German: "nicht klassifizierbar")

Annotations were serialized in a standardized JSON format, incorporating categorical classifications (tissue type, inflammation status) and geometric data (ROI coordinates). This format ensures reproducibility and facilitates programmatic access during model development.

5.3 Automated Tile Generation

The final phase implemented an automated tile extraction pipeline utilizing OpenSlide (Sepulveda and Patil, 2008a) and PIL libraries. The pipeline generates standardized 256×256 pixel tiles with 64-pixel overlap between adjacent sections, representing 2560×2560 pixel regions of the original slide at a downsampling factor of 10. See Section 4.4 for more information on tiles. The resulting dataset comprises 59,612 tiles, representing diverse tissue characteristics and pathological conditions.

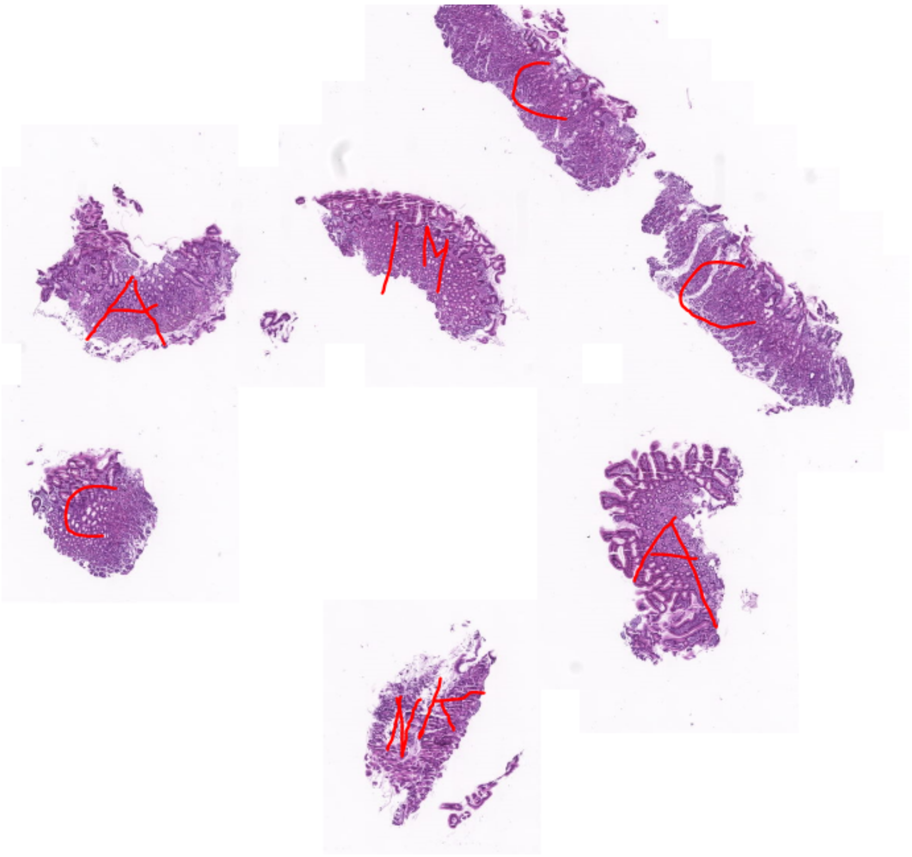


Figure 6: Digital annotation example showing pathologist-marked tissue classification on a gastric biopsy slide.

6 Patient Selection and Cohort Definition

6.1 Patient Flow Analysis

The initial cohort consisted of 274 unique patients with histological slides across multiple scanners, stain types, and annotations. A systematic selection process was applied to ensure data consistency and validity for this study's specific aims (Table 1, Figure 7).

Table 1: Patient Selection Flow

Selection Stage	Patient Count	Percentage
Initial Cohort	274	100%
Valid Annotation Cohort	252	92.0%
Scanner1 Cohort	222	81.0%
Scanner1 HE Staining Cohort	210	76.6%
Scanner1 Tissue Analysis Cohort	201	73.4%
Scanner1 Inflammation Analysis Cohort	204	74.4%
Overlap (in both analysis cohorts)	195	71.2%

6.2 Selection Process and Criteria

The patient selection process followed a systematic approach with specific exclusion criteria at each stage.

6.2.1 Annotation Validity Assessment

The annotation validity assessment ensured data quality and reliable model training. Valid annotations were determined using a two-tier validation system independently assessing tissue type and inflammation status. This approach allowed slides to be valid for one task but not necessarily the other, maximizing the usable data for each specific analysis. For tissue classification, annotations were considered valid if they explicitly identified regions as either "corpus" or "antrum". Each particle within a slide was evaluated independently, allowing slides with multiple tissue types to contribute to the analysis. Importantly, slides containing particles with intermediate or unclear tissue types were not entirely excluded; instead, only the specifically marked corpus and antrum regions were used for tissue classification. For inflammation classification, validity was assessed at the slide-level. Only slides with clear "inflamed" or "noninflamed" status designations were considered valid. All particles within a valid slide inherited their inflammation status, regardless of their tissue type classification.

Following these criteria:

- Excluded 22 patients lacking valid annotations for either classification task (See Section A.6, *Patient Exclusion*)
- Retained partially valid slides (valid for one task but not the other) to maximize usable data
- Implemented consistent validation rules across both scanner cohorts

Section A.5, *Valid Annotations Example*, provides a detailed example of this validation process, demonstrating how these criteria were applied to specific cases.

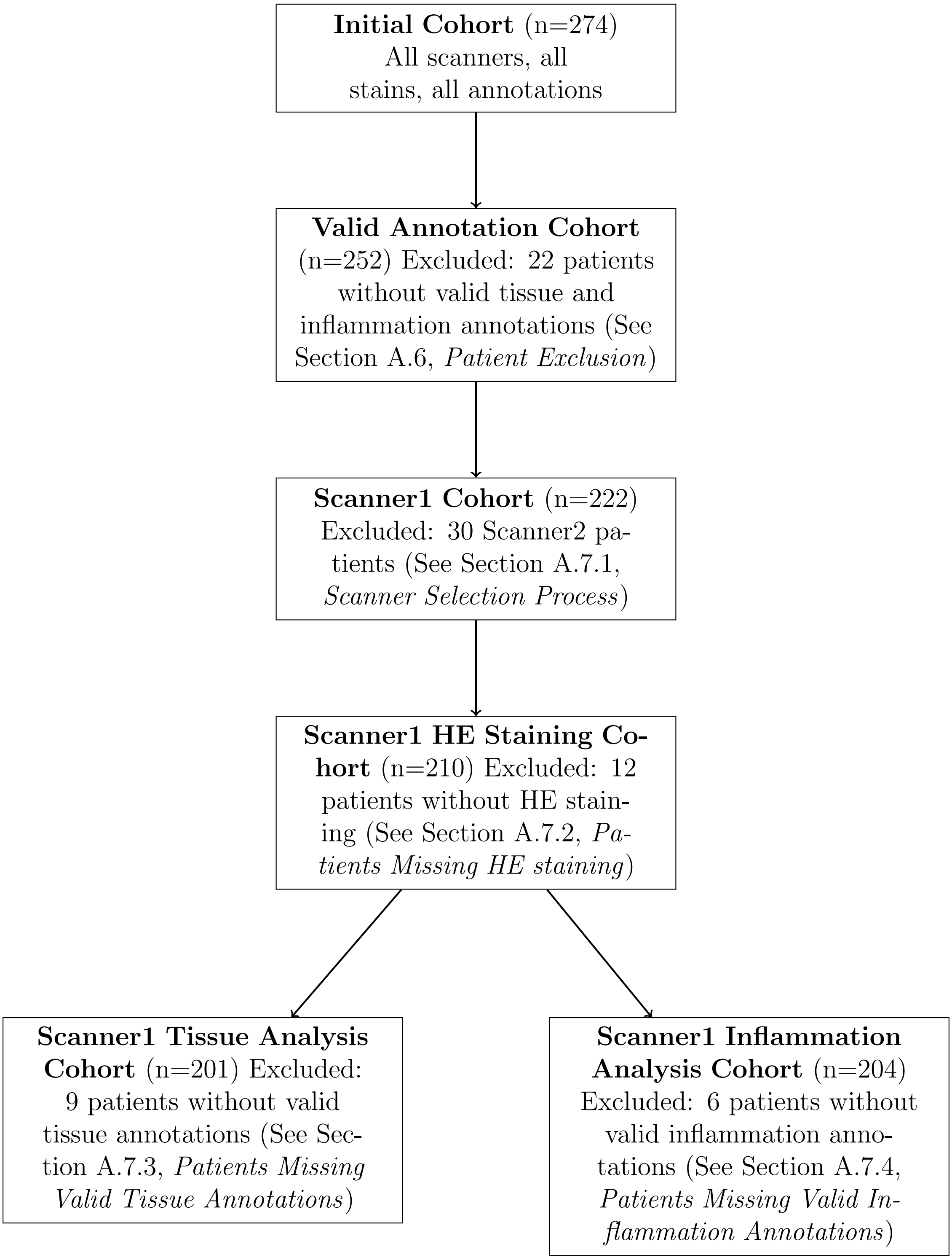

Initial Cohort (n=274) All scanners, all stains, all annotations
Valid Annotation Cohort (n=252) Excluded: 22 patients without valid tissue and inflammation annotations (See Section A.6, *Patient Exclusion*)
Scanner1 Cohort (n=222) Excluded: 30 Scanner2 patients (See Section A.7.1, *Scanner Selection Process*)
Scanner1 HE Staining Cohort (n=210) Excluded: 12 patients without HE staining (See Section A.7.2, *Patients Missing HE staining*)
Scanner1 Tissue Analysis Cohort (n=201) Excluded: 9 patients without valid tissue annotations (See Section A.7.3, *Patients Missing Valid Tissue Annotations*)
Scanner1 Inflammation Analysis Cohort (n=204) Excluded: 6 patients without valid inflammation annotations (See Section A.7.4, *Patients Missing Valid Inflammation Annotations*)

Figure 7: Patient flow diagram showing the systematic selection process for study inclusion. The process resulted in two partially overlapping cohorts for tissue and inflammation analysis. Detailed information about excluded cases can be found in the referenced sections.

6.2.2 Scanner Selection Strategy

- Excluded 30 patients with slides from Scanner2 (See Section A.7.1, *Scanner Selection Process*)
- Retained for later generalizability testing

The dataset comprises slides from two scanner generations (Pannoramic MIDI I and II), presenting a unique opportunity for assessing model generalizability across different imaging devices. Scanner2 patients (n=30) were deliberately reserved as a separate evaluation set to assess how well the trained models generalize to images from different acquisition hardware. This separation is crucial because scanner variations in factors like color calibration and image processing can significantly impact model performance.

6.2.3 Stain Selection Strategy

- Limited to HE staining
- Excluded 12 patients without HE staining (See Section A.7.2, *Patients Missing HE staining*)

All gastric biopsies were routinely stained with all three stains (HE, PAS, MG) to reveal different tissue structures, representing standard pathological practice. While all three stain types exist for each case, our study focuses primarily on HE-stained slides as they represent the foundational staining method for histopathological assessment.

However, the Scanner1 dataset contains slides that were inadvertently scanned with PAS or MG staining instead of HE. These misscans occurred due to two specific factors identified through consultation with our clinical collaborator:

- Medical technical assistants occasionally placed slides in an order different from the standard sequence (HE, PAS, MG) on the scanning trays, while the scanning protocol was configured to capture only the first slide in each sequence
- Höfling (2023) and Hempel (2023) performing the scanning, being unfamiliar with identifying stain types through visual inspection, did not recognize when non-HE stains were being digitized

In contrast, the Scanner2 cohort maintains complete staining consistency, with all three stain types (HE, PAS, MG) available for each case, providing a valuable opportunity for potential multi-stain analysis in future work.

6.2.4 Task-Specific Selection Strategy

Following the initial annotation validity assessment, we implemented task-specific selection criteria to ensure optimal dataset quality for each classification task. This additional filtering step was necessary to address the distinct requirements of tissue type and inflammation classification while maintaining the highest possible data quality for each task.

We established a cohort of 201 patients (73.4% of the initial dataset) with definitive anatomical annotations for tissue type classification. This selection required the presence of clearly identified antrum or corpus particles within each slide, ensuring reliable ground truth for the tissue classification task. Nine patients were excluded at this stage due to the abscence of any antrum or corpus tissue type annotations, with detailed exclusion rationales provided in Section A.7.3, *Patients Missing Valid Tissue Annotations.*

Similarly, we identified 204 patients (74.4% of the initial cohort) with reliable inflammation status annotations for inflammation classification. This cohort was established by excluding six patients whose slides lack definitive inflammation status classifications. The specific reasons for these exclusions are documented in Section A.7.4, *Patients Missing Valid Inflammation Annotations.*

There is substantial overlap between these two cohorts, with 195 (71.2% of the initial cohort) patients qualifying for both classification tasks. This significant intersection provides several advantages for our analysis. It enables robust comparative studies between tissue type and inflammation classification performance. Finally, this overlap validates our annotation and selection methodology, demonstrating that most properly annotated slides can meet the quality standards for both classification tasks while maintaining task-specific integrity through distinct annotation requirements.

This validation approach provided multiple benefits for dataset quality. By enforcing rigorous annotation standards, we ensured the data was sufficiently reliable for model training. We deliberately maintained a separation between tissue type and inflammation classification tasks, which allowed us to evaluate each independently. The careful documentation of our selection criteria makes our dataset generation process transparent and reproducible. Perhaps most importantly, our flexible approach maximized data utilization without compromising quality standards, allowing slides to be valid for one task while invalid for another. This practical solution helped us retain valuable samples that would otherwise have been excluded entirely, ultimately strengthening our analyses while maintaining methodological integrity.

6.3 Dataset Notation Convention

We employ a standardized notation system to ensure clarity and reproducibility when presenting results. Following the valid annotation criteria defined in Section 4.3.1, we denote datasets as $Dataset_{task,val}^{scanner-stain}$, where:

The superscript position indicates data source specifications:

$$scanner\text{-}stain = \begin{cases} \text{scanner} & \in \{S1, S2, S12\} \text{ (Scanner1, Scanner2, Both)} \\ \text{stain} & \in \{HE, PAS, MG\} \end{cases}$$

The subscript position indicates both analysis task and annotation validity:

$$task, val = \begin{cases} \text{task} & \in \{I, T, IT\} \text{ (Inflammation, Tissue, Both)} \\ \text{val} & \in \{v, a\} \text{ (valid, all)} \end{cases}$$

The 'val' subscript distinguishes between:

$$\begin{cases} v : \text{slides with any valid annotations per Section 4.3.1} \\ a : \text{all annotated slides, including those with exclusively non-valid annotations} \end{cases}$$

For examples, see Section A.3.

This notation system enables clear differentiation between single-scanner and combined-scanner datasets while maintaining the distinction between different annotation types and validity criteria. The $S12$ designation specifically indicates datasets that combine slides from both scanners, facilitating transparent reporting of dataset composition and scanner origin.

6.4 Dataset Analysis

6.4.1 Initial Cohort Analysis

Table 2: Dataset Summary Statistics (Dataset $^{S12-HE,PAS,MG}_{IT,a}$)

Characteristic	Scanner1	Scanner2
Total Slides	270	90
Unique Patients	244	30
Stain Distribution (slides)		
HE	244 (90.37%)	30 (33.33%)
PAS	13 (4.81%)	30 (33.33%)
MG	13 (4.81%)	30 (33.33%)
Inflammation Status (slides)		
Inflamed	138 (51.11%)	60 (66.67%)
Non-inflamed	99 (36.67%)	30 (33.33%)
Other	33 (12.22%)	0 (0.0%)

Table 2 presents the raw composition of our initial dataset before any quality filtering or validation procedures. This unfiltered cohort includes all digitized slides, irrespective of content quality or annotation validity, representing the complete collection of available data. The "Other" inflammation status category (12.22% of Scanner1 slides) encompasses slides with ambiguous inflammation patterns, insufficient diagnostic tissue, non-gastric tissue samples, or unclear expert annotations. Notably, the Scanner1 cohort contains an uneven staining distribution due to selective digitization practices, while Scanner2 exhibits a balanced representation across all three staining methods. It is important to emphasize that this initial dataset serves as the starting point for our systematic filtering pipeline, which subsequently removes non-representative samples, ambiguous annotations, and tiles that fail to meet our quality control criteria, as detailed in the following sections.

6.4.2 Valid Annotation Cohort Analysis

Following rigorous data acquisition and preprocessing, we conducted a comprehensive statistical analysis of the Valid Annotation Cohort ($\text{Dataset}_{IT,v}^{S12-HE,PAS,MG}$), which includes all slides with any valid annotations across both scanners and all staining methods. Table 3 provides a comprehensive overview of this cohort.

The Valid Annotation Cohort encompasses 252 patients and 338 slides, totaling 3,803 tissue particles. The dataset presents a balanced distribution of tissue types, with 44.8% corpus and 46.8% antrum particles, alongside 8.4% other tissue particles. Similarly, the inflammation status displays an appropriate distribution with 58.6% inflamed and 38.2% non-inflamed slides, alongside 3.2% other inflammation slides. This balanced representation is crucial for developing robust classification models.

6.5 Dataset Splits Creation

Robust train-validation-test partitioning is essential for unbiased model evaluation in computational pathology. We implemented a principled splitting strategy to address key challenges unique to histopathological data:

- Patient-level separation to prevent information leakage
- Systematic handling of heterogeneous staining modalities
- Stratification across imaging hardware to assess generalizability

To ensure methodological rigor and mitigate data leakage, we developed a patient-centric splitting protocol that separates patients across splits. While our dataset contains slides with multiple staining modalities (HE, PAS, MG), this study focuses exclusively on HE-stained slides based on the rationale detailed in Section A.7.2. Our preprocessing pipeline maintains a strict one-to-one correspondence between patients and slides in $\text{Dataset}_{I,T}^{S1-HE}$, with each patient represented by precisely one HE-stained slide. This design enables interchangeable use of patient and slide counts in statistical reporting while ensuring partition independence.

Table 3: Dataset Summary Statistics (Dataset$_{IT,v}^{S12-HE,PAS,MG}$)

Characteristic	**Total**	**Train**	**Val**	**Test**	**Test Scanner2**
SLIDE-LEVEL					
Total Slides	338	149	46	53	90
Scanner1 Slides	248	149	46	53	0
Scanner2 Slides	90	0	0	0	90
Total Patients	252	133	44	45	30
Total Slides	338	149	46	53	90
HE stained Slides	240	125	42	43	30
PAS stained Slides	54	15	3	6	30
MG stained Slides	44	9	1	4	30
Total Slides	338	149	46	53	90
Inflamed Slides	198	82	22	34	60
Non-inflamed Slides	130	60	21	19	30
Other Inflammation Status	10	7	3	0	0
PARTICLE-LEVEL					
Total Particles	3803	1638	563	511	1091
Corpus Particles	1704	853	260	195	396
Antrum Particles	1779	675	247	284	573
Other Tissue Particles	320	110	56	32	122
TILE-LEVEL					
Total Tiles	59,612	24,460	8,640	7,917	18,595
Inflammation Classification					
Inflamed Tiles	38,920	15,717	5,250	5,838	12,115
Non-inflamed Tiles	19,010	7,513	2,938	2,079	6,480
Other Inflammation Status	1,682	1,230	452	0	0
Tissue Classification					
Corpus Tiles	28,472	13,137	4,532	3,445	7,358
Antrum Tiles	26,409	10,251	3,249	4,009	8,900
Other Tissue Tiles	4,731	1,072	859	463	2,337

6.5.1 Data Splitting Implementation Methodology

The data splitting process was implemented through a specialized Python `DataSplitter` class adhering to the following methodological principles:

- **Patient-Level Stratification**: Partitioning was performed at the patient level rather than at individual slide or tile granularity. This approach ensures that all slides from a single patient remain within the same partition, preventing cross-contamination during model validation.

- **Scanner Stratification**: We explicitly separated Dataset$_{IT,v}^{S1-HE,MG,PAS}$ and Dataset$_{IT,v}^{S2-HE,MG,PAS}$ to enable rigorous generalizability assessment. Dataset$_{IT,v}^{S1-HE,MG,PAS}$ was distributed across train/validation/test partitions for the primary analysis, while Dataset$_{IT,v}^{S2-HE,MG,PAS}$ was reserved exclusively for generalizability evaluation.

- **Reproducibility Control**: A fixed random seed (42) ensured deterministic partitioning while maintaining stochastic patient distribution across splits. The data splitting script (create_splits.py) uses the base seed plus 2 (seed 42+2=44) to ensure balanced class distribution. This offset produced a better balance between train/validation/test sets than the default seed alone, as detailed in Appendix A.8.

6.5.2 Split Ratios and Validation

To optimize the trade-off between model development and evaluation, we implemented a systematic data partitioning strategy. The valid annotations dataset (Dataset$_{IT,v}^{S1-HE,MG,PAS}$) was divided into three non-overlapping partitions:

- Training set: 60% for model development and parameter estimation
- Validation set: 20% for hyperparameter and threshold optimization
- Test set: 20% for unbiased performance evaluation

While our preprocessing pipeline supports multi-stain analysis (HE, MG, PAS), we focused exclusively on HE-stained slides to maintain methodological consistency. The Scanner2 cohort (Dataset$_{IT,v}^{S2-HE,MG,PAS}$) was deliberately isolated as a separate evaluation set to assess model performance across different imaging devices, a critical factor in clinical deployment scenarios. This cross-scanner generalizability assessment is analyzed separately in Sections 9.3 and 9.4.

7 Dataset Splits Distribution

7.1 HE Staining Cohort Analysis

Table 4 provides a detailed breakdown of the Scanner1 and Scanner2 HE Staining Cohort across partitions. It illustrates the distribution of this cohort across training, validation, and test partitions. The Scanner1 HE cohort comprises 210 slides (excluding Scanner2), distributed in a 60:20:20 ratio across training, validation, and test sets. This partition ensures sufficient data for model training while allowing for robust validation and unbiased testing. This resulted in the allocation of 125 slides to the training set, 42 to the validation set, and 43 to the test set, maintaining our

Table 4: Detailed Composition of HE Dataset Partitions (Dataset$_{IT,v}^{S12-HE}$)

Characteristic	Training	Validation	Test	Test Scanner2	Total
HE stained					
Slides	125	42	43	30	240
Patients	125	42	43	30	240
Inflammation Status					
Inflamed	65	19	24	20	128
Non-inflamed	57	20	19	10	106
Tissue Particles					
Corpus	737	232	163	144	1276
Antrum	524	215	200	205	1144
Intermediate	67	48	32	42	189

targeted 60:20:20 distribution ratio. The balanced distribution of slides and particles across partitions ensures consistent evaluation conditions. The distribution of tissue types and inflammation status is maintained across partitions, with minor variations that reflect natural biological diversity.

7.2 Tile-Level Distribution Analysis

Following the hierarchical organization of our dataset, we analyzed the distribution of extracted tiles across different dataset splits and staining types. We analyzed the comprehensive tile counts for the entire dataset, revealing a substantial volume of image data available for model training and evaluation. Our dataset comprises 41,107 tiles across all stains within Scanner1, with 34,141 tiles derived from Dataset$_{IT,v}^{S12-HE}$. This distribution ensures sufficient data volume for deep learning model development.

The tile distribution follows a consistent 60:20:20 split ratio, with approximately 60% allocated to training (60.2% for HE stains, 59.5% for all stains), 21% to validation (21.7% for HE stains, 21.0% for all stains), and 19% to testing (18.1% for HE stains, 19.3% for all stains). This balanced distribution supports robust model development and unbiased evaluation within the Scanner1 domain.

It is important to note that these figures represent the total extracted tiles before quality filtering. During the model training and evaluation phases, additional filtering is applied to exclude tiles with insufficient tissue content (less than 10% tissue area), as detailed in Section 8 (Table 5, 6). This filtering step ensures that only diagnostically relevant tiles are used for model development, enhancing the reliability of our classification approach.

7.3 Cross-Scanner Generalization Dataset

We implemented a systematic approach for handling the Scanner2 cohort to facilitate rigorous cross-scanner generalization testing. This dedicated evaluation Dataset$_{IT,v}^{S2-HE,MG,PAS}$) was kept separate from all training and validation procedures. It serves as an independent test set for assessing model performance across different scanning hardware environments. Figure 8 illustrates the patient selection workflow for this generalizability cohort.

As confirmed by Dr. Braunecker, the Scanner2 cohort differs from Scanner1 in the scanning hardware and slide selection methodology. While Scanner1 slides were systematically created from archived patient blocks, Scanner2 slides were specifically curated from routine diagnostic work to include: (1) underrepresented case types such as non-inflamed antrum particles, (2) intermediate zone mucosa examples, and (3) particularly clear exemplary cases. This purposeful selection may have introduced a distribution bias that partially explains performance differences observed in cross-scanner generalization experiments.

Unlike the Scanner1 cohort, all Scanner2 patients had the complete staining sets with all three stain types (HE, PAS, MG) scanned for each slide. All 30 Scanner2 patient slides provided valid annotations for the inflammation classification task. However, two patients (IDs 260 and 263) were excluded for tissue classification because they contained only intermediate tissue regions without clear corpus or antrum annotations, resulting in a 28-patient tissue analysis cohort.

8 Deep Learning Model Development

8.1 Technical Contributions and Overview

While previous work in the research group (Hempel, 2023; Höfling, 2023) provided a valuable foundation, our work makes several distinct contributions:

- **Annotation Framework:** Implementation of a hierarchical annotation system that supports both tissue type and inflammation status classification
- **Quality Control:** Introduction of more stringent quality control measures, including more detailed expert pathologist review
- **Generalizability:** Explicit consideration of scanner variability through the inclusion of data from multiple scanner generations
- **Preprocessing Pipeline:** Development of a more sophisticated preprocessing workflow that handles varying scanner outputs and staining patterns

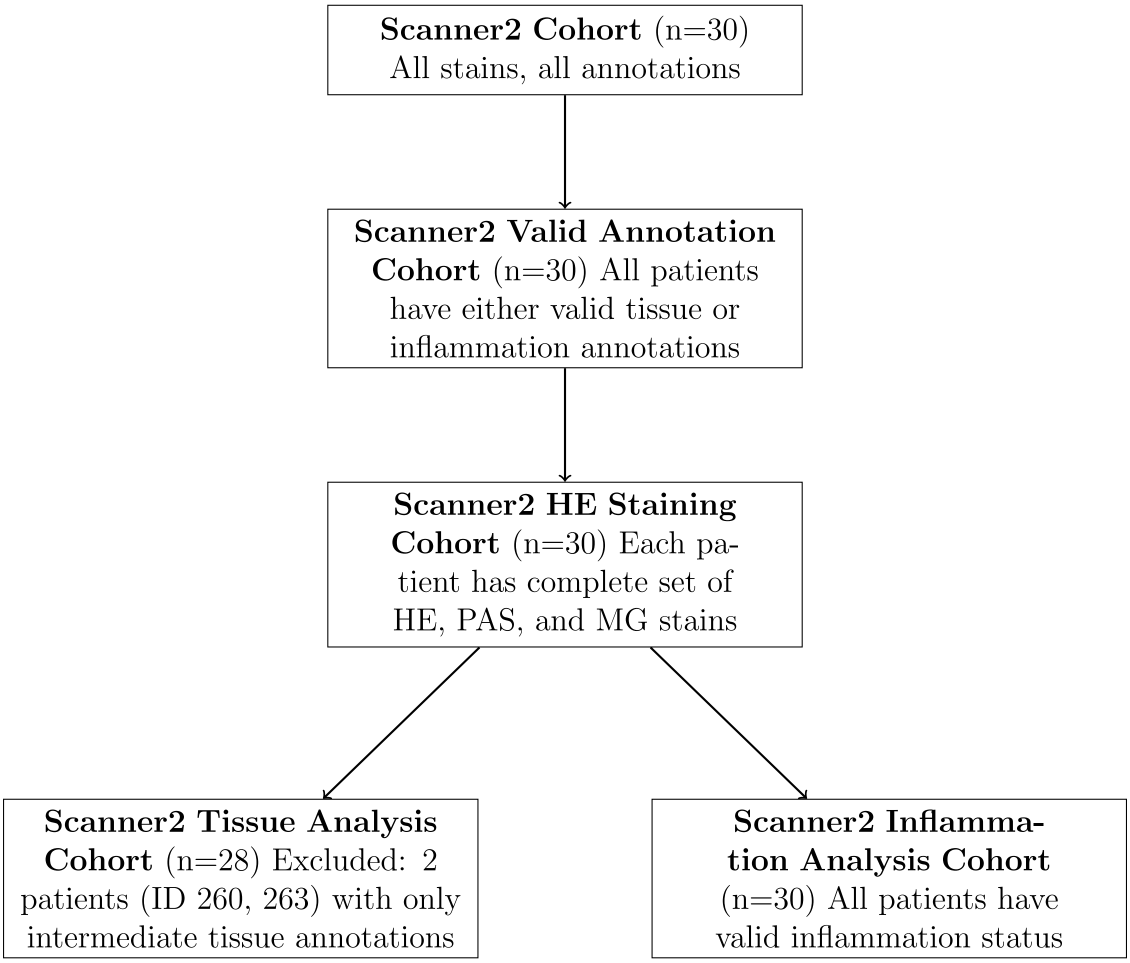

Scanner2 Cohort (n=30) All stains, all annotations
Scanner2 Valid Annotation Cohort (n=30) All patients have either valid tissue or inflammation annotations
Scanner2 HE Staining Cohort (n=30) Each patient has complete set of HE, PAS, and MG stains
Scanner2 Tissue Analysis Cohort (n=28) Excluded: 2 patients (ID 260, 263) with only intermediate tissue annotations
Scanner2 Inflammation Analysis Cohort (n=30) All patients have valid inflammation status

Figure 8: Patient flow diagram showing the systematic selection process for inclusion in the Scanner2 cohort study. The process resulted in two partially overlapping cohorts for tissue and inflammation analysis, with all patients having complete staining sets (HE, PAS, MG) but varying annotation validity.

Developing our deep learning models for histological image classification followed a systematic pipeline comprising data handling, model architecture design, hyperparameter optimization, and evaluation. All experiments were conducted on a dual-GPU system with two NVIDIA RTX A5000 graphics cards, each featuring 24 GB of GDDR6X memory. This high-performance computing infrastructure enabled efficient parallel training and extensive hyperparameter optimization for both models.

Our development process focused on two independent classification models:

1. **Tissue Type Classification Model**: Specialized in distinguishing between antrum and corpus tissue regions, requiring sensitivity to subtle architectural differences in tissue organization and cell distribution patterns.

2. **Inflammation Status Detection Model**: Designed to identify inflammatory patterns in tissue samples, focusing on detecting cellular infiltrates and other inflammation-associated morphological changes.

Each model was independently trained and optimized, although both shared the same architectural foundations to ensure consistent methodology while allowing for task-specific adaptations in hyperparameters and training protocols.

8.2 Dataset Implementation

The foundation of our model development pipeline is a custom PyTorch `HistologyDataset` class that handles efficient loading and preprocessing of histological image tiles. This implementation addresses several key challenges specific to histopathology through a comprehensive data management approach.

Our dataset architecture supports a hierarchical data organization that mirrors the natural structure of histopathological slides: WSIs contain multiple tissue particles, which are further divided into individual tiles. The implementation provides built-in support for both inflammation and tissue classification tasks.

To ensure data quality, the implementation incorporates an automated tissue content validation methodology. This process employs grayscale intensity analysis to quantify the proportion of tissue-containing regions within each tile. The validation algorithm converts input images to grayscale space and applies an intensity threshold of 240 on the 8-bit scale (0-255) to differentiate between tissue and background regions. Tiles are retained for analysis only if the background proportion remains below 90%, effectively filtering out regions dominated by non-tissue content while preserving sufficient diagnostic information. Table 5 presents the resulting distribution of filtered tiles (containing >10% tissue) across both classification tasks, demonstrating the balanced composition of our training, validation, and test sets after quality filtering.

The dataset implementation also supports domain adaptation research through flexible filtering capabilities between Scanner1 and Scanner2. This feature allows selective training and evaluation of images from specific scanner devices, enabling the

Table 5: Dataset Distribution Across Classification Tasks

	Particles	Total Tiles			
Inflammation Classification			Non/Inflamed		% Inf.
Training (122)	1,335	19,747	7,107	12,640	64.0
Validation (39)	467	6,852	2,716	4,136	60.4
Test (43)	395	6,120	2,049	4,071	66.5
Tissue Classification			Antrum/Corpus		% Corp.
Training (121)	1,261	19,099	7,839	11,260	59.0
Validation (38)	447	6,425	2,610	3,842	59.5
Test (42)	363	5,668	2,766	2,902	51.2

Note: Numbers in parentheses indicate unique slides. Percentages show the proportion of positive class (inflamed/corpus) within respective splits.

investigation of scanner-specific characteristics and their impact on model generalization. Our implementation handles multiple data splits (train/validation/test/test_scanner2) with consistent processing, facilitating reproducible evaluation across development stages.

A key feature of our implementation is comprehensive metadata handling. Each image tile is associated with rich contextual information, including slide origin, particle identifier, and class-specific annotations (inflammation status or tissue type). This metadata is propagated through the data pipeline and made available during training and inference, enabling hierarchical aggregation of predictions from tile to particle to slide-level. The metadata integration allows for nuanced evaluation across different levels of the histological hierarchy, better reflecting real-world diagnostic workflows.

The dataset's modular design and efficient memory management facilitate large-scale processing, which is critical for WSI analysis, where individual slides may yield thousands of tiles. This implementation establishes a standardized baseline by systematically applying consistent quality control measures across all extracted tiles, while significantly reducing manual curation requirements.

8.3 Model Architecture

Due to the complex nature of tissue morphology, selecting an appropriate deep learning architecture for histological image analysis presents unique challenges. We systematicly evaluated modern architectures, starting with an established baseline and exploring state-of-the-art models.

8.3.1 Architectural Selection and Design

We established ResNet18 (He et al., 2015) as our foundational baseline architecture, strategically building upon established precedent in gastric histopathology classification. This choice was supported by the successful implementations documented in prior research within our group. Hempel (2023) demonstrated ResNet18's effectiveness for tissue classification, achieving a tile-level accuracy of 89.1% and inflammation classification accuracy of 90.3% on their test set. Subsequently, these findings were validated by Höfling (2023) with independent implementations, reaching 88.6% accuracy for tissue classification, though they showed more modest results (72.7%) for the more challenging inflammation detection task. These consistent performance metrics across two separate investigations provided a reliable benchmark against which we could evaluate our architectural innovations and methodological refinements. ResNet18's computational efficiency and established performance characteristics made it an ideal reference point before exploring more complex architectures.

In addition to this baseline, we systematically evaluated five state-of-the-art architectures, each representing different approaches to computer vision:

1. ConvNeXt Large (Liu et al., 2022b): A modern evolution of traditional convolutional networks incorporating transformer-inspired design elements.

2. Swin Transformer V2-B (Liu et al., 2022a): A hierarchical vision transformer with shifted windows for efficient self-attention computation.

3. Prov-GigaPath (Xu et al., 2024): A foundation model pre-trained on 1.3 billion histopathology image tiles spanning 171,189 WSIs.

4. DenseNet121 (Huang et al., 2018): A convolutional network with dense connectivity patterns, where each layer receives feature maps from all preceding layers, enhancing feature reuse and gradient flow.

5. DenseNet169 (Huang et al., 2018): A deeper variant of DenseNet with 169 layers, maintaining the same dense connectivity principle while offering increased representational capacity.

8.3.2 Implementation and Transfer Learning

Our model implementation centers on a flexible `HistologyClassifier` class that supports multiple state-of-the-art convolutional neural network backbones with architecture-specific adaptations:

- **Standard CNN backbones**: ResNet18, DenseNet121, and DenseNet169 were initialized with pre-trained ImageNet weights from torchvision, providing robust general-purpose feature extraction capabilities.

- **Modern architectures**: ConvNeXt Large and Swin Transformer V2-B were similarly initialized with ImageNet weights but required specialized head adaptations to accommodate their unique feature extraction patterns.
- **Domain-specific foundation model**: For Prov-GigaPath, we employed a unique transfer learning strategy by utilizing a frozen encoder pre-trained on 1.3 billion histopathology images. This encoder outputs 1536-dimensional feature vectors which are then processed by a trainable classification head consisting of dimensionality reduction (1536→512), ReLU activation, dropout regularization, and final binary classification.

The architecture implementation includes these key components:

- **Input preprocessing pipeline**:
  - Spatial dimensions: 224×224 pixels (resized and center-cropped from original tiles)
  - Normalization using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$)
- **Task-specific classifier head modifications**:
  - Architecture-aware feature handling (preserving spatial information where beneficial)
  - Task-calibrated dropout layers with rates individually optimized per architecture
  - Binary classification output for both tissue and inflammation tasks
- **Training optimization**:
  - AdamW optimizer with architecture-specific learning rates and weight decay
  - Mixed precision training support via PyTorch's GradScaler
  - BCE with logits loss function with class weighting capabilities

For GigaPath specifically, we employed a highly efficient learning approach by keeping the extensive pre-trained encoder frozen during training, allowing optimization to focus exclusively on the classification head parameters. This strategy significantly reduced training time while leveraging the rich histopathological features captured by the foundation model's extensive pre-training. In contrast, the other models were fine-tuned end-to-end, allowing all parameters to adapt to our specific histological classification tasks.

8.4 Training Methodology

We implemented a training methodology optimized for histological image classification tasks, focusing on hyperparameter optimization, efficient training pipelines, and robust validation strategies.

8.4.1 Hyperparameter Optimization Framework

We optimized hyperparameters using the Optuna framework with the Tree-structured Parzen Estimators (TPE) sampler (Akiba et al., 2019). This Bayesian optimization approach efficiently navigates high-dimensional parameter spaces by constructing probabilistic models of objective function values. We employed a structured search space with architecture-specific parameter ranges to ensure comprehensive architectural comparison while maintaining computational feasibility.

The optimization space included the following dimensions:

- **Model architecture**: Six distinct architectures were evaluated (ResNet18, ConvNeXt Large, Swin Transformer V2-B, GigaPath, DenseNet121, and DenseNet169), each representing different approaches to feature extraction
- **Architecture-aware batch sizes**: Dynamically adjusted based on model memory requirements, with smaller batches for memory-intensive models (16-24 for GigaPath, 24-64 for DenseNet169) and larger batches for lightweight models (64-128 for ResNet18)
- **Learning rates**: Model-specific ranges on log-scale, with foundation models using higher rates (5e-4 to 1e-2 for GigaPath, leveraging frozen encoder properties) and deeper networks using lower rates (5e-5 to 2e-4 for DenseNet variants)
- **Regularization parameters**: Weight decay (1e-5 to 1e-3, log-scale) and dropout rates (0.0 to 0.3) to control overfitting
- **Class balancing**: Adaptive positive class weight ranges centered around class distribution-derived values, dynamically adjusted within ± 0.15 of the optimal theoretical value

To maximize computational efficiency and ensure robust evaluation, we implemented several advanced optimization mechanisms:

- **Hyperband pruning**: Early termination of underperforming trials based on performance trajectory, with minimum evaluation of 5 epochs before pruning decision
- **Early stopping**: Training terminated after 3 epochs without validation loss improvement

- **Gradient clipping**: Applied max norm of 1.0 to prevent gradient explosion during exploration of sensitive parameter regions
- **Multi-GPU parallelization**: Distributed batch processing across available GPUs using PyTorch's DataParallel

Each trial ran for up to 20 epochs with validation loss as the optimization objective. To mitigate the risk of local minima solutions, we conducted 100 trials per task, requiring approximately 2-3 days of computation on dual NVIDIA RTX A5000 GPUs. Following hyperparameter optimization, we selected the best configuration for each model architecture and performed extended training for 50 epochs. This comprehensive approach resulted in 12 fully trained models, six architectures for each classification task, enabling a thorough comparison of architectural performance across both tissue type and inflammation detection objectives.

8.4.2 Training Pipeline Implementation

Our training pipeline incorporated several technical optimizations to ensure efficient and reproducible model development. We implemented automatic mixed precision (AMP) training to accelerate computation while maintaining numerical stability. For multi-GPU environments, DataParallel was employed to distribute batch processing across available devices. The pipeline supports both SGD with momentum and AdamW optimizers, with optimization parameters derived from hyperparameter tuning.

Learning rate scheduling was implemented using either cosine annealing or reduction on plateau strategies, depending on model architecture and task requirements. All models were equipped with gradient clipping (max norm 1.0) to prevent gradient explosion and early stopping mechanisms to mitigate overfitting. Each training session maintained comprehensive metadata, including random seeds, CUDA configuration, and timestamp information to ensure reproducibility of results.

8.4.3 Data Processing and Augmentations

Data augmentation was tailored to histological image characteristics and implemented as configurable transformation pipelines. For training, we employed random resized crops (scale 0.8-1.0), random horizontal and vertical flips, and task-specific augmentations. For inflammation classification random rotations and Gaussian blur were found to improve performance, while tissue classification benefited primarily from rotational augmentations. All images underwent standard normalization using ImageNet statistics to facilitate transfer learning from pre-trained models.

Input data was processed using optimized DataLoader configurations with eight worker processes and pin memory to maximize GPU utilization. Based on memory constraints and convergence behavior, batch sizes were individually optimized for each architecture, ranging from 24 for DenseNet169 to 128 for ResNet18.

8.4.4 Validation Strategy and Model Selection

We implemented a hierarchical validation strategy that evaluated model performance at multiple levels of granularity. For inflammation classification, metrics were calculated at both tile and slide-levels, while tissue classification was evaluated at tile and particle-levels. This approach ensured that models were optimized for real-world clinical applications rather than just tile-level accuracy.

During training, models were validated after each epoch using a comprehensive set of metrics, including loss, accuracy, sensitivity, specificity, F1 score, and AUC-ROC. Based on validation loss, the best-performing model weights were saved, and training continued until either the maximum epoch count was reached or early stopping was triggered. This methodology ensured robust model selection while preventing overfitting to the validation data.

All training sessions were documented with logging and performance visualization to facilitate experimental tracking and reproduction. The final models underwent evaluation on separate test sets to assess generalization performance in both same-scanner and cross-scanner scenarios.

8.5 Experimental Configurations

8.5.1 Tissue Classification Configuration

We conducted comprehensive hyperparameter optimization by systematically exploring 100 trials utilizing the Optuna framework with Tree-structured Parzen Estimators (TPE). This Bayesian optimization approach efficiently navigated the high-dimensional parameter space while prioritizing promising regions. The computational demands of this process were substantial, requiring over 72 hours of continuous computation on dual NVIDIA A5000 GPUs for the tissue classification task alone, underscoring the resource-intensive nature of neural architecture optimization.

Figure 9 illustrates the progression of validation loss (objective value) throughout the tissue task optimization process. Starting from approximately 0.3, the validation loss exhibits a consistent downward trend during early trials, with significant improvements occurring within the first 25 trials. The optimization curve demonstrates classic diminishing returns, with only marginal improvements observed during the subsequent 75 trials despite extensive computational investment. This plateau indicates effective convergence toward optimal hyperparameter configurations, suggesting that the search space was thoroughly explored.

The optimal hyperparameter configurations for each architecture are detailed in the Results section (Table 9). ConvNeXt Large achieved the lowest validation loss and highest performance on most metrics across all architectures, while GigaPath demonstrated efficient transfer learning capabilities with comparable validation AUC to more complex models despite its lower validation accuracy.

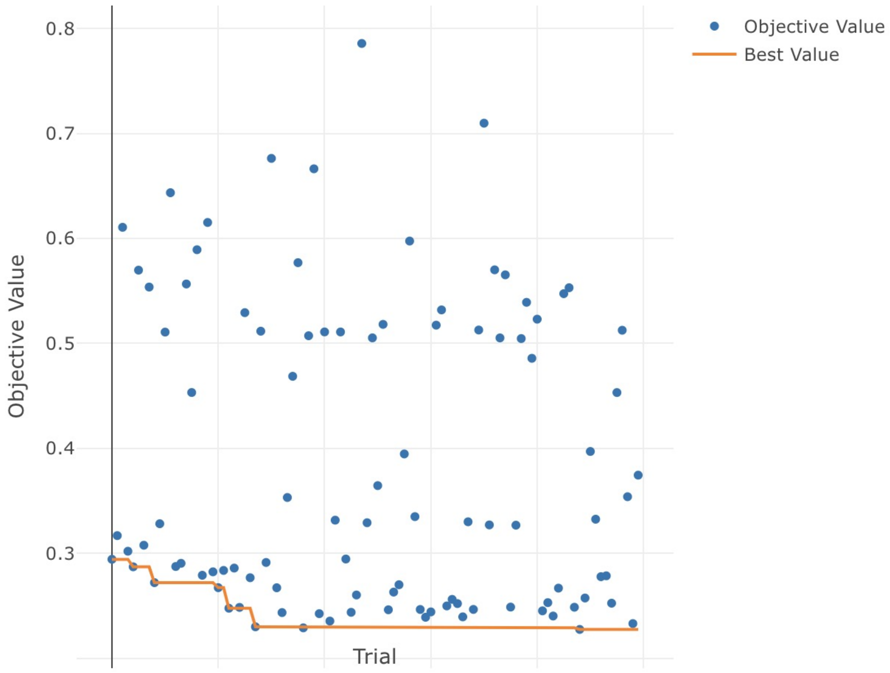


Figure 9: Hyperparameter optimization trajectory across 100 trials for tissue classification task. The blue points represent individual trial validation losses, while the orange line tracks the best validation loss achieved to date. The optimization demonstrates rapid initial improvement before reaching diminishing returns around trial 25, with minimal improvement thereafter despite extensive additional exploration.

8.5.2 Inflammation Classification Configuration

We employed a similar hyperparameter optimization approach through 100 trials using the Optuna framework with Tree-structured Parzen Estimators (TPE) for the inflammation detection task. While maintaining the same methodological rigor as the tissue classification optimization, the inflammation classification process required less computational resources, completing in around 60 hours on the same dual NVIDIA A5000 GPU system.

Figure 10 depicts the optimization trajectory for the inflammation classification task. The validation loss began at approximately 0.5 and rapidly improved to 0.3 by the second trial, followed by further refinement to 0.27 at trial 13. The optimization process exhibited an extended plateau before achieving the optimal validation loss of 0.244 at trial 66 with the Swin-V2-B architecture. This pattern suggests that the parameter space for inflammation classification presented different optimization challenges than tissue classification, potentially requiring more extensive exploration to identify optimal configurations.

The detailed hyperparameter configurations for the inflammation classification task are presented in Table 12. Interestingly, while ConvNeXt Large demonstrated superior performance for tissue classification, the Swin-V2-B architecture achieved the

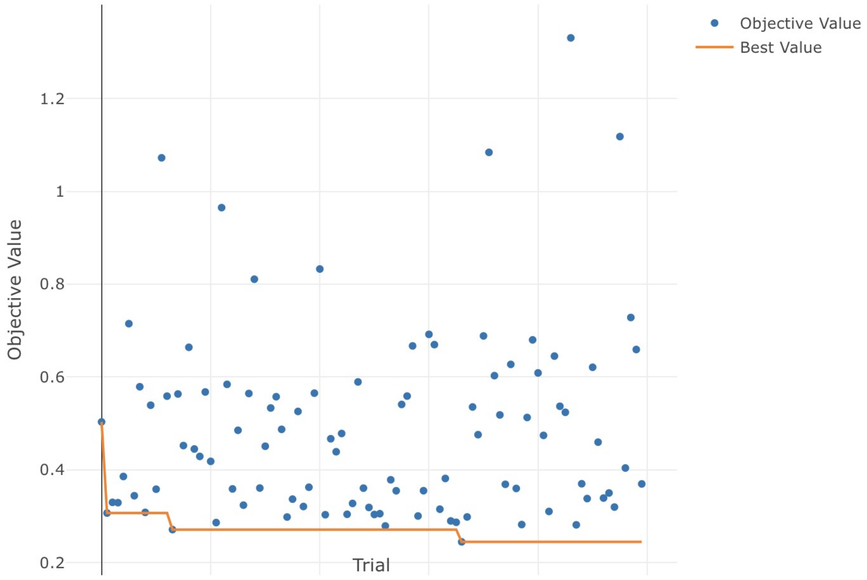


Figure 10: Hyperparameter optimization trajectory across 100 trials for inflammation classification task. The blue points represent individual trial validation losses, while the orange line tracks the best validation loss achieved to date. After initial rapid improvement, the optimization process required extensive exploration before achieving the optimal configuration at trial 66.

lowest validation loss for inflammation detection, suggesting that different architectural designs may be optimal for capturing the distinct visual patterns associated with each classification task.

8.6 Hierarchical Prediction Aggregation

A critical challenge in WSI classification is effectively aggregating numerous tile-level predictions into a unified particle- or slide-level diagnosis. This hierarchical aggregation process is essential in histopathological analysis, where diagnostically relevant features may be concentrated in specific regions rather than distributed uniformly throughout the tissue. To address this challenge, we systematically evaluated multiple aggregation strategies that capture different statistical properties of the tile prediction distributions, enabling robust inference at higher hierarchical levels.

8.6.1 Mathematical Formulation

Let $S = \{t_1, t_2, \ldots, t_n\}$ represent a particle or slide containing n tiles, and $P = \{p_1, p_2, \ldots, p_n\}$ denote the corresponding probability predictions for each tile, where $p_i \in [0, 1]$ represents the model's confidence that tile t_i belongs to the positive class (corpus for tissue classification or inflamed for inflammation classification). We define the particle-/slide-level aggregation function $f : P \to [0, 1]$ that maps the set of tile predictions to a single aggregated prediction score, which is subsequently thresholded to obtain the final binary classification.

We implemented and systematically evaluated the following aggregation strategies:

1. **Mean Aggregation**: The arithmetic mean of all tile predictions, providing an unweighted average across the entire particle/slide:

$$f_{\text{mean}}(P) = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{1}$$

2. **Median Aggregation**: The median value of all tile predictions, offering robustness against outlier predictions:

$$f_{\text{median}}(P) = \text{median}(p_1, p_2, \ldots, p_n) \tag{2}$$

3. **Top-K Percentile Mean**: The mean of the top $k\%$ of tile predictions, where $k \in \{10, 20, 30\}$. This approach focuses on the most confident positive predictions, which is particularly relevant when diagnostically significant features appear in only a small portion of the particle/slide:

$$f_{\text{top-}k\%}(P) = \frac{1}{|P_k|}\sum_{p \in P_k} p \tag{3}$$

where P_k represents the subset containing the top $k\%$ of values from P, formally defined as:

$$P_k = \{p_i \in P \mid p_i \geq P_{(1-k/100)}\} \tag{4}$$

with $P_{(1-k/100)}$ representing the $(1 - k/100)$-quantile of set P.

8.6.2 Comparative Analysis of Aggregation Strategies

Each aggregation strategy exhibits distinct properties that make it suitable for specific scenarios in histopathological analysis:

- **Mean Aggregation** is computationally efficient and provides a global perspective of the tissue sample. This strategy performs optimally when diagnostically relevant features are distributed relatively uniformly throughout the tissue but may be suboptimal when significant features are localized to small regions or when the presence of artifacts leads to outlier predictions.

- **Median Aggregation** offers superior robustness to outliers compared to mean aggregation, which is particularly valuable in scenarios with noisy predictions or artifacts. However, this approach may be less sensitive to small regions of diagnostic significance when these regions constitute less than 50% of the sample.

- **Top-K Percentile Mean** strategies are specifically designed to capture instances where pathological features occupy only a small fraction of the slide. For example, a top-10% strategy would be particularly suitable for detecting focal inflammation or isolated regions of distinctive tissue architecture that may be present in just a small portion of the sample. However, this approach disregards the majority of predictions and may be more susceptible to false positives if artifacts are present in the sample.

The optimal aggregation strategy is inherently task-dependent and influenced by the classification target's specific biological and morphological characteristics. Different strategies might be optimal for inflammation detection, where inflammatory infiltrates may be focal or diffuse, compared to tissue type classification, where architectural patterns tend to be more globally distributed. The empirical evaluation of these strategies is presented in Section 9.

8.7 Cross-Scanner Generalization and Clinical Validation

To evaluate cross-scanner generalization capabilities of our models, we employed a dedicated set of slides from Scanner2 that were deliberately excluded from the training process. This separate cohort of 30 patients (described in Section A.7.1) provides a critical test environment for assessing model robustness when deployed on different imaging hardware. Our generalizability analysis follows the same rigorous quality control standards applied to the primary dataset. The automated tissue content validation methodology ensures consistency across scanner platforms by applying identical filtering criteria. Table 6 presents the resulting distribution of filtered tiles (containing >10% tissue content) from Scanner2 across both classification tasks.

Table 6: Dataset Distribution Across Classification Tasks

	Particles	Total Tiles			
Inflammation Classification			Non/Inflamed		% Inf.
Training (122)	1,335	19,747	7,107	12,640	64.0
Validation (39)	467	6,852	2,716	4,136	60.4
Test Scanner2 (30)	397	6,829	2,194	4,635	67.9
Tissue Classification			Antrum/Corpus		% Corp.
Training (121)	1,261	19,099	7,839	11,260	59.0
Validation (38)	447	6,425	2,610	3,842	59.5
Test Scanner2 (28)	349	5,922	3,212	2,710	45.8

Note: Numbers in parentheses indicate unique slides. Percentages show the proportion of positive class (inflamed/corpus) within respective splits.
The Test Scanner2 dataset represents a completely separate test set, while all training and validation was performed on Scanner1 data.

The present study extends previous work by addressing critical challenges in model generalization and clinical deployment. Central to this research is integrating data from multiple scanner types, specifically incorporating an additional 90 WSIs from a Pannoramic MIDI II scanner (3DHISTECH Ltd., Budapest, Hungary) alongside the existing MIDI I dataset. This multi-scanner approach enables systematic investigation of model robustness across different digitization platforms.

Our methodology encompasses several key research objectives:

1. **Dataset Expansion and Standardization:** Development of a comprehensive dataset incorporating WSIs from multiple scanners, with standardized protocols for:
    - Systematic quality assessment across scanner types
    - Unified annotation procedures ensure consistency
2. **Generalizability Assessment Framework:** Design of rigorous validation protocols including cross-scanner performance evaluation

9 Results

Our experimental evaluation assesses the performance of various deep learning architectures across two distinct classification tasks: inflammation detection and tissue type classification. We present a comprehensive analysis following a rigorous testing methodology that simulates real clinical deployment scenarios. All performance metrics and ROC curves presented in this section are derived exclusively from test set data, using Scanner1 test set for within-scanner evaluations and Scanner2 test set for cross-scanner generalization assessments.

To ensure methodological rigor, all classification thresholds were optimized strictly on the Scanner1 validation set and applied directly to the respective test sets without further adjustment. This approach replicates clinical reality where model deployment occurs on previously unseen data. The validation-optimized thresholds and corresponding aggregation strategies for tissue and inflammation classification can be found in Table 9 and Table 12, respectively.

For each classification task, we report hierarchical evaluation results at multiple levels of granularity: tile-level, particle-level (for tissue classification), and slide-level (for inflammation detection). This section highlights key findings with particular emphasis on comparing our models with previous approaches and evaluating within-scanner performance and cross-scanner generalization capabilities.

9.1 Within-Scanner Performance (Scanner1): Tissue Classification

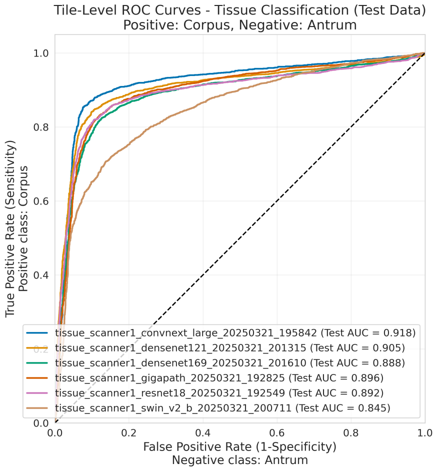


Figure 11: ROC curves comparing model performance for tissue classification at the tile-level on Scanner1 test-set

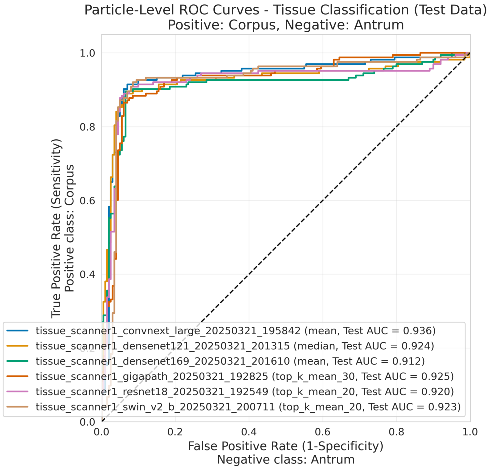


Figure 12: ROC curves comparing model performance for tissue classification at the particle-level on Scanner1 test-set

When comparing our results with previous work (Table 7), our models show competitive performance at the tile-level test performance, with our ResNet18 tissue classification model achieving 85.20% accuracy compared to Hempel's 89.10% and Höfling's 87.00%. This comparison is not entirely fair, as different dataset splits were used across studies. Furthermore, the metrics for previous works were calculated based on their published confusion matrices, as their models discriminated between corpus, antrum, and intermediate tissue types, while our approach focused on binary classification (corpus/antrum). Significantly, our study extends beyond previous work by reporting comprehensive particle-level metrics and cross-scanner generalization performance.

Table 8 presents the performance of different models for gastric tissue classification. Hyperparameters were optimized using Optuna with validation loss as the objective. The optimization process completed in 3 days, 2 hours, and 21 minutes on dual NVIDIA A5000 GPUs. All models, excluding ResNet18, were subsequently trained for 50 epochs.

The optimal hyperparameter configurations for each architecture are detailed in Table 8 under ”Implementation Details”. These were determined through a systematic Bayesian optimization process using Optuna (Akiba et al., 2019) with Tree-structured Parzen Estimators (TPE), which efficiently explored the high-dimensional parameter space to minimize validation loss.

Table 7: Performance Comparison of ResNet18 Models for Gastric Tissue Classification Across Studies

	Current Work	**Previous Work**†	
Study	**This Study**	**Hempel (2023)**	**Höfling (2023)**
Validation Performance			
Tile-Level Accuracy	86.81%	87.50%	88.56%
Tile-Level Sensitivity	88.76%	—	—
Tile-Level Specificity	83.95%	—	—
Tile-Level F1 Score	88.91%	—	—
Tile-Level AUC	92.23%	—	—
Test Performance			
Tile-Level Accuracy	85.20%	89.10%	87.00%
Tile-Level Sensitivity	85.32%	89.80%	91.10%
Tile-Level Specificity	85.08%	88.00%	85.20%
Tile-Level F1 Score	84.91%	90.40%	90.11%
Tile-Level AUC	89.20%	95.00%	95.00%
Aggregation Strategy			
Strategy	top_k_mean_20	majority vote	probabilistic avg
Implementation Details			
Batch Size	64	32	64
Dropout Rate	0.000	0.000	0.000
Epochs	20	10	11
Learning Rate	1e-3	1e-4	1e-3
Weight Decay	0.000	0.000	0.000
Positive Class Weight	1.0	1.0	1.0
Optimizer	SGD	SGD	SGD

† Direct numerical comparison is not straightforward due to different dataset splits, multi-class vs. binary classification approaches, and metrics recalculated from published confusion matrices.

Note: All results are from ResNet18 models.

Note: Previous studies also classified intermediate tissue types. For comparison with our binary classification, only their antrum and corpus predictions were considered.

Note: Missing values (—) indicate metrics not reported in previous studies.

Table 9 highlights the impact of different aggregation strategies on model performance.

Table 8: Comprehensive Performance Comparison of Deep Learning Models for Gastric Tissue Type Classification (Scanner1)

Model **Metric**	**ConvNeXt** **Large**	**DenseNet121**	**DenseNet169**	**GigaPath**	**ResNet18**	**Swin-V2** **Base**
Validation Performance						
Tile-Level Accuracy	90.47%	89.09%	87.99%	87.74%	86.81%	81.18%
Tile-Level Sensitivity	93.75%	92.24%	89.28%	90.45%	88.76%	81.70%
Tile-Level Specificity	85.63%	84.44%	86.09%	83.75%	83.95%	80.42%
Tile-Level F1 Score	92.13%	90.97%	89.85%	89.78%	88.91%	83.80%
Tile-Level AUC	92.91%	94.26%	94.05%	93.04%	92.23%	86.97%
Particle-Level Accuracy	96.35%	94.62%	95.45%	96.33%	95.47%	93.91%
Particle-Level Sensitivity	98.28%	94.83%	97.41%	98.71%	96.98%	95.26%
Particle-Level Specificity	94.42%	94.42%	93.49%	93.95%	93.95%	92.56%
Particle-Level F1 Score	96.61%	94.83%	95.76%	96.62%	95.74%	94.24%
Particle-Level AUC	97.05%	98.06%	98.15%	97.70%	97.38%	95.96%
Test Performance						
Tile-Level Accuracy	88.71%	87.07%	84.23%	84.30%	85.20%	78.41%
Tile-Level Sensitivity	87.06%	84.60%	84.16%	76.46%	85.32%	68.91%
Tile-Level Specificity	90.28%	89.42%	84.29%	91.76%	85.08%	87.46%
Tile-Level F1 Score	88.27%	86.46%	83.89%	82.62%	84.91%	75.69%
Tile-Level AUC	91.83%	90.52%	88.79%	89.63%	89.20%	84.55%
Particle-Level Accuracy	92.01%	90.91%	91.46%	90.36%	91.46%	91.46%
Particle-Level Sensitivity	90.18%	87.73%	89.57%	87.12%	90.80%	90.80%
Particle-Level Specificity	93.50%	93.50%	93.00%	93.00%	92.00%	92.00%
Particle-Level F1 Score	91.02%	89.66%	90.40%	89.03%	90.52%	90.52%
Particle-Level AUC	93.61%	92.43%	91.15%	92.46%	92.03%	92.33%
Optimal Aggregation Strategy						
Strategy	mean	median	mean	top_k_mean_30	top_k_mean_20	top_k_mean_20
Implementation Details						
Batch Size	64	64	56	48	64	64
Dropout Rate	0.201	0.061	0.150	0.236	0.000	0.224
Epochs	50	50	50	50	20	50
Learning Rate	4.285e-4	1.894e-4	1.903e-4	9.328e-4	1.000e-3	4.372-4
Weight Decay	2.130e-4	1.077e-5	5.859e-4	9.532e-5	0.000	4.600e-5
Positive Class Weight	0.551	0.654	0.5704	0.551	1.000	0.713
Optimizer	AdamW	AdamW	AdamW	AdamW	SGD	AdamW

Note: Optimal aggregation strategies were determined through validation data optimization.

Note: Aggregation strategies determine how tile-level predictions are combined to make particle-level decisions.

Table 9: Comparison of Aggregation Strategies based on Validation Set for Gastric Tissue Classification Across Models

Model	**Strategy**	**Threshold**	**Balanced Accuracy**	**Sensitivity**	**Specificity**	**F1 Score**
ConvNeXt Large	**mean**	0.470	**96.35%**	98.28%	94.42%	96.61%
	median	0.497	96.13%	97.84%	94.42%	96.39%
	top_k_mean_10	0.900	95.87%	98.71%	93.02%	96.22%
	top_k_mean_20	0.900	96.33%	98.71%	93.95%	96.62%
	top_k_mean_30	0.875	96.35%	98.28%	94.42%	96.61%
DenseNet121	mean	0.595	94.41%	94.40%	94.42%	94.60%
	median	0.621	**94.62%**	94.83%	94.42%	94.83%
	top_k_mean_10	0.900	93.13%	97.41%	88.84%	93.78%
	top_k_mean_20	0.894	93.86%	96.55%	91.16%	94.32%
	top_k_mean_30	0.900	94.09%	96.55%	91.63%	94.51%
DenseNet169	**mean**	0.551	**95.45%**	97.41%	93.49%	95.76%
	median	0.597	95.45%	97.41%	93.49%	95.76%
	top_k_mean_10	0.900	92.23%	96.55%	87.91%	92.95%
	top_k_mean_20	0.896	93.86%	96.55%	91.16%	94.32%
	top_k_mean_30	0.896	94.55%	96.55%	92.56%	94.92%
GigaPath	mean	0.362	95.65%	98.28%	93.02%	96.00%
	median	0.253	95.38%	99.14%	91.63%	95.83%
	top_k_mean_10	0.759	94.70%	98.71%	90.70%	95.22%
	top_k_mean_20	0.756	96.10%	98.71%	93.49%	96.42%
	top_k_mean_30	0.705	**96.33%**	98.71%	93.95%	96.62%
ResNet18	mean	0.485	93.89%	95.69%	92.09%	94.27%
	median	0.616	93.30%	93.10%	93.49%	93.51%
	top_k_mean_10	0.900	94.07%	96.98%	91.16%	94.54%
	top_k_mean_20	0.900	**95.47%**	96.98%	93.95%	95.74%
	top_k_mean_30	0.873	95.24%	96.98%	93.49%	95.54%
Swin_v2_b	mean	0.432	92.78%	94.40%	91.16%	93.19%
	median	0.428	92.38%	92.67%	92.09%	92.67%
	top_k_mean_10	0.678	93.03%	93.97%	92.09%	93.36%
	top_k_mean_20	0.632	**93.91%**	95.26%	92.56%	94.24%
	top_k_mean_30	0.588	93.87%	96.12%	91.63%	94.29%

Note: Aggregation strategies represent different methods for combining tile-level predictions to make particle-level decisions.
Values shown are from the aggregation strategies based on the validation set performance.
Threshold values were optimized on validation data to maximize balanced accuracy.
”mean”: Average of all tile predictions within a particle
”median”: Median of all tile predictions within a particle
”top_k_mean_n”: Average of the top n% tile predictions within a particle

9.2 Within-Scanner Performance (Scanner1): Inflammation Classification

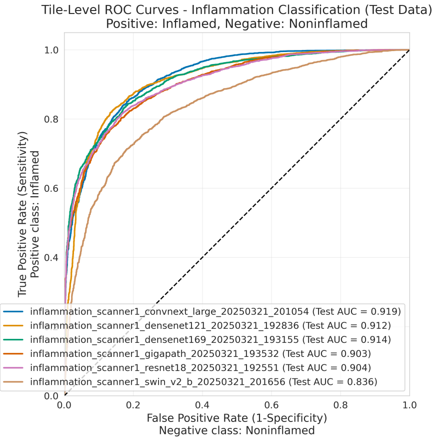


Figure 13: ROC curves comparing model performance for inflammation classification at the tile-level on Scanner1 test-set

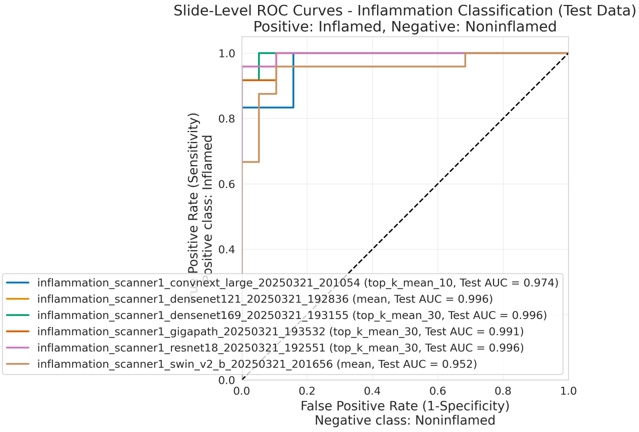


Figure 14: ROC curves comparing model performance for inflammation classification at the slide-level on Scanner1 test-set

Table 10 shows how our models compare to previous work. While Hempel (2023) reported higher tile-level metrics (90.30% accuracy vs. our 81.70%), direct comparisons are challenging due to different dataset splits, evaluation methodologies, and metrics recalculated from published confusion matrices.

Table 11 presents the performance of different models for gastric inflammation classification. Hyperparameters were optimized using Optuna with validation loss as the objective. The optimization process completed in 2 days, 12 hours, and 46 minutes on dual NVIDIA A5000 GPUs. All models, excluding ResNet18 which was trained for 25 epochs, were subsequently trained for 50 epochs.

The optimal hyperparameter configurations for each architecture are also detailed in Table 11 under "Implementation Details".

The comparison of aggregation strategies in Table 12 reveals that selective aggregation methods like top_k_mean tend to outperform simpler approaches for inflammation detection across all models, likely because they can focus on the most informative regions of the slide.

Table 10: Performance Comparison of ResNet18 Models for Gastric Inflammation Classification Across Studies

	Current Work	**Previous Work**†	
Study	**This Study**	**Hempel (2023)**	**Höfling (2023)**
Validation Performance			
Tile-Level Accuracy	82.19%	—	92.41%
Tile-Level Sensitivity	81.65%	—	—
Tile-Level Specificity	83.03%	—	—
Tile-Level F1 Score	84.70%	—	—
Tile-Level AUC	91.19%	—	—
Slide-Level Accuracy	100.00%	—	—
Slide-Level Sensitivity	100.00%	—	—
Slide-Level Specificity	100.00%	—	—
Slide-Level F1 Score	100.00%	—	—
Slide-Level AUC	100.00%	—	—
Test Performance			
Tile-Level Accuracy	81.70%	90.30%	72.73%
Tile-Level Sensitivity	80.42%	88.30%	67.90%
Tile-Level Specificity	84.24%	94.70%	88.10%
Tile-Level F1 Score	85.39%	92.60%	79.08%
Tile-Level AUC	90.43%	98.00%	86.00%
Slide-Level Accuracy	95.35%	100.00%	94.74%
Slide-Level Sensitivity	100.00%	100.00%	—
Slide-Level Specificity	89.47%	100.00%	—
Slide-Level F1 Score	96.00%	100.00%	—
Slide-Level AUC	99.56%	100.00%	97.00%
Aggregation Strategy			
Strategy	top_k_mean_30	majority vote	probabilistic avg
Implementation Details			
Batch Size	128	32	128
Dropout rate	0.000	0.000	0.000
Epochs	25	10	25
Learning Rate	1e-3	1e-4	1e-3
Weight Decay	0.000	0.000	0.000
Positive Class Weight	1.0	1.0	1.0
Optimizer	SGD	SGD	SGD

† Direct numerical comparison is not straightforward due to different dataset splits.

Note: All results are from ResNet18 models.

Note: Missing values (—) indicate metrics not reported in previous studies.

Table 11: Comprehensive Performance Comparison of Deep Learning Models for Gastric Inflammation Classification (Scanner1)

Model	**ConvNeXt**	**DenseNet121**	**DenseNet169**	**GigaPath**	**ResNet18**	**Swin-V2**
Metric	**Large**					**Base**
Validation Performance						
Tile-Level Accuracy	83.95%	86.06%	81.16%	82.57%	82.19%	74.21%
Tile-Level Sensitivity	82.16%	85.61%	79.55%	82.08%	81.65%	75.58%
Tile-Level Specificity	86.67%	86.75%	83.62%	83.32%	83.03%	72.13%
Tile-Level F1 Score	86.07%	88.12%	83.60%	85.05%	84.70%	77.96%
Tile-Level AUC	91.78%	93.61%	89.56%	90.61%	91.19%	81.34%
Slide-Level Accuracy	95.00%	97.37%	97.37%	97.50%	100.00%	92.24%
Slide-Level Sensitivity	100.00%	94.74%	94.74%	100.00%	100.00%	89.47%
Slide-Level Specificity	90.00%	100.00%	100.00%	95.00%	100.00%	95.00%
Slide-Level F1 Score	95.00%	97.30%	97.30%	97.44%	100.00%	91.89%
Slide-Level AUC	99.74%	100.00%	99.74%	100.00%	100.00%	97.11%
Test Performance						
Tile-Level Accuracy	83.73%	76.93%	82.76%	80.44%	81.70%	75.88%
Tile-Level Sensitivity	84.65%	68.88%	82.63%	77.79%	80.42%	74.92%
Tile-Level Specificity	81.89%	92.92%	83.02%	85.70%	84.24%	77.79%
Tile-Level F1 Score	87.37%	79.89%	86.44%	84.11%	85.39%	80.52%
Tile-Level AUC	91.93%	91.20%	91.36%	90.29%	90.43%	83.63%
Slide-Level Accuracy	90.70%	95.35%	93.02%	95.35%	95.35%	90.70%
Slide-Level Sensitivity	95.83%	91.67%	91.67%	100.00%	100.00%	91.67%
Slide-Level Specificity	84.21%	100.00%	94.74%	89.47%	89.47%	89.47%
Slide-Level F1 Score	92.00%	95.65%	93.62%	96.00%	96.00%	91.67%
Slide-Level AUC	97.37%	99.56%	99.56%	99.12%	99.56%	95.18%
Optimal Aggregation Strategy						
Strategy	top_k_mean_10	mean	top_k_mean_30	top_k_mean_30	top_k_mean_30	mean
Implementation Details						
Batch Size	64	32	64	64	128	96
Dropout Rate	0.049	0.161	0.228	0.208	0.000	0.286
Epochs	50	50	50	50	25	50
Learning Rate	4.117e-4	1.777e-4	1.860e-4	9.145e-4	1.000e-3	3.887e-4
Weight Decay	8.027e-5	3.438e-5	1.462e-4	2.035e-4	0.000	5.240e-5
Positive Class Weight	0.572	0.413	0.417	0.465	1.000	0.574
Optimizer	AdamW	AdamW	AdamW	AdamW	SGD	AdamW

Note: All metrics are calculated on the test datasets using consistent methodology. Optimal aggregation strategies were determined through validation data optimization. Note: Slide-level refers to WSI, where all tiles from the same slide share the same inflammation label. Aggregation strategies determine how tile-level predictions are combined to make slide-level decisions.

Table 12: Comparison of Aggregation Strategies based on Validation Set for Gastric Inflammation Classification Across Models

Model	Strategy	Threshold	Balanced Accuracy	Sensitivity	Specificity	F1 Score
ConvNeXt Large	mean	0.428	94.87%	94.74%	95.00%	94.74%
	median	0.526	94.74%	89.47%	100.00%	94.44%
	top_k_mean_10	0.900	**95.00%**	100.00%	90.00%	95.00%
	top_k_mean_20	0.870	94.87%	94.74%	95.00%	94.74%
	top_k_mean_30	0.782	94.87%	94.74%	95.00%	94.74%
DenseNet121	**mean**	0.691	**97.37%**	94.74%	100.00%	97.30%
	median	0.815	97.37%	94.74%	100.00%	97.30%
	top_k_mean_10	0.900	90.00%	100.00%	80.00%	90.48%
	top_k_mean_20	0.900	92.50%	100.00%	85.00%	92.68%
	top_k_mean_30	0.900	95.00%	100.00%	90.00%	95.00%
DenseNet169	mean	0.435	94.87%	94.74%	95.00%	94.74%
	median	0.340	94.87%	94.74%	95.00%	94.74%
	top_k_mean_10	0.900	85.00%	100.00%	70.00%	86.36%
	top_k_mean_20	0.900	95.00%	100.00%	90.00%	95.00%
	top_k_mean_30	0.900	**97.37%**	94.74%	100.00%	97.30%
GigaPath	mean	0.587	97.37%	94.74%	100.00%	97.30%
	median	0.614	97.37%	94.74%	100.00%	97.30%
	top_k_mean_10	0.900	92.50%	100.00%	85.00%	92.68%
	top_k_mean_20	0.900	92.50%	100.00%	85.00%	92.68%
	top_k_mean_30	0.900	**97.50%**	100.00%	95.00%	97.44%
ResNet18	mean	0.617	97.37%	94.74%	100.00%	97.30%
	median	0.665	97.37%	94.74%	100.00%	97.30%
	top_k_mean_10	0.900	90.00%	100.00%	80.00%	90.48%
	top_k_mean_20	0.900	97.50%	100.00%	95.00%	97.44%
	top_k_mean_30	0.900	**100.00%**	100.00%	100.00%	100.00%
Swin_v2_b	**mean**	0.494	**92.24%**	89.47%	95.00%	91.89%
	median	0.478	89.74%	89.47%	90.00%	89.47%
	top_k_mean_10	0.818	89.61%	84.21%	95.00%	88.89%
	top_k_mean_20	0.732	89.74%	89.47%	90.00%	89.47%
	top_k_mean_30	0.741	92.11%	84.21%	100.00%	91.43%

Note: Aggregation strategies represent different methods for combining tile-level predictions to make slide-level decisions.
Values shown are from the aggregation strategies based on the validation set performance.
Threshold values were optimized on validation data to maximize balanced accuracy.
"mean": Average of all tile predictions within a slide
"median": Median of all tile predictions within a slide
"top_k_mean_n": Average of the top n% tile predictions within a slide

9.3 Cross-Scanner Performance (Scanner2): Tissue Classification

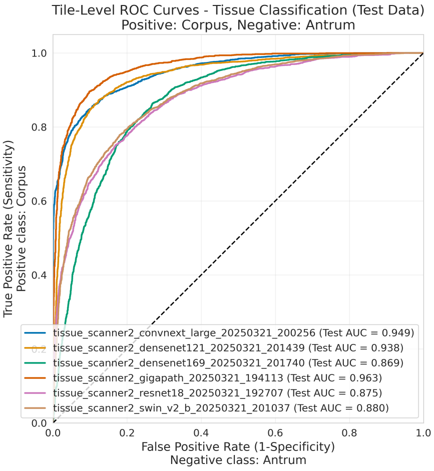


Figure 15: ROC curves comparing model performance for tissue classification at the tile-level on Scanner2 test-set

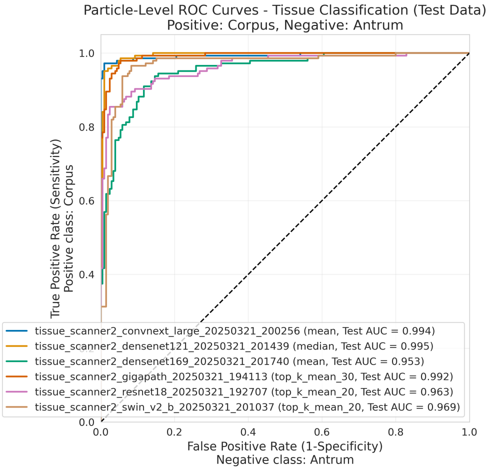


Figure 16: ROC curves comparing model performance for tissue classification at the particle-level on Scanner2 test-set

This section evaluates the models' ability to generalize to a different scanner environment for tissue classification. Using the same validation-optimized thresholds from Scanner1, we assess how well these models perform when applied to Scanner2 data without recalibration. Tables 13 and 14 present comprehensive metrics for this cross-scanner evaluation, while Figures 15 and 16 visualize the ROC curves for tile-level and particle-level performance respectively.

Table 13: Cross-Scanner Generalization for Gastric Tissue Classification - Part 1

	ConvNeXt Large		**DenseNet121**		**DenseNet169**	
Metric	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**
Validation Performance						
Tile-Level Accuracy	90.47%	—	89.09%	—	87.99%	—
Tile-Level Sensitivity	93.75%	—	92.24%	—	89.28%	—
Tile-Level Specificity	85.63%	—	84.44%	—	86.09%	—
Tile-Level F1 Score	92.13%	—	90.97%	—	89.85%	—
Tile-Level AUC	92.91%	—	94.26%	—	94.05%	—
Particle-Level Accuracy	96.35%	—	94.62%	—	95.45%	—
Particle-Level Sensitivity	98.28%	—	94.83%	—	97.41%	—
Particle-Level Specificity	94.42%	—	94.42%	—	93.49%	—
Particle-Level F1 Score	96.61%	—	94.83%	—	95.76%	—
Particle-Level AUC	97.05%	—	98.06%	—	98.15%	—
Test Performance						
Tile-Level Accuracy	88.71%	79.80%	87.07%	80.40%	84.23%	66.65%
Tile-Level Sensitivity	87.06%	56.09%	84.60%	60.04%	84.16%	32.03%
Tile-Level Specificity	90.28%	99.81%	89.42%	97.57%	84.29%	95.86%
Tile-Level F1 Score	88.27%	71.77%	86.46%	73.70%	83.89%	46.78%
Tile-Level AUC	91.83%	94.91%	90.52%	93.81%	88.79%	86.87%
Particle-Level Accuracy	92.01%	84.24%	90.91%	81.09%	91.46%	65.62%
Particle-Level Sensitivity	90.18%	61.81%	87.73%	54.17%	89.57%	16.67%
Particle-Level Specificity	93.50%	100.00%	93.50%	100.00%	93.00%	100.00%
Particle-Level F1 Score	91.02%	76.39%	89.66%	70.27%	90.40%	28.57%
Particle-Level AUC	93.61%	99.36%	92.43%	99.48%	91.15%	95.27%
Generalization Gap*	-5.75%		-7.05%		-4.12%	
Optimal Aggregation Strategy						
Strategy	mean		median		mean	

* Generalization Gap = Scanner1 Aggregated AUC - Scanner2 Aggregated AUC. Negative values indicate better performance on Scanner2.

Note: Aggregation strategy optimized on validation data from Scanner1, then applied to both test datasets.

Note: Scanner1 results are from the held-out test set of the same scanner used for training. Scanner2 results are from a completely different scanner representing real-world domain shift.

Note: Missing values (—) denote the absence of validation metrics for Scanner2, as it is exclusively used as a test set.

Table 14: Cross-Scanner Generalization for Gastric Tissue Classification - Part 2

	GigaPath		**ResNet18**		**Swin_v2_b**	
Metric	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**
Validation Performance						
Tile-Level Accuracy	87.74%	—	86.81%	—	81.18%	—
Tile-Level Sensitivity	90.45%	—	88.76%	—	81.70%	—
Tile-Level Specificity	83.75%	—	83.95%	—	80.42%	—
Tile-Level F1 Score	89.78%	—	88.91%	—	83.80%	—
Tile-Level AUC	93.04%	—	92.23%	—	86.97%	—
Particle-Level Accuracy	96.33%	—	95.47%	—	93.91%	—
Particle-Level Sensitivity	98.71%	—	96.98%	—	95.26%	—
Particle-Level Specificity	93.95%	—	93.95%	—	92.56%	—
Particle-Level F1 Score	96.62%	—	95.74%	—	94.24%	—
Particle-Level AUC	97.70%	—	97.38%	—	95.96%	—
Test Performance						
Tile-Level Accuracy	84.30%	79.92%	85.20%	66.18%	78.41%	74.47%
Tile-Level Sensitivity	76.46%	57.34%	85.32%	27.12%	68.91%	48.86%
Tile-Level Specificity	91.76%	98.97%	85.08%	99.13%	87.46%	96.08%
Tile-Level F1 Score	82.62%	72.33%	84.91%	42.33%	75.69%	63.65%
Tile-Level AUC	89.63%	96.26%	89.20%	87.48%	84.55%	88.01%
Particle-Level Accuracy	90.36%	91.98%	91.46%	62.75%	91.46%	87.39%
Particle-Level Sensitivity	87.12%	81.94%	90.80%	9.72%	90.80%	73.61%
Particle-Level Specificity	93.00%	99.02%	92.00%	100.00%	92.00%	97.07%
Particle-Level F1 Score	89.03%	89.39%	90.52%	17.72%	90.52%	82.81%
Particle-Level AUC	92.46%	99.23%	92.03%	96.31%	92.33%	96.89%
Generalization Gap*	-6.77%		-4.28%		-4.56%	
Optimal Aggregation Strategy						
Strategy	top_k_mean_30		top_k_mean_20		top_k_mean_20	

* Generalization Gap = Scanner1 Aggregated AUC - Scanner2 Aggregated AUC. Negative values indicate better performance on Scanner2.

Note: Aggregation strategy optimized on validation data from Scanner1, then applied to both test datasets.

Note: Scanner1 results are from the held-out test set of the same scanner used for training. Scanner2 results are from a completely different scanner representing real-world domain shift.

Note: Missing values (—) denote the absence of validation metrics for Scanner2, as it is exclusively used as a test set.

9.4 Cross-Scanner Performance (Scanner2): Inflammation Classification

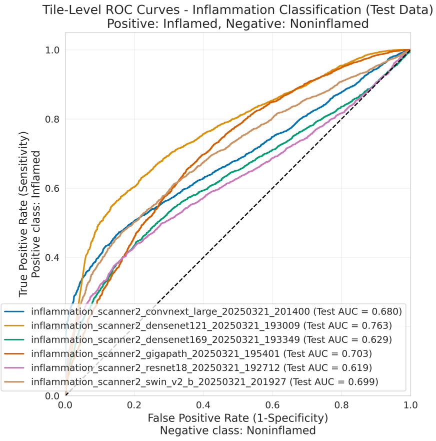


Figure 17: ROC curves comparing model performance for inflammation classification at the tile-level on Scanner2 test-set

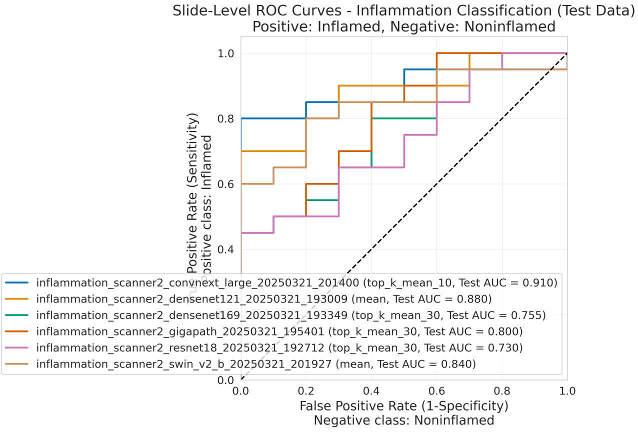


Figure 18: ROC curves comparing model performance for inflammation classification at the slide-level on Scanner2 test-set

Following our assessment of tissue classification, we evaluate the cross-scanner generalization capabilities of our models for inflammation detection. This analysis applies Scanner1-optimized thresholds and aggregation strategies to Scanner2 data, mirroring a real-world clinical scenario where models trained on one scanner are deployed to another without recalibration. The performance metrics are presented in Tables 15 and 16, with ROC curves for tile-level and slide-level performance illustrated in Figures 17 and 18 respectively.

Table 15: Cross-Scanner Generalization for Gastric Inflammation Classification - Part 1

	ConvNeXt Large		**DenseNet121**		**DenseNet169**	
Metric	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**
Validation Performance						
Tile-Level Accuracy	83.95%	—	86.06%	—	81.16%	—
Tile-Level Sensitivity	82.16%	—	85.61%	—	79.55%	—
Tile-Level Specificity	86.67%	—	86.75%	—	83.62%	—
Tile-Level F1 Score	86.07%	—	88.12%	—	83.60%	—
Tile-Level AUC	91.78%	—	93.61%	—	89.56%	—
Slide-Level Accuracy	95.00%	—	97.37%	—	97.37%	—
Slide-Level Sensitivity	100.00%	—	94.74%	—	94.74%	—
Slide-Level Specificity	90.00%	—	100.00%	—	100.00%	—
Slide-Level F1 Score	95.00%	—	97.30%	—	97.30%	—
Slide-Level AUC	99.74%	—	100.00%	—	99.74%	—
Test Performance						
Tile-Level Accuracy	83.73%	68.11%	76.93%	71.12%	82.76%	60.51%
Tile-Level Sensitivity	84.65%	99.50%	68.88%	93.92%	82.63%	66.71%
Tile-Level Specificity	81.89%	1.78%	92.92%	22.97%	83.02%	47.40%
Tile-Level F1 Score	87.37%	80.90%	79.89%	81.53%	86.44%	69.63%
Tile-Level AUC	91.93%	67.97%	91.20%	76.31%	91.36%	62.88%
Slide-Level Accuracy	90.70%	66.67%	95.35%	76.67%	93.02%	66.67%
Slide-Level Sensitivity	95.83%	100.00%	91.67%	100.00%	91.67%	95.00%
Slide-Level Specificity	84.21%	0.00%	100.00%	30.00%	94.74%	10.00%
Slide-Level F1 Score	92.00%	80.00%	95.65%	85.11%	93.62%	79.17%
Slide-Level AUC	97.37%	91.00%	99.56%	88.00%	99.56%	75.50%
Generalization Gap*	6.37%		11.56%		24.06%	
Optimal Aggregation Strategy						
Strategy	top_k_mean_10		mean		top_k_mean_30	

* Generalization Gap = Scanner1 Aggregated AUC - Scanner2 Aggregated AUC. Negative values indicate better performance on Scanner2.

Note: Aggregation strategy optimized on validation data from Scanner1, then applied to both test datasets.

Note: Scanner1 results are from the held-out test set of the same scanner used for training. Scanner2 results are from a completely different scanner representing real-world domain shift.

Note: Missing values (—) denote the absence of validation metrics for Scanner2, as it is exclusively used as a test set.

Table 16: Cross-Scanner Generalization for Gastric Inflammation Classification - Part 2

	GigaPath		ResNet18		Swin_v2_b	
Metric	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**	**Scanner1**	**Scanner2**
Validation Performance						
Tile-Level Accuracy	82.57%	—	82.19%	—	74.21%	—
Tile-Level Sensitivity	82.08%	—	81.65%	—	75.58%	—
Tile-Level Specificity	83.32%	—	83.03%	—	72.13%	—
Tile-Level F1 Score	85.05%	—	84.70%	—	77.96%	—
Tile-Level AUC	90.61%	—	91.19%	—	81.34%	—
Slide-Level Accuracy	97.50%	—	100.00%	—	92.24%	—
Slide-Level Sensitivity	100.00%	—	100.00%	—	89.47%	—
Slide-Level Specificity	95.00%	—	100.00%	—	95.00%	—
Slide-Level F1 Score	97.44%	—	100.00%	—	91.89%	—
Slide-Level AUC	100.00%	—	100.00%	—	97.11%	—
Test Performance						
Tile-Level Accuracy	80.44%	70.95%	81.70%	64.75%	75.88%	67.78%
Tile-Level Sensitivity	77.79%	91.07%	80.42%	91.05%	74.92%	94.56%
Tile-Level Specificity	85.70%	28.44%	84.24%	9.21%	77.79%	11.21%
Tile-Level F1 Score	84.11%	80.97%	85.39%	77.81%	80.52%	79.94%
Tile-Level AUC	90.29%	70.29%	90.43%	61.85%	83.63%	69.87%
Slide-Level Accuracy	95.35%	66.67%	95.35%	66.67%	90.70%	66.67%
Slide-Level Sensitivity	100.00%	100.00%	100.00%	100.00%	91.67%	100.00%
Slide-Level Specificity	89.47%	0.00%	89.47%	0.00%	89.47%	0.00%
Slide-Level F1 Score	96.00%	80.00%	96.00%	80.00%	91.67%	80.00%
Slide-Level AUC	99.12%	80.00%	99.56%	73.00%	95.18%	84.00%
Slide-Level Generalization Gap*	19.12%		26.56%		11.18%	
Optimal Aggregation Strategy						
Strategy	top_k_mean_30		top_k_mean_30		mean	

* Generalization Gap = Scanner1 Aggregated AUC - Scanner2 Aggregated AUC. Negative values indicate better performance on Scanner2.

Note: Aggregation strategy optimized on validation data from Scanner1, then applied to both test datasets.

Note: Scanner1 results are from the held-out test set of the same scanner used for training. Scanner2 results are from a completely different scanner representing real-world domain shift.

Note: Missing values (—) denote the absence of validation metrics for Scanner2, as it is exclusively used as a test set.

10 Discussion

10.1 Within-Scanner Performance (Scanner1)

Our experiments on Scanner1 demonstrate robust performance across multiple architectures for tissue and inflammation classification tasks, as evidenced by the comprehensive metrics in Tables 8 and 11. These results reflect the models' ability to generalize within the same scanner domain, leveraging optimized hyperparameters determined through Bayesian optimization with Optuna (Akiba et al., 2019).

10.1.1 Tissue Classification

ConvNeXt Large emerges as the top performer for tissue classification, achieving a particle-level test accuracy of 92.01% and an AUC of 93.61% (Table 8). ConvNeXt Large outperforms simpler models like ResNet18 (test particle-level AUC of 92.03%) by approximately 1.58 percentage points, as visualized in the ROC curves at tile and particle-levels (Figures 11 and 12). The superior performance of ConvNeXt Large is likely attributable to its advanced convolutional design, which enhances feature extraction from histological images. GigaPath also performs competitively with a particle-level test AUC of 92.46%, slightly outperforming ResNet18, while DenseNet169 shows slightly lower performance with a particle-level test AUC of 91.15%. These results suggest that modern convolutional architectures and deeper networks can effectively capture tissue-specific patterns.

10.1.2 Inflammation Classification

Multiple models achieved excellent slide-level test performance for inflammation classification, with DenseNet121, DenseNet169, and ResNet18 all reaching an AUC of 99.56% (Table 11). GigaPath follows closely with a slide-level test AUC of 99.12%, while ConvNeXt Large (97.37%) and Swin-V2 Base (95.18%) show slightly lower performance. The slide-level test accuracy results follow a similar pattern, with DenseNet121, GigaPath, and ResNet18 all achieving 95.35%, followed by DenseNet169 (93.02%), ConvNeXt Large and Swin-V2 Base at 90.70%. These results are confirmed by the ROC curves (Figures 13 and 14), where multiple models exhibit near-perfect discrimination.

This pattern differs from tissue classification, where ConvNeXt Large led the performance metrics. For inflammation detection, the simpler ResNet18 achieves comparable performance to more complex architectures, suggesting that inflammation features may be more readily detectable without requiring advanced architectural design.

10.1.3 Aggregation Strategies

The impact of different aggregation strategies varies significantly between models, as detailed in Tables 9 and 12. For tissue classification, ConvNeXt Large achieves optimal performance with a simple ‘mean‘ strategy (balanced validation accuracy 96.35%), while GigaPath and ResNet18 benefit from more selective approaches like ‘top_k_mean_30‘ (96.33%) and ‘top_k_mean_20‘ (95.47%), respectively. This suggests that focusing on high-confidence predictions can enhance model robustness.

For inflammation classification, selective aggregation dominates across most models, with ‘top_k_mean‘ variants showing the best performance. ResNet18 achieves perfect validation balanced accuracy (100.00%) with ‘top_k_mean_30‘, while GigaPath (97.50%) and DenseNet169 (97.37%) also excel with this approach. These results indicate that slide-level inflammation classification benefits substantially from focusing on the most informative regions of a slide, which aligns with clinical intuition where inflammation often presents in localized areas rather than uniformly across the entire slide (Pennelli et al., 2020; Sepulveda and Patil, 2008b).

10.2 Cross-Scanner Generalization (Scanner2)

Cross-scanner generalization tests the real-world applicability of our models by evaluating performance on Scanner2, a distinct domain from the training Scanner1. Results in Tables 13–16 reveal varying degrees of domain adaptation, visualized in Figures 15–18.

10.2.1 Tissue Classification

Our models demonstrate remarkable cross-scanner generalization for tissue classification, with all models performing better on Scanner2 than on their original training domain. This unexpected finding is evidenced by the negative generalization gaps on the test sets across all models, ranging from -4.12% (DenseNet169) to -7.05% (DenseNet121), as shown in Tables 13 and 14.

GigaPath exhibits exceptional cross-domain performance with a particle-level test AUC of 99.23% on Scanner2, considerably higher than its 92.46% on Scanner1, yielding a negative generalization gap of -6.77%. ConvNeXt Large and DenseNet121 also demonstrate strong generalization capabilities with Scanner2 test AUCs of 99.36% and 99.48%, respectively. These results suggest that these architectures effectively capture scanner-invariant tissue features, likely due to their sophisticated feature extraction capabilities and, in GigaPath’s case, possible benefits from its pretraining on diverse pathology datasets.

However, a notable pattern emerges when examining sensitivity and specificity metrics. Across all models, we observe substantial decreases in tile-level test sensitivity on Scanner2 compared to Scanner1 (e.g., from 90.18% to 61.81% for ConvNeXt

Large, and most dramatically from 90.80% to 9.72% for ResNet18), offset by near-perfect specificity values (approaching 100% for most models). This significant imbalance suggests a shift in classification thresholds when transitioning between scanner domains, potentially driven by differences in staining intensity and image characteristics between scanners.

This trade-off is most pronounced with ResNet18, which, despite achieving a relatively good Scanner2 AUC of 96.31%, shows a sensitivity drop to just 9.72%. Such extreme performance characteristics indicate that while the model can rank predictions effectively (as reflected in the high AUC), its decision boundary is severely miscalibrated for the new domain. Swin_v2_b maintains a more balanced performance with a Scanner2 sensitivity of 73.61% and AUC of 96.89%, suggesting its transformer-based architecture may provide some inherent robustness to domain shifts.

The negative generalization gaps observed across all architectures highlight an intriguing phenomenon wherein models trained on Scanner1 exhibit enhanced discriminative power on Scanner2 data. This could potentially be attributed to Scanner2's image characteristics providing greater contrast between tissue types, making the classification task easier despite the domain shift.

10.2.2 Inflammation Classification

In stark contrast to tissue classification, inflammation classification presents a significantly more challenging cross-scanner generalization scenario. All models demonstrate substantial performance degradation when applied to Scanner2, with positive generalization gaps ranging from 6.37% (ConvNeXt Large) to 26.56% (ResNet18) on the test sets, as detailed in Tables 15 and 16.

At the slide-level, ConvNeXt Large maintains the strongest cross-domain performance with an AUC of 91.00% on the Scanner2 test set compared to 97.37% on the Scanner1 test set, representing the smallest generalization gap of 6.37%. DenseNet121 follows with a Scanner2 AUC of 88.00%, while GigaPath (80.00%), Swin_v2_b (84.00%), DenseNet169 (75.50%), and ResNet18 (73.00%) show more significant performance drops. These results suggest that detecting inflammation features is susceptible to scanner-specific characteristics, with more complex architectures like ConvNeXt Large demonstrating greater resilience to these domain shifts.

A concerning pattern emerges in the test specificity metrics across all models. At the slide-level, test specificity values on Scanner2 drop dramatically, reaching as low as 0.00% for several models, including ResNet18, GigaPath, and Swin_v2_b. This is accompanied by consistently high sensitivity values (95.00%-100.00%), indicating that models systematically overpredict inflammation on Scanner2 samples. This severe decision boundary shift results in models effectively classifying nearly all Scanner2 samples as inflamed regardless of their true label.

The tile-level test performance mirrors this trend, with specificity values ranging from just 1.78% (ConvNeXt Large) to 47.40% (DenseNet169), while sensitivities remain high (66.71%-99.50%). This pervasive misclassification pattern suggests that Scanner2's image characteristics likely contain features that closely resemble inflammation signals learned from Scanner1, leading to widespread false positive predictions.

These findings have significant implications for clinical deployment, indicating that inflammation models trained on a single scanner would require substantial domain adaptation before reliable use on images from different scanning devices. This is consistent with the findings of Howard et al. (2021), who demonstrated that site-specific digital histology signatures can lead to biased accuracy in deep learning models, highlighting the importance of accounting for scanner-specific variations. Furthermore, Shi et al. (2022) successfully applied domain adaptation techniques to improve cross-hospital diagnosis of gastric dysplasia, suggesting that similar approaches could benefit cross-scanner generalizability in our context. The stark difference in cross-scanner generalization between tissue and inflammation classification suggests that tissue morphology features are inherently more scanner-agnostic than the more subtle textural and intensity patterns that characterize inflammation.

10.3 Comparative Analysis with Previous Studies

This thesis builds on previous studies from 2023, offering insights into the progression of deep learning for gastric histological classification. Direct comparisons with prior work by Höfling (2023) and Hempel (2023), as detailed in Tables 7 and 10, must account for methodological differences, including dataset refinements (e.g., corrected annotations) and varied data splits in the current study.

Comparisons with the previous studies (Table 7) show that our tissue classification ResNet18 model (tile-level test accuracy 85.20%) is slightly outperformed by Hempel (2023) (89.10%) and Höfling (2023) (87.00%). However, direct comparisons are confounded by differences in dataset splits and classification scope, as well as our binary corpus/antrum task versus their multi-class approaches including intermediate types. Our study extends beyond prior work by reporting particle-level metrics, offering a more granular assessment of model performance critical for clinical applications.

Compared to prior work (Table 10), our inflammation classification ResNet18 tile-level accuracy (81.70%) is lower than Hempel's (90.30%) but higher than Höfling's (72.73%). At the slide-level, our model's test accuracy (95.35%) approaches Hempel's (100.00%) and slightly exceeds Höfling's (94.74%). These comparisons highlight the advantage of our hierarchical evaluation, where slide-level performance can outshine tile-level metrics, aligning with clinical needs for whole-slide diagnoses.

Höfling (2023) established human expert benchmarks, reporting tile-level test accuracies of 80.00-86.44% for tissue classification and 70.00-73.33% for inflammation classification. These benchmarks contextualize our computational results: our

ResNet18 achieves a tile-level test accuracy of 85.20% for tissue (Table 7) and 81.70% for inflammation (Table 10), surpassing human performance in the latter and approaching the upper bound in the former. Slide-level accuracies (e.g., 95.35% for ResNet18 inflammation) further suggest that our models may exceed human capabilities at the WSI level, a critical advancement for clinical diagnostics.

The architectural evolution across studies is notable. Höfling relied solely on ResNet18, while Hempel demonstrated improved performance with the same architecture. Our study extends this by evaluating ResNet18 alongside advanced models like ConvNeXt Large and GigaPath on a refined dataset. For tissue classification, ConvNeXt Large achieves a tile-level test accuracy of 88.71% and specificity of 90.28%, compared to ResNet18's 85.20% and 85.08%, and Hempel's 89.10% and 88.00% (Table 7). For inflammation, ConvNeXt Large reaches 83.73% tile-level accuracy, surpassing Höfling's 72.73% but trailing Hempel's 90.30% (Table 10). These gains reflect dataset improvements and architectural advancements, though comparisons remain nuanced due to differing classification scopes on the tissue classification task (binary vs. multi-class).

While Höfling's work showed lower tile-level test performance for inflammation classification (72.73% vs. Hempel's 90.30%), both achieved comparable slide-level test accuracy (94.74% vs. 100.00%), highlighting the critical role of effective aggregation strategies. Our selective 'top_k_mean' approach (yielding 95.35% slide-level accuracy) offers a middle ground between Höfling's probabilistic averaging and Hempel's majority voting, with stronger robustness to outlier tiles.

Particle-level tissue test classification shows consistent robustness, with accuracies ranging from 90.36% (GigaPath) to 92.01% (ConvNeXt Large) on Scanner1, aligning with the range implied by prior confusion matrices. This stability underscores the effectiveness of our hierarchical approach across architectures.

Our contributions advance gastric histopathology analysis in several key ways:

- **Cross-Scanner Validation:** Unlike prior single-scanner evaluations, our cross-scanner analysis (Tables 13–16) exposes real-world deployment challenges, a methodological leap.
- **Architecture Exploration:** Beyond ResNet18, we demonstrate ConvNeXt Large's superior feature extraction (e.g., AUC 91.83% for tissue) and GigaPath's scanner-agnostic robustness (e.g., negative generalization gap -6.77%), insights unavailable from earlier single-architecture studies.
- **Aggregation Strategy Optimization:** Our systematic evaluation of aggregation strategies, e.g., 'top_k_mean_30' yielding 100.00% balanced accuracy for ResNet18 inflammation, offers a novel framework for hierarchical prediction that builds upon the probabilistic approach (Höfling, 2023).
- **Task-Specific Deployment Guidelines:** The divergent generalization patterns (e.g., tissue robustness vs. inflammation sensitivity, Section 10.2) sug-

gest tailored clinical strategies, providing critical insights not available in prior studies' uniform approaches.

These findings affirm that modern architectures can match or exceed human-level performance at tile and WSI levels, with ConvNeXt Large showing particular promise. Coupled with dataset refinements and selective aggregation strategies, our work systematically enhances histopathological classification, paving the way for robust computer-aided diagnosis in clinical practice.

10.4 Augmentation Strategies

We implemented and evaluated five distinct augmentation strategies to enhance model robustness, particularly for cross-scanner generalization: MedMNIST, ColorJitter, ModelConfig, Medmnist_ColorJitter (combined approach), and NormalizationOnly. Each approach addresses different aspects of histopathological image variability, to improve performance across scanners.

10.4.1 Theoretical Foundation and Implementation

Our approach to stain augmentation draws inspiration from prior work by Tellez et al. (2018, 2019), who established that direct manipulation of HE color channels generates diverse, realistic variations that substantially enhance cross-center generalization in computational pathology. Similarly, Salvo et al. (2024) demonstrated that domain-specific augmentations, tailored to address modality-specific artifacts, consistently outperform generic augmentation strategies across a range of imaging modalities.

We implemented the following augmentation variants:

- **MedMNIST**: Leveraging the MedMNISTC framework, this approach applied pathology-specific corruptions derived from the CORRUPTIONS_DS registry for PathMNIST (Salvo et al., 2024).
- **ColorJitter**: This strategy implemented a novel hematoxylin and eosin-specific color transformation based on a stain separation matrix. The `StainColorJitter` class performs the following operations:
    1. Converts RGB values to the optical density domain using a logarithmic transformation.
    2. Applies the inverse of a predefined stain matrix M to separate the hematoxylin and eosin components.
    3. Introduces controlled perturbations through scaling (α) and shifting (β) operations with stochasticity controlled by σ (set to 0.05).
    4. Recombines the perturbed stain components and transforms back to the RGB space.

- **Medmnist_ColorJitter**: This combined approach sequentially applies both the MedMNIST pathology-specific corruptions (Salvo et al., 2024) and the StainColorJitter transformation, potentially offering complementary benefits of both domain-specific augmentation techniques.

- **NormalizationOnly**: This approach serves as a controlled comparison, applying only standard normalization using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) without additional augmentations.

- **ModelConfig**: The baseline approach used the standard augmentation configuration specified in the model configuration files, including random resized cropping (scale 0.8-1.0), random horizontal and vertical flips, and standard normalization.

10.4.2 Experimental Configuration

The augmentation strategies were systematically evaluated using DenseNet121 as the base architecture to ensure controlled comparison. For each augmentation variant, we maintained identical hyperparameters derived from our Bayesian optimization process:

- **For inflammation classification:** Batch size of 32, dropout rate of 0.161, AdamW optimizer with learning rate of 0.00018, weight decay of 3.44e-05, and positive class weight of 0.413.

- **For tissue classification:** Batch size of 64, dropout rate of 0.061, AdamW optimizer with learning rate of 0.00019, weight decay of 1.08e-05, and positive class weight of 0.654.

All models were trained using early stopping with a patience of 10 epochs and cosine annealing learning rate scheduling to ensure fair comparison across augmentation strategies. To maintain experimental integrity, we standardized the validation process by using pre-computed thresholds and maintaining consistent aggregation strategies when evaluating performance.

Experiments were structured to assess both within-scanner performance (Scanner1 test set) and cross-scanner generalization (Scanner2 test set) across both classification tasks. To ensure comprehensive evaluation, we assessed performance at multiple hierarchical levels:

- Tile-level metrics (accuracy, sensitivity, specificity, F1 score, AUC).

- Aggregated metrics (particle-level for tissue classification, slide-level for inflammation).

For both tasks, we implemented rigorous evaluation procedures by adopting previously determined optimal thresholds from our DenseNet121 experiments (as documented in Tables 9 and 12). These validation-optimized thresholds were applied consistently to test data across all augmentation variants, ensuring a realistic and fair assessment of model generalization capabilities.

Note on Training Time: Due to time constraints, all models were trained for only 20 epochs, which may result in slightly different performance compared to previous DenseNet121 "ModelConfig" results that used more extensive training. However, the relative performance across augmentation strategies still provides valuable insights into their effectiveness for WSI images, particularly for H&E stained slides.

10.4.3 Results and Analysis

Our experimental results reveal that augmentation strategies have differential effects based on the classification task and evaluation context:

Tissue Classification:

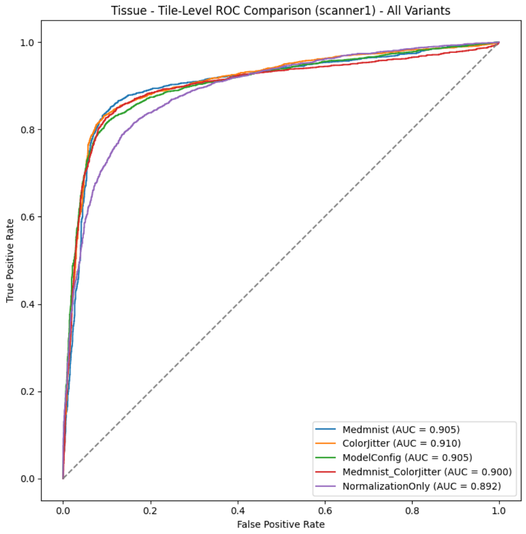


Figure 19: ROC curves comparing augmentation results for tissue classification at the tile-level on Scanner1 test-set

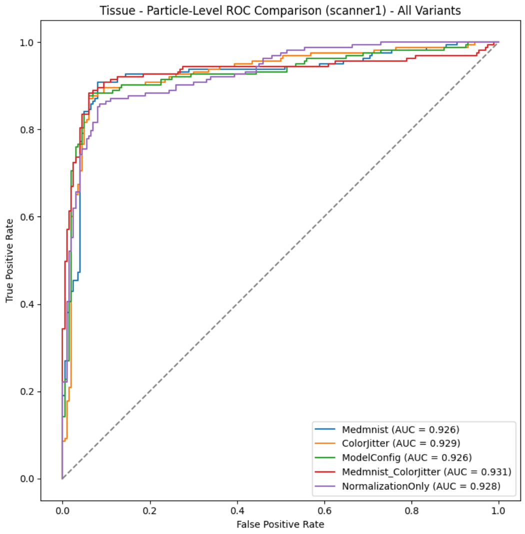


Figure 20: ROC curves comparing augmentation results for tissue classification at the particle-level on Scanner1 test-set

- Within Scanner1, the **Medmnist_ColorJitter** augmentation demonstrated the highest particle-level test AUC (93.13%), slightly outperforming the baseline ModelConfig approach (92.61% test AUC), also outperforming the particle-level test accuracy (91.46% vs. ModelConfig's 90.91%).

- For cross-scanner generalization to Scanner2, **Medmnist_ColorJitter** augmentation again showed the strongest performance, with a particle-level AUC of 100.00% and particle-level accuracy of 96.56%.

Table 17: Comprehensive Results for Tissue Classification Augmentation Results on Scanner1 Test-Set

Metric	**Medmnist**	**ColorJitter**	**ModelConfig**	**Medmnist_ColorJitter**	**NormalizationOnly**
Test Performance					
Tile-level Accuracy	87.07%	86.20%	85.62%	85.57%	82.67%
Tile-level Sensitivity	84.45%	78.81%	82.75%	86.08%	79.21%
Tile-level Specificity	89.56%	93.25%	88.35%	85.08%	85.98%
Tile-level F1 Score	86.44%	84.79%	84.89%	85.34%	81.69%
Tile-level AUC	90.48%	90.95%	90.49%	90.03%	89.16%
Particle-level Accuracy	89.81%	89.26%	90.91%	91.46%	87.33%
Particle-level Sensitivity	85.28%	83.44%	88.34%	88.34%	79.75%
Particle-level Specificity	93.50%	94.00%	93.00%	94.00%	93.50%
Particle-level F1 Score	88.25%	87.46%	89.72%	90.28%	84.97%
Particle-level AUC	92.59%	92.88%	92.61%	93.13%	92.77%

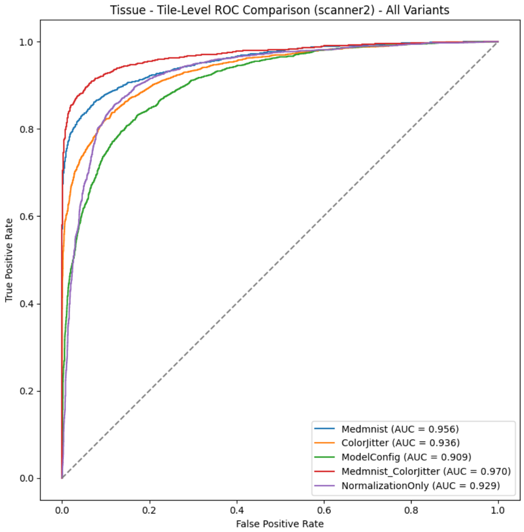


Figure 21: ROC curves comparing augmentation results for tissue classification at the tile-level on Scanner2 test-set

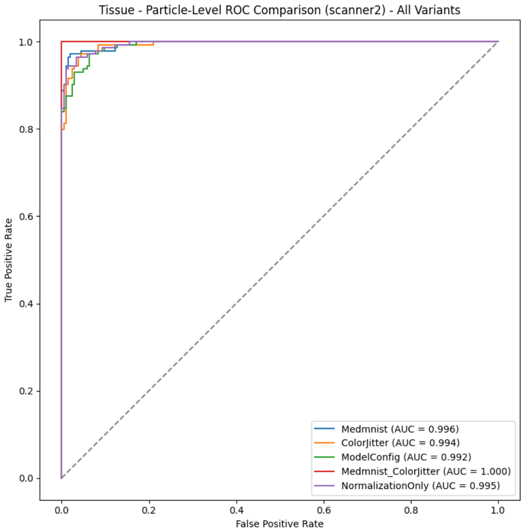


Figure 22: ROC curves comparing augmentation results for tissue classification at the particle-level on Scanner2 test-set

Table 18: Comprehensive Results for Tissue Classification Augmentation Results on Scanner2 Test-Set

Metric	**Medmnist**	**ColorJitter**	**ModelConfig**	**Medmnist_ColorJitter**	**NormalizationOnly**
Test Performance					
Tile-level Accuracy	87.89%	81.12%	75.36%	91.88%	76.65%
Tile-level Sensitivity	74.65%	59.82%	48.86%	84.35%	52.36%
Tile-level Specificity	99.07%	99.10%	97.73%	98.23%	97.14%
Tile-level F1 Score	84.95%	74.36%	64.48%	90.48%	67.24%
Tile-level AUC	95.58%	93.60%	90.92%	97.04%	92.88%
Particle-level Accuracy	91.12%	79.94%	73.93%	96.56%	71.06%
Particle-level Sensitivity	78.47%	51.39%	36.81%	91.67%	29.86%
Particle-level Specificity	100.00%	100.00%	100.00%	100.00%	100.00%
Particle-level F1 Score	87.94%	67.89%	53.81%	95.65%	45.99%
Particle-level AUC	99.59%	99.39%	99.24%	100.0%	99.51%

- Interestingly, the **NormalizationOnly** approach showed a nuanced pattern: it performed similar to the **Medmnist_ColorJitter** on Scanner1 and maintained excellent discriminative power on Scanner2 (99.51% particle-level test AUC) despite lower particle-level test accuracy (71.06%). This suggests that while the model can distinguish between tissue types across scanners, the optimal decision threshold shifts significantly depending on the augmentation strategy. This highlights that threshold recalibration, rather than complete retraining, might be sufficient for cross-scanner adaptation with simple normalization for tissue classification tasks.

Inflammation Classification:

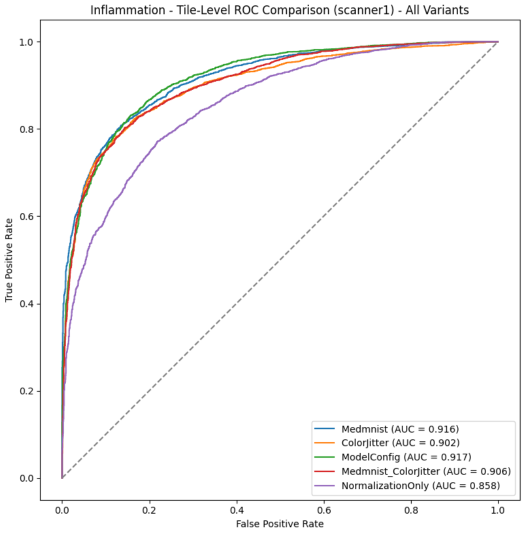


Figure 23: ROC curves comparing augmentation results for inflammation classification at the tile-level on Scanner1 test-set

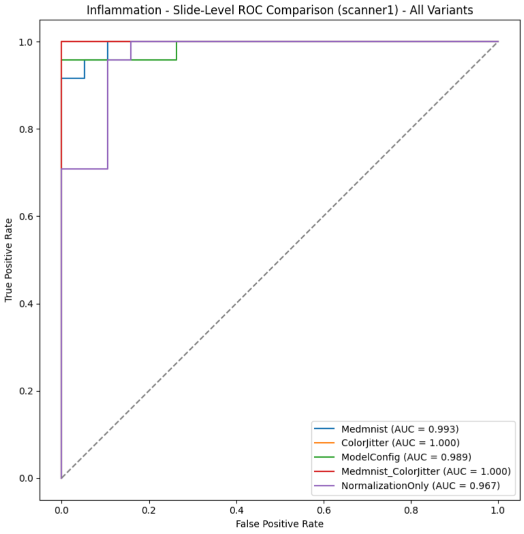


Figure 24: ROC curves comparing augmentation results for inflammation classification at the slide-level on Scanner1 test-set

Table 19: Comprehensive Results for Inflammation Classification Augmentation Results on Scanner1 Test-Set

Metric	**Medmnist**	**ColorJitter**	**ModelConfig**	**Medmnist_ColorJitter**	**NormalizationOnly**
Test Performance					
Tile-level Accuracy	72.09%	68.38%	73.24%	74.04%	63.14%
Tile-level Sensitivity	59.49%	53.89%	61.88%	63.18%	46.75%
Tile-level Specificity	97.12%	97.17%	95.80%	95.61%	95.71%
Tile-level F1 Score	73.93%	69.40%	75.46%	76.40%	62.78%
Tile-level AUC	91.62%	90.22%	91.70%	90.60%	85.80%
Slide-level Accuracy	86.05%	81.40%	83.72%	83.72%	79.07%
Slide-level Sensitivity	75.00%	66.67%	70.83%	70.83%	62.50%
Slide-level Specificity	100.00%	100.00%	100.00%	100.00%	100.00%
Slide-level F1 Score	85.71%	80.00%	82.93%	82.93%	76.92%
Slide-level AUC	99.34%	100.00%	98.90%	100.00%	96.71%

- Within Scanner1, **Medmnist_ColorJitter** and **ColorJitter** achieved perfect slide-level AUC (100.00%), with **Medmnist** showing the highest slide-

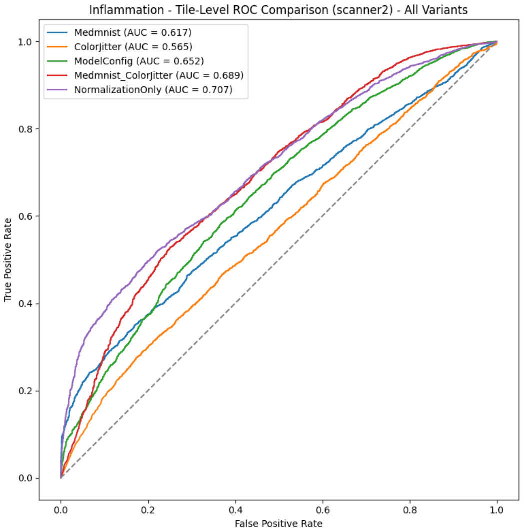


Figure 25: ROC curves comparing augmentation results for inflammation classification at the tile-level on Scanner2 test-set

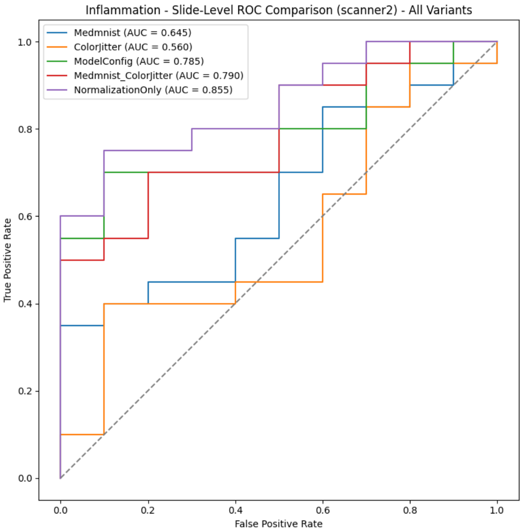


Figure 26: ROC curves comparing augmentation results for inflammation classification at the slide-level on Scanner2 test-set

Table 20: Comprehensive Results for Inflammation Classification Augmentation Results on Scanner2 Test-Set

Metric	Medmnist	ColorJitter	ModelConfig	Medmnist_ColorJitter	NormalizationOnly
Test Performance					
Tile-level Accuracy	59.85%	62.22%	51.69%	66.25%	59.85%
Tile-level Sensitivity	65.13%	79.72%	38.96%	73.25%	51.05%
Tile-level Specificity	48.68%	25.25%	78.58%	51.46%	78.44%
Tile-level F1 Score	68.77%	74.12%	52.26%	74.66%	63.31%
Tile-level AUC	61.66%	56.53%	65.16%	68.92%	70.66%
Slide-level Accuracy	63.33%	66.67%	53.33%	76.67%	70.00%
Slide-level Sensitivity	75.00%	95.00%	30.00%	90.00%	60.00%
Slide-level Specificity	40.00%	10.00%	100.00%	50.00%	90.00%
Slide-level F1 Score	73.17%	79.17%	46.15%	83.72%	72.73%
Slide-level AUC	64.50%	56.00%	78.50%	79.00%	85.50%

level test accuracy (86.05%) compared to the baseline ModelConfig approach (98.90% AUC, 83.72% accuracy).

- For cross-scanner inflammation detection, **NormalizationOnly** achieved the highest slide-level AUC (85.50%), while **Medmnist_ColorJitter** demonstrated the best slide-level accuracy (76.67%).

The combination of **Medmnist** and **ColorJitter** augmentations (i.e., **Medmnist_ColorJitter**) yielded the most balanced performance across both tasks and scanner environments, suggesting that domain-specific augmentations targeting both morphological variations and stain characteristics are essential for robust cross-scanner performance.

These findings align with the principles established by Tellez et al. (2018), confirming that stain-specific augmentations particularly benefit inflammation detection by normalizing color variations across scanners. Our results further validate Salvo et al. (2024)'s observation that domain-specific augmentation strategies consistently outperform generic approaches, with the performance differential being most pronounced in cross-scanner generalization scenarios.

10.5 Limitations and Improvements

This thesis advances beyond prior work by addressing several limitations, yet it is not without constraints that warrant consideration. Below, we outline how we have improved upon previous research and identify remaining challenges, preempting potential critiques by providing context and mitigation strategies.

- **Scanner Variability:** Unlike single-scanner studies (Höfling, 2023; Hempel, 2023), we incorporate data from Scanner1 and Scanner2 (Tables 13–16), enabling a more realistic assessment of generalizability across imaging devices. This addresses a key gap in prior work, where scanner-specific biases remained untested.
- **Annotation Quality:** We systematically corrected annotation errors, enhancing dataset reliability over earlier studies that relied on noisier labels.
- **Architectural Exploration:** By evaluating five architectures (e.g., ConvNeXt Large, GigaPath, ResNet18) and the GigaPath model, we offer a broader performance spectrum than the ResNet18-centric focus of previous investigations.
- **Aggregation Analysis:** Our novel analysis of aggregation strategies (Tables 9 and 12) provides actionable insights into optimizing hierarchical predictions, absent in prior work.

Despite these advancements, several limitations persist, which we address proactively to guide interpretation and future research:

- **Scanner Diversity and Domain Adaptation:** A common critique might highlight our testing on only two scanners, limiting broader generalization to diverse manufacturers or imaging conditions. Chen et al. (2024) provide a comprehensive review, emphasizing the need for domain adaptation to ensure model robustness across diverse imaging environments, which aligns with our findings. This is evident in Scanner2's performance drops (e.g., ResNet18 inflammation tile-level AUC falls to 61.85%, Table 16), with generalization gaps up to 26.56%. We did not implement domain adaptation techniques in the main study, such as stain normalization (Tellez et al., 2021) or adversarial training (Ganin et al., 2016), to align feature distributions across scanners.

This omission likely exacerbates sensitivity-specificity shifts (e.g., ResNet18 specificity drops from 84.24% to 9.21%, Table 16), reflecting potentially unaddressed scanner-specific artifacts (e.g., color or resolution variances). Future work could mitigate this by applying unsupervised domain adaptation, potentially reducing gaps by aligning Scanner2's intensity distributions to Scanner1's, as demonstrated in similar histopathology studies (Stacke et al., 2019). While our study highlights the need for models that generalize across different scanners, future work could explore domain adaptation techniques to mitigate these challenges. For instance, Shi et al. (2022) demonstrated the use of deep learning with domain adaptation to improve cross-hospital diagnosis of gastric dysplasia, which could be adapted to our context. Additionally, a comprehensive review of AI applications in digital pathology for gastric cancer Chen et al. (2024) suggests that integrating such techniques could significantly enhance model performance and clinical applicability.

- **Sample Size and Statistical Power:** The Scanner2 test set, particularly for inflammation classification (30 slides), may draw scrutiny for its limited size, potentially inflating generalization gaps (e.g., 26.56% for ResNet18) and constraining statistical power. While Scanner1's larger dataset supports robust within-scanner results, expanding Scanner2's sample size (e.g., to >100 slides) would enhance confidence in cross-scanner findings, a priority for future validation.

- **Binary Classification Constraints:** Reviewers may note that the binary tasks (corpus/antrum, inflamed/non-inflamed) oversimplify clinical reality, omitting intermediate tissue types (e.g., transitional zones) and inflammation severity grades (e.g., mild, moderate, severe per Sydney System (Dixon et al., 1996)). For instance, our models cannot distinguish subtle gastritis subtypes critical for treatment decisions, limiting diagnostic granularity. A multi-class approach, though computationally intensive, could leverage GigaPath's high AUC (99.23% on Scanner2, Table 14) to address this, albeit requiring expanded annotations.

- **Lack of Interpretability:** A potential critique is the absence of explainability tools (e.g., saliency maps (Selvaraju et al., 2017), SHAP values (Lundberg and Lee, 2017)) to elucidate model decisions. While ConvNeXt Large excels (Scanner1 test tile-level AUC 91.83%, Table 8), its black-box nature may undermine clinical trust. Integrating attention mechanisms could reveal whether high performance stems from pathologically relevant features or scanner-specific noise, enhancing adoption.

- **External Validation:** Our reliance on internal datasets might prompt concerns about external generalizability. Without testing on independent cohorts (e.g., from different institutions), Scanner2 results (e.g., GigaPath's -6.77% gap) may reflect dataset-specific biases rather than universal scanner robustness. Collaborating with external pathology centers for validation would address this, though resource constraints precluded such efforts here.

To improve, we propose expanding scanner diversity (e.g., including scanners from different manufacturers), increasing Scanner2 sample sizes, and transitioning to multi-class frameworks. Implementing domain adaptation, such as cycle-consistent adversarial networks (Zhu et al., 2020), could mitigate cross-scanner degradation, while attention-based visualizations would enhance interpretability. These steps, grounded in our findings (e.g., Figures 15-18), aim to bridge current gaps and tackle concerns, ensuring robustness for clinical translation.

10.6 Clinical Implications and Future Directions

Our models demonstrate significant potential for transforming pathology practice, with high accuracies in tissue and inflammation classification, 92.01% particle-level accuracy for tissue on Scanner1 test data using ConvNeXt Large (Table 8) and 95.35% slide-level accuracy for inflammation on Scanner1 test data using DenseNet121, GigaPath, or ResNet18 (Table 11). These results suggest several clinical benefits:

- **Diagnostic Workflow Enhancement:** High accuracies enable initial screening, prioritizing urgent cases and reducing turnaround times amid rising workloads.
- **Standardization of Assessment:** Objective predictions, with slide-level AUCs nearing 100% for inflammation classification on Scanner2 test data (Figures 14), can minimize inter-observer variability in gastritis diagnosis, enhancing consistency.
- **Resource Optimization:** Robust tissue classification, exemplified by GigaPath's negative generalization gap of -6.77% (Table 14), supports triage in resource-scarce settings, optimizing expert review.

However, while expert-level performance is achieved within the same scanner domain (Scanner1), cross-scanner challenges, particularly for inflammation detection (Section 10.2), necessitate careful consideration for deployment. The stark difference in generalization between tissue and inflammation classification highlights the need for task-specific strategies:

- **Task-Specific Deployment:** Tissue classification systems appear more immediately deployable across varied scanner environments, whereas inflammation detection requires sophisticated domain adaptation techniques to ensure reliability.
- **Foundation Model Advantage:** The exceptional cross-scanner performance of GigaPath suggests that histopathology-specific foundation models may provide a more robust starting point for clinical applications than general computer vision architectures.

- **Threshold Recalibration:** High AUC values maintained across scanners indicate that collecting a small calibration set from each new scanner to optimize decision thresholds could be a practical solution for clinical deployment.

To address these challenges and build on our findings, future work should focus on three key areas:

- **Technical Improvements:**
  - Developing specialized domain adaptation techniques, such as stain normalization or adversarial training, and implementing scanner harmonization to improve cross-scanner generalization, particularly for inflammation detection (e.g., addressing ResNet18's 26.56% slide-level inflammation classification generalization gap).
  - Integrating explainable AI methods, such as attention mechanisms (Vaswani et al., 2023), to enhance model interpretability and foster clinical trust (e.g., visualizing features from Figure 11).
- **Clinical Relevance:**
  - Moving beyond binary classification to capture inflammation severity and intermediate tissue types for greater diagnostic granularity.
  - Developing integrated systems that combine multiple diagnostic tasks (e.g., tissue classification and inflammation detection) into unified platforms.
- **Validation and Deployment:**
  - Expanding evaluations to additional scanner types and manufacturers to assess broader generalization capabilities.
  - Conducting prospective clinical validation studies to quantify the impact on diagnostic workflows and patient outcomes in real-world settings.

10.7 Conclusion

Our comprehensive evaluation of deep learning approaches for gastric histopathology classification has yielded significant insights into both architectural performance and cross-scanner generalization capabilities. By systematically assessing multiple state-of-the-art models across two distinct classification tasks with a rigorously curated dataset, we have established several key findings:

1. **Architectural Performance**: We demonstrated that modern convolutional architectures, particularly ConvNeXt Large, excel at tissue classification (92.01% particle-level accuracy, 93.61% AUC), outperforming traditional models like ResNet18. For inflammation detection, multiple architectures achieved

excellent slide-level performance (DenseNet121, DenseNet169, and ResNet18 all reaching 99.56% AUC), suggesting that inflammation features may be more readily detectable without requiring complex architectural designs.

2. **Task-Specific Aggregation Strategies**: Our systematic evaluation revealed that optimal prediction aggregation varies by task and model architecture. While tissue classification with ConvNeXt Large benefited from simple mean aggregation (96.35% validation accuracy), inflammation detection consistently improved with selective aggregation approaches like `top_k_mean_30`, which reached 100% validation accuracy with ResNet18. This pattern aligns with the clinical understanding that inflammation often presents in localized regions rather than uniformly across slides.

3. **Divergent Cross-Scanner Generalization**: Perhaps our most significant finding was the stark contrast in cross-scanner generalization between tasks. Tissue classification exhibited remarkable robustness, with all models performing better on previously unseen Scanner2 data than on their training domain (negative generalization gaps ranging from -4.12% to -7.05%). GigaPath demonstrated exceptional cross-scanner capabilities with a particle-level AUC of 99.23% on Scanner2. Conversely, inflammation classification suffered severe degradation when deployed cross-scanner, with specificity collapsing to near-zero levels despite maintained sensitivity, resulting in generalization gaps up to 26.56%.

4. **Foundation Model Advantage**: The histopathology-specific foundation model GigaPath consistently demonstrated superior cross-scanner robustness, particularly for tissue classification (-6.77% generalization gap), vastly outperforming general vision models. This suggests that domain-specific pretraining on diverse histopathological data confers significant advantages for real-world deployment scenarios.

5. **Domain-Specific Augmentation Benefits**: Our augmentation experiments revealed that a combined approach of pathology-specific corruptions with stain-specific color transformations (Medmnist_ColorJitter) yielded the most balanced performance across both tasks and scanner environments. This supports prior research indicating that domain-specific augmentations targeting both morphological variations and stain characteristics are essential for robust cross-scanner performance.

These findings have important implications for clinical implementation. Tissue classification systems appear more immediately deployable across varied scanner environments, while inflammation detection may require either sophisticated domain adaptation techniques or scanner-specific recalibration. The foundation model approach exemplified by GigaPath offers a promising direction for developing scanner-agnostic models.

Future work should focus on expanding scanner diversity to assess broader generalization capabilities, implementing targeted domain adaptation techniques for inflammation detection, transitioning to multi-class frameworks that capture intermediate tissue types and inflammation severity gradients, and integrating explainable AI methods to enhance clinical interpretability and trust.

In conclusion, while we have demonstrated that deep learning can achieve expert-level performance for gastric histopathology classification within controlled environments, achieving robust performance across diverse scanning devices remains a significant challenge, particularly for inflammation detection. By identifying these task-dependent generalization challenges and evaluating potential mitigation strategies, our work provides a foundation for developing computational pathology tools that can be reliably deployed in varied clinical settings.

A Appendix

A.1 Dataset and Code Availability

To facilitate reproducibility and further research in this area, we have made all materials used in this study publicly accessible. The complete dataset, including all images and annotations, is available for download through our institutional repository. The implementation code, including all analysis scripts and documentation, has been released under the MIT License and can be accessed via our GitHub repository at `https://www.github.com/DominicLiebel/GSDB`. We encourage the scientific community to build upon our work.

A.2 Original Dataset Naming Schemes

The dataset contains slides from two different scanners, each initially using distinct and less systematic naming conventions before the implementation of our unified naming scheme (`[slideID]_[patientID]_[scanner]_[stain]`):

A.2.1 Scanner1 Naming Scheme

Slides from the first scanner followed a naming scheme:

- Basic format: `[NUMBER][TYPE][STAIN]`, e.g., `1BHE`
- `HE`: Hematoxylin and Eosin stain
- `PAS`: Periodic acid–Schiff stain
- `modGiem`: Modified Giemsa stain
- Type indicators:
  - No letter: Non-inflamed tissue
  - `B`: Type B gastritis
  - `C`: Type C gastritis
  - `S`: Other inflammation types
  - `K`: Non-classifiable tissue

These slides were specifically prepared for the project from archived patient samples, ensuring consistent preparation by the same medical technical assistant. The original glass slides remain available for potential rescanning.

A.2.2 Scanner2 Naming Scheme

The second scanner dataset originally used a sequential numbering scheme:

- Format: `[NUMBER]-[STAIN] [TYPE]`, e.g., `201-1 c`
- Each case includes three stains: HE, PAS, and modified Giemsa (indicated by -1, -2, -3 respectively)
- Inflammation indicators:
  - `c`: Type C gastritis
  - `b`: Type B gastritis
  - `ef`: Non-inflamed tissue

This dataset contains 90 total scans from 30 patients, with each case represented by three staining methods (HE, PAS, modified Giemsa), resulting in 30 scans per staining type.

The limitations of these initial naming schemes, particularly their inconsistency and lack of explicit patient-slide relationships, motivated the development of our current unified naming convention. For this work, we introduced this new, more systematic naming convention to improve:

- Reproducibility: Clear patient and slide associations
- Data organization: Structured identifiers for database management
- Cross-scanner analysis: Scanner-specific identifiers for studying scanner effects
- Stain tracking: Explicit stain type in filename

A.2.3 Dataset Identifier Mapping Protocol

To ensure systematic data management and reproducibility, we implemented an identifier mapping protocol for the dataset$_{IT,a}^{S12-HE,PAS,MG}$. The protocol establishes a bidirectional mapping between legacy identifiers and standardized naming convention, facilitating consistent data organization while maintaining traceability to original annotations.

Mapping Implementation The mapping schema was implemented as a CSV file, maintained at `data/raw/archive/mapping.csv`. The mapping follows a one-to-one correspondence principle with two primary fields:

- `old_id`: Legacy identifier from original data collection
- `slide_name`: Standardized identifier following the convention defined in Section 4.1

Example Mapping Entries Table 21 presents representative examples from the mapping schema, illustrating the transformation from legacy to standardized identifiers across different staining types.

Table 21: Example entries from the Slide Identifier Mapping Protocol

old_id	slide_name
1BHE	1_1_1_HE
1CHE	2_2_1_HE
1CmodGiem	3_2_1_MG
1CPAS	4_2_1_PAS
...	...

Applications and Significance The mapping protocol primarily serves as a traceability mechanism between the original slide identifiers and our standardized naming convention. By maintaining a clear record of identifier transformations, researchers can trace any slide back to its original designation in the source dataset. This documentation is particularly crucial for verifying data processing steps and validating analyses against the original expert annotations. The systematic approach ensures that despite the implementation of a more structured naming convention, no historical context or reference capability is lost. This transparency in data transformation supports auditability of the research pipeline while enabling efficient cross-referencing between the original and standardized naming systems.

A.3 Dataset Notation Convention Examples

$Dataset_{IT,a}^{S12-HE,PAS,MG}$:	Initial Cohort
$Dataset_{IT,v}^{S12-HE,PAS,MG}$:	Valid Annotation Cohort
$Dataset_{IT,v}^{S1-HE,PAS,MG}$:	Scanner1 Cohort
$Dataset_{IT,v}^{S1-HE}$:	Scanner1 HE Staining Cohort
$Dataset_{T,v}^{S1-HE}$:	Scanner1 Tissue Analysis Cohort
$Dataset_{I,v}^{S1-HE}$:	Scanner1 Inflammation Analysis Cohort
$Dataset_{IT,a}^{S2-HE,PAS,MG}$:	Scanner2 Cohort
$Dataset_{IT,v}^{S2-HE,PAS,MG}$:	Scanner2 Valid Annotation Cohort
$Dataset_{IT,v}^{S2-HE}$:	Scanner2 HE Staining Cohort
$Dataset_{T,v}^{S2-HE}$:	Scanner2 Tissue Analysis Cohort
$Dataset_{I,v}^{S2-HE}$:	Scanner2 Inflammation Analysis Cohort

A.4 Annotation Software Pipeline

The annotation process involves two key software components:

A.4.1 Slide Preparation

Prior to annotation, slides are downsampled using a Python script that leverages the OpenSlide library. The script processes MRXS format slides with a downsample factor of 16, converting them to more manageable PNG files while maintaining sufficient detail for accurate annotation. Key features include:

- Downsampling MRXS files with specified factor
- Conversion to RGB color space
- Progressive processing with error handling
- Comprehensive logging system

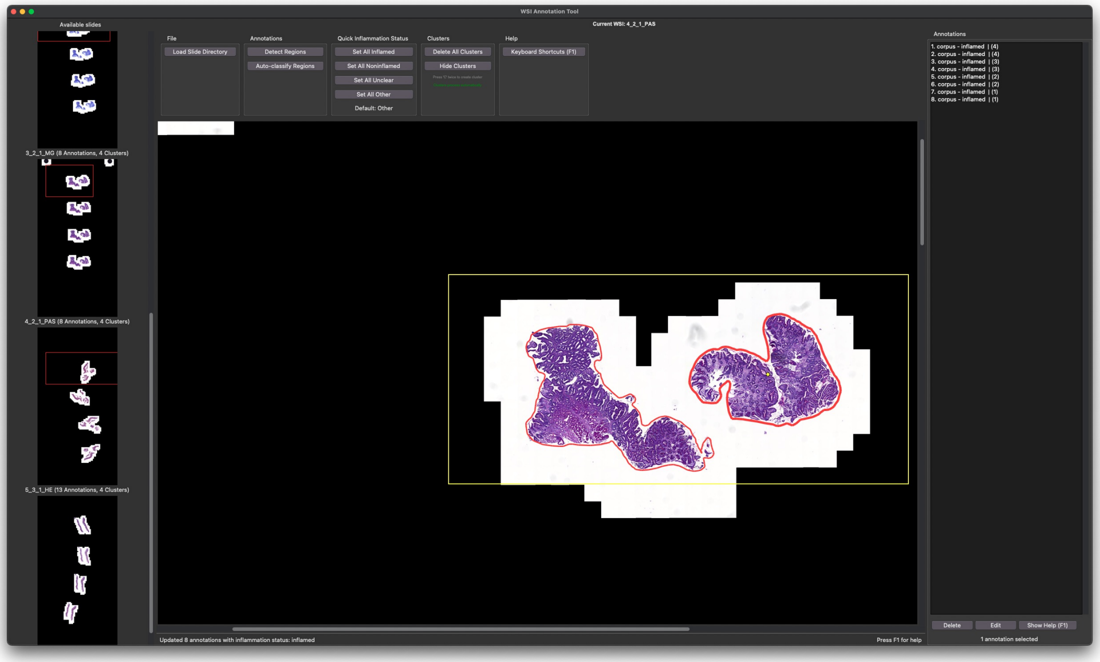
A.4.2 Annotation Tool

Annotations are created using a custom-developed Python-based tool that provides a graphical user interface for precise particle marking. The tool features:

- Interactive polygon drawing interface
- Automatic detection of particles
- Support for multiple tissue types and inflammation states
- GeoJSON format export
- Automatic backup system
- Cluster management
- Multi-level undo/redo functionality

The left panel shows available slides with their thumbnail previews and annotation counts. The main viewport displays the current slide with manually drawn annotations (red outlines) marking distinct tissue particles. The yellow rectangle indicates a cluster. The right panel lists all annotations for the current slide, showing tissue type (corpus) and inflammation status (inflamed) along with their cluster assignments (numbers in parentheses).

Full documentation and installation instructions are maintained in the repository.



Current WSI: 4_2_1_PAS

1. corpus - inflamed | (4)
2. corpus - inflamed | (4)
3. corpus - inflamed | (3)
4. corpus - inflamed | (3)
5. corpus - inflamed | (2)
6. corpus - inflamed | (2)
7. corpus - inflamed | (1)
8. corpus - inflamed | (1)

Updated 8 annotations with inflammation status: inflamed

1 annotation selected

Figure 27: Screenshot of the WSI Annotation Tool interface showing the interactive polygon drawing system

A.5 Valid Annotations Example

Example: Consider slide 93_67_1_HE which demonstrates independent validity for both inflammation and tissue tasks:

Inflammation Task Validity:

- Slide is classified as ”inflamed” (valid inflammation status)
- All particles and their tiles inherit this inflammation status:
  - 4 corpus particles → used in inflammation analysis
  - 8 antrum particles → used in inflammation analysis
  - 4 intermediate particles → used in inflammation analysis, as inflammation status is independent of tissue type

Tissue Task Validity:

- Contains:
  - 4 corpus particles → used in tissue analysis
  - 8 antrum particles → used in tissue analysis
  - 4 intermediate particles → reserved for future analysis (outside thesis scope)

- Valid for tissue analysis because it contains at least one corpus or antrum annotated particle

Key Point: The 4 intermediate particles:

- Are included in inflammation analysis (inherit slide's "inflamed" status)
- Are not included in the tissue analysis of this thesis, which focuses specifically on binary classification between corpus and antrum. These intermediate tissue regions represent a valuable dataset for future multi-class classification models or transitional zone analysis

Table 22: Annotation Analysis Validity Example (Slide 93_67_1_HE)

Particle ID	Inflammation	Valid for Inflammation	Tissue Type	Valid for tissue*
c909943[...]	inflamed	✓	antrum	✓
0deed90[...]	inflamed	✓	antrum	✓
fae3b18[...]	inflamed	✓	antrum	✓
f48e8c0[...]	inflamed	✓	antrum	✓
bec7ad1[...]	inflamed	✓	antrum	✓
968cc72[...]	inflamed	✓	antrum	✓
d98c507[...]	inflamed	✓	antrum	✓
2599d8f[...]	inflamed	✓	antrum	✓
b6952e0[...]	inflamed	✓	corpus	✓
e7c4343[...]	inflamed	✓	corpus	✓
fef9075[...]	inflamed	✓	corpus	✓
e596f28[...]	inflamed	✓	corpus	✓
c49c66a[...]	inflamed	✓	intermediate	†
65c11ab[...]	inflamed	✓	intermediate	†
cf226a9[...]	inflamed	✓	intermediate	†
d4c48b7[...]	inflamed	✓	intermediate	†

Note: All particles inherit slide's inflammation status ("inflamed")

* Valid for tissue column reflects thesis scope (corpus/antrum binary classification)

† Intermediate tissue particles are preserved in the dataset for future research

A.6 Patient Exclusion

Prior to dataset splitting, 22 patients were removed based on expert pathological review by Dr. Bettina Braunecker and Dr. med. Volker Mordstein (Table 23). These patients were excluded as they could compromise both primary analysis and generalizability testing. These quality control exclusions were applied before any subsequent steps to ensure a foundation of high-quality data for all analyses.

Table 23: Patients excluded from splitting

Patient ID	Exclusion Reason
4	No gastric biopsies but rather biopsies from other organs
9	No gastric biopsies but rather biopsies from other organs
14	No gastric biopsies but rather biopsies from other organs
18	Non-representative sample
24	No gastric biopsies but rather biopsies from other organs
29	No gastric biopsies but rather biopsies from other organs
34	No gastric biopsies but rather biopsies from other organs
39	No gastric biopsies but rather biopsies from other organs
44	No gastric biopsies but rather biopsies from other organs
49	No gastric biopsies but rather biopsies from other organs
54	No gastric biopsies but rather biopsies from other organs
59	No gastric biopsies but rather biopsies from other organs
64	No gastric biopsies but rather biopsies from other organs
69	No gastric biopsies but rather biopsies from other organs
74	No gastric biopsies but rather biopsies from other organs
75	No gastric biopsies but rather biopsies from other organs
79	No gastric biopsies but rather biopsies from other organs
83	No gastric biopsies but rather biopsies from other organs
180	Non-representative sample
197	Non-representative sample
213	Too superficial for reliable classification
217	Non-representative sample

Note: These patients were excluded based on expert pathological review by Dr. Bettina Braunecker and Dr. med. Volker Mordstein. These cases were removed prior to creating splits to ensure data quality.

A.7 Thesis Specific Excluded Patient Analysis

This section provides a comprehensive analysis of patients excluded from the main study, ensuring transparency and reproducibility of the patient selection process. Each exclusion category is documented with its specific criteria and rationale.

A.7.1 Scanner Selection Process

The initial dataset included patients from two distinct scanning devices (Scanner1 and Scanner2). To maintain methodological consistency and minimize technical variability, only Scanner1 patients were used for the primary analysis. Scanner2 patients (n=30) were reserved for subsequent generalizability testing.

Table 24: Scanner2 Dataset Distribution

Stain Type	Patients	Slides
HE staining	10	30
PAS staining	10	30
MG staining	10	30
Total	30	90

Note: Each patient (n=30, patientIDs 245-274) has a complete set of three stains
Scanner2 patients were reserved for future generalizability testing

A.7.2 Patients Missing HE Staining

From the Scanner1 cohort, 12 patients were excluded due to having only non-HE stained slides. These cases comprised:

Table 25: Scanner1 non-HE stained patients

Patient ID	Stain Type
70	PAS
80	PAS
181	PAS
183	PAS
185	PAS
234	MG
239	PAS
240	PAS
241	PAS
242	PAS
243	PAS
244	PAS

Note: These patients were excluded from analysis as they lacked HE staining.

A.7.3 Patients Missing Valid Tissue Annotations

Scanner1 HE patients lacking valid tissue type annotations:

Table 26: Patients with No Valid Tissue (HE Scanner1)

Patient ID	Exclusion Reason
62	Esophageal mucosa, tissue type unclassifiable
73	Intermediate tissue
104	Intermediate tissue
110	Intermediate tissue
147	Intermediate tissue
165	Esophageal mucosa, tissue type unclassifiable
229	Intermediate tissue
232	Intermediate tissue
236	Partially blurry, intermediate tissue

Note: This table shows HE slides that lack valid tissue type annotations but have inflammation annotations.

A.7.4 Patients Missing Valid Inflammation Annotations

Scanner1 HE patients lacking valid inflammation status annotations:

Table 27: Patients with No Valid Inflammation (HE Scanner1)

Patient ID	Exclusion Reason
5	Glandular cysts, unclear inflammation
25	Glandular cysts, unclear inflammation
30	Glandular cysts, unclear inflammation
60	Non-representative inflammation
65	Non-representative inflammation
204	Glandular cysts, unclear inflammation

Note: This table contains HE slides that lack valid inflammation annotations but have tissue type annotations.

A.7.5 Summary of Excluded Patients

Table 28: Summary of Excluded Patients

Exclusion Reason	Patient Count
Missing valid annotations	22
Scanner2 (reserved for generalizability testing)	30
Missing HE staining	12
Missing valid tissue annotations	9
Missing valid inflammation annotations	6
Total Excluded	79
Fully Included	195

A.8 Impact of Random Seeds on Dataset Distribution

Tables 29, 30, and 31 demonstrate the effect of different random seeds on class distribution across dataset splits. These tables illustrate why seed 44 was selected as the optimal seed for the data splitting process. When comparing the distribution of inflammation and tissue types across train, validation, and test sets, seed 44 provides the most balanced allocation.

It should be noted that the percentages in these tables do not sum to 100% because the "Other" category tiles have been omitted for clarity. For inflammation classification, the remaining 2-3% of tiles belong to slides with ambiguous inflammation status. Similarly, for tissue classification, the remaining 8-10% represent intermediate zones and non-classifiable tissue regions that don't clearly belong to either corpus or antrum categories. These "Other" category particles/slides were excluded from model training and evaluation but are included in the total tile counts for completeness.

Table 29: Seed: 42 Dataset Summary Statistics (Dataset$^{S12-HE}_{IT,v}$)

Characteristic	Total	Train	Val	Test
TILE-LEVEL				
Total Tiles	40,997	19,787	6,915	7,439
Inflammation Classification				
Inflamed Tiles	25,875 (63.1%)	12,227 (61.8%)	4,812 (69.6%)	4,181 (56.2%)
Non-inflamed Tiles	14,250 (34.8%)	7,319 (37.0)%	2,103 (30.4%)	2,627 (35.3%)
Tissue Classification				
Corpus Tiles	20,903 (51.0%)	9,693 (49.0%)	3,285 (47.5%)	5,197 (69.86%)
Antrum Tiles	16,787 (40.9%)	8,211 (41.5%)	3,366 (48.7%)	1,995 (26.8%)

Table 30: Seed: 43 Dataset Summary Statistics (Dataset$^{S12-HE}_{IT,v}$)

Characteristic	Total	Train	Val	Test
TILE-LEVEL				
Total Tiles	40,997	19,295	8,015	6,831
Inflammation Classification				
Inflamed Tiles	25,875 (63.1%)	12,386 (64.2%)	5,472 (68.3%)	3,362 (49.2%)
Non-inflamed Tiles	14,250 (34.8%)	6,768 (35.1%)	2,015 (25.1%)	3,266 (47.8%)
Tissue Classification				
Corpus Tiles	20,903 (51.0%)	9,929 (51.5%)	4,685 (58.5%)	3,561 (52.1%)
Antrum Tiles	16,787 (40.9%)	7,659 (39.7%)	2,896 (36.1%)	3,017 (44.2%)

Table 31: Seed: 44 Dataset Summary Statistics (Dataset$^{S12-HE}_{IT,v}$)

Characteristic	Total	Train	Val	Test
Total Tiles	40,997	20,542	7,420	6,179
Inflammation Classification				
Inflamed Tiles	25,875(63.1%)	12,923 (60.9%)	4,197 (56.56%)	4,100 (66.4%)
Non-inflamed Tiles	14,250 (34.8%)	7,199 (35.0%)	2,771 (37.3%)	2,079 (33.65%)
Tissue Classification				
Corpus Tiles	20,903 (51.0%)	11,475 (55.9%)	3,917 (52.79%)	2,783 (45.0%)
Antrum Tiles	16,787 (40.9%)	7,995 (38.9%)	2,644 (35.6%)	2,933 (47.5%)

Bibliography

Famke Aeffner, Mark D Zarella, Nathan Buchbinder, Marilyn M Bui, Matthew R Goodman, Douglas J Hartman, Giovanni M Lujan, Mariam A Molani, Anil V Parwani, Kate Lillard, Oliver C Turner, Venkata N P Vemuri, Ana G Yuil-Valdes, and Douglas Bowman. Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association. *J Pathol Inform*, 10: 9, 2019.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Trans Med Imaging*, 38(2): 550–560, Feb 2019.

Matthew Banks, David Graham, Marnix Jansen, Takuji Gotoda, Sergio Coda, Massimiliano di Pietro, Noriya Uedo, Pradeep Bhandari, D Mark Pritchard, Ernst J Kuipers, et al. British society of gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut*, 68(9):1545–1575, 2019.

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. *Nat Rev Clin Oncol*, 16(11):703–715, Nov 2019.

S Chen, P Ding, H Guo, L Meng, Q Zhao, and C Li. Applications of artificial intelligence in digital pathology for gastric cancer. *Frontiers in Oncology*, 14: 1437252, 2024. doi: 10.3389/fonc.2024.1437252.

Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Lab Invest*, 101(4):412–422, Apr 2021.

M F Dixon, R M Genta, J H Yardley, and P Correa. Classification and grading of gastritis. the updated sydney system. international workshop on the histopathology of gastritis, houston 1994. *Am J Surg Pathol*, 20(10):1161–1181, Oct 1996.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318 (22):2199–2210, 12 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL `https://doi.org/10.1001/jama.2017.14585`.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL `http://jmlr.org/papers/v17/15-239.html`.

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, Jan 2017. ISSN 1365-2559 (Electronic); 0309-0167 (Linking). doi: 10.1111/his.12993.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Tom Hempel. Development of a dataset and AI-based proof-of-concept algorithm for the classification of digitized whole slide images of gastric tissue. Bachelor's thesis, Otto-Friedrich-University Bamberg, Chair of Explainable Machine Learning, 2023.

F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. R. Heij, ..., and A. T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12:4423, 2021. doi: 10.1038/s41467-021-24698-1.

Weiming Hu, Chen Li, Xiaoyan Li, Md Mamunur Rahaman, Jiquan Ma, Yong Zhang, Haoyuan Chen, Wanli Liu, Changhao Sun, Yudong Yao, Hongzan Sun, and Marcin Grzegorzek. Gashissdb: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer, 2021. URL `https://arxiv.org/abs/2106.02473`.

Weiming Hu, Haoyuan Chen, Wanli Liu, Xiaoyan Li, Hongzan Sun, Xinyu Huang, Marcin Grzegorzek, and Chen Li. A comparative study of gastric histopathology sub-size image classification: from linear regression to visual transformer. 2022. URL `https://arxiv.org/abs/2205.12843`.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL `https://arxiv.org/abs/1608.06993`.

Gazi Husain, Jonathan Mayer, Molly Bekbolatova, Prince Vathappallil, Mihir Matalia, and Milan Toma. Machine learning for medical image classification. *Academia Medicine*, 1(4), 2024. doi: 10.20935/AcadMed7444. URL `https://doi.org/10.20935/AcadMed7444`.

Philipp Andreas Höfling. Development of an AI-based algorithm for the classification of gastric tissue in computational pathology. Master's thesis, Otto-Friedrich-University Bamberg, Chair of Explainable Machine Learning, 2023.

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 40(10):2845–2856, 2021. doi: 10.1109/TMI.2021.3056023.

Geert Litjens, Péter Bándi, Babak Ehteshami Bejnordi, N Burlutskiy, M Ghafoorian, Jeroen van der Laak, B Ginneken, Jeroen van der Laak, B Ginneken, C N Vasconcelos, et al. Camelyon17: Grand challenge on cancer metastasis detection in lymph nodes. *IEEE Transactions on Medical Imaging*, 42(9):2565–2583, 2018.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022a. URL `https://arxiv.org/abs/2111.09883`.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022b. URL `https://arxiv.org/abs/2201.03545`.

Shenghan Lou, Jianxin Ji, Huiying Li, Xuan Zhang, Yang Jiang, Menglei Hua, Kexin Chen, Kaiyuan Ge, Qi Zhang, Liuying Wang, Peng Han, and Lei Cao. A large histological images dataset of gastric cancer with tumour microenvironment annotation for ai. *Scientific Data*, 12(1):138, 2025.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL `https://arxiv.org/abs/1705.07874`.

Amirreza Mahbod, Christine Polak, Katharina Feldmann, Rumsha Khan, Katharina Gelles, Georg Dorffner, Ramona Woitek, Sepideh Hatamikia, and Isabella Ellinger. Nuinsseg: A fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. *Scientific Data*, 11(1):295, 2024. doi: 10.1038/s41597-024-03117-2. URL `https://doi.org/10.1038/s41597-024-03117-2`.

Omar S. M. El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J. Wagner, Katherine J. Hewitt, Firas Khader, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. From whole-slide image to biomarker prediction: A protocol for end-to-end deep learning in computational pathology. 2023. URL `https://arxiv.org/abs/2312.10944`.

Camilla Nero, Luca Boldrini, Jacopo Lenkowicz, Maria Teresa Giudice, Alessia Piermattei, Frediano Inzani, Tina Pasciuto, Angelo Minucci, Anna Fagotti, Gianfranco Zannoni, Vincenzo Valentini, and Giovanni Scambia. Deep-learning to predict brca mutation and survival from digital he slides of epithelial ovarian cancer. *International Journal of Molecular Sciences*, 23(19), 2022. ISSN 1422-0067.

doi: 10.3390/ijms231911326. URL `https://www.mdpi.com/1422-0067/23/19/11326`.

Rafał Obuchowicz, Michał Strzelecki, and Adam Piórkowski. Clinical applications of artificial intelligence in medical imaging and image processing-a review. *Cancers (Basel)*, 16(10), May 2024.

Gianmaria Pennelli, Federica Grillo, Francesca Galuppini, Giuseppe Ingravallo, Emanuela Pilozzi, Massimo Rugge, Roberto Fiocca, Matteo Fassan, and Luca Mastracci. Gastritis: update on etiological features and histological practical approach. *Pathologica*, 112(3):153–165, Sep 2020.

Massimo Rugge, Gianmaria Pennelli, Emanuela Pilozzi, Matteo Fassan, Giuseppe Ingravallo, Valentina M Russo, Francesco Di Mario, A Covacci, DY Graham, M Iascone, et al. Gastritis: the histology report. *Digestive and Liver Disease*, 43: S373–S384, 2011.

Taro Sakamoto, Tomoi Furukawa, Kris Lami, Hoa Hoang Ngoc Pham, Wataru Uegami, Kishio Kuroda, Masataka Kawai, Hidenori Sakanashi, Lee Alex Donald Cooper, Andrey Bychkov, and Junya Fukuoka. A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. *Transl Lung Cancer Res*, 9(5):2255–2276, Oct 2020.

Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig. Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions, 2024. URL `https://arxiv.org/abs/2406.17536`.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Antonia R Sepulveda and Madhavi Patil. Practical approach to the pathologic diagnosis of gastritis. *Arch Pathol Lab Med*, 132(10):1586–1593, Oct 2008a.

Antonia R Sepulveda and Madhavi Patil. Practical approach to the pathologic diagnosis of gastritis. *Arch Pathol Lab Med*, 132(10):1586–1593, Oct 2008b.

Z Shi, C Zhu, Y Zhang, Y Wang, W Hou, X Li, J Lu, X Guo, F Xu, X Jiang, Y Wang, J Liu, and M Jin. Deep learning for automatic diagnosis of gastric dysplasia using whole-slide histopathology images in endoscopic specimens. *Gastric Cancer*, 25 (4):751–760, 2022. doi: 10.1007/s10120-022-01294-w.

Pentti Sipponen and Heidi-Ingrid Maaroos. Chronic gastritis. *Scand J Gastroenterol*, 50(6):657–667, Jun 2015.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A closer look at domain shift for deep learning in histopathology, 2019. URL `https://arxiv.org/abs/1909.11575`.

David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136, 2018. doi: 10.1109/TMI.2018.2820199.

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019. doi: 10.1016/j.media.2019.101544.

David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):567–578, 2021. doi: 10.1109/TPAMI.2019.2936841.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL `https://arxiv.org/abs/1706.03762`.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. URL `https://arxiv.org/abs/1703.10593`.

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.

____
Place, Date

____
Signature