Automated classification of celiac disease in histopathological images: a multi-scale approach

Conference Paper · April 2024

DOI: 10.1117/12.3006669

12 authors, including:

Lluis Borras Ferris
HES-SO Valais-Wallis
5 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Niccolò Marini
University of Applied Sciences and Arts Western Switzerland
34 PUBLICATIONS **295** CITATIONS
SEE PROFILE

Iris Nagtegaal
Radboud University
679 PUBLICATIONS **44,153** CITATIONS
SEE PROFILE

Francesco Ciompi
Radboud University Medical Centre (Radboudumc)
110 PUBLICATIONS **17,537** CITATIONS
SEE PROFILE

Automated classification of celiac disease in histopathological images: a multi-scale approach

Simon Püttmann[a,b], Lluis Borras Ferris[a], Niccoló Marini[a], Witali Aswolinsky[d], Simona Vatrano[e], Filippo Fragetta[e], Iris Nagtegaal[d], Chella van der Post[d], Francesco Ciompi[d], Manfredo Atzori[a, c], Christoph Friedrich[b, g], and Henning Müller[a, f]

[a]University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
[b]University of Applied Sciences and Arts Dortmund, Dortmund, Germany
[c]Department of Neuroscience, University of Padua, Padua, Italy
[d]Radboud University Medical Center, Nijmegen, The Netherlands
[e]Pathology Unit, Gravina Hospital Caltagirone ASP, Catania, Italy
[f]The Sense Innovation and Research Center, Sion and Lausanne, Switzerland
[g]Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Essen, Germany

ABSTRACT

With a prevalence of 1-2% Celiac Disease (CD) is one of the most commonly known genetic and autoimmune diseases, which is induced by the intake of gluten in genetically predisposed persons. Diagnosing CD involves the analysis of duodenum biopsies to determine the small intestine condition. In this study, we propose a single-scale pipeline and the combination of two single-scale pipelines, forming a multi-scale approach, to accurately classify CD signs in histopathology whole slide images with automatically generated labels. The automatic classification of CD signs in histopathological images of these biopsies has not been extensively studied, resulting in the absence of a standardized guidelines or best-practices for this purpose. To fill this gap, we evaluated different magnifications and architectures, including a pre-trained MoCov2 model, for both single- and multi-scale approaches. Furthermore, for the multi-scale approach, methods for aggregating feature vectors from several magnifications are explored.

For the single-scale pipeline we achieved an AUC of 0.9975 and a weighted F1-score of 0.9680, while for the multi-scale Pipeline an AUC of 0.9966 and a weighted F1-score of 0.9250 was achieved. On large datasets, no significant differences were observed; however, with only 10% of the dataset, the multi-scale framework outperforms the single-scale framework significantly. Moreover, the multi-scale approach requires only half of the dataset and half of the time compared to the best single-scale result to identify the optimal model. In conclusion, the multi-scale framework emerges as an exceptionally efficient solution, capable of delivering superior results with minimal data and resource demands.

Keywords: Celiac Disease, Self-/Weakly-Supervised, Multi-Scale, Multiple-Instance-Learning, Histopathology

Further author information: (Send correspondence to S.P.)
S.P.: E-mail: simon.puettmann001@stud.fh-dortmund.de, Telephone: +49 1578 7365315

1. INTRODUCTION

Histopathology plays a crucial role in a wide range of diseases including almost all types of cancer.[1] This diagnostic approach involves the meticulous examination of tissue sections to identify microscopic disease manifestations. Tissue samples obtained by biopsies or surgical resections are carefully prepared for pathologists for microscopic examination. However, this manual analysis is a time-consuming task and the process might be affected by challenges such as different tissue morphologies, arbitrary selection of regions for detailed analysis, and subjective interpretation of findings.[2] As a result, inter-pathologist agreement on diagnoses can be low.[3–5] Moreover, technological advancements and enhancements in healthcare systems have led to a steady rise in the global collection of biopsies over the years. Unfortunately, the growth in the number of pathologists has not kept pace, resulting in an imbalanced workload within the sector.[6]

Although digital pathology is becoming increasingly important, clinical practice still relies heavily on traditional methods.[7] Digital pathology involves the acquisition and management of digitized tissue samples, known as whole slide images (WSIs). WSIs are digital scans of entire tissue sections on glass slides, offering comprehensive analysis, remote viewing, and collaborative possibilities among specialists. The transformation of physical tissue samples into digital images involves a complex process of dehydration, fixation, staining and subsequent scanning. Typically, whole slide scanners capture images with high optical magnification (20–40x), yielding spatial high-resolution images (0.25–0.5 μm per pixel). WSIs are usually stored in a versatile multi-scale format, allowing pathologists the flexibility to inspect different levels of detail, ranging from the lowest to the highest magnification. These digital images can be used by experts to identify specific symptoms in tissue sections, which are then documented in pathology reports. While structured synoptic reports are becoming more common,[8] semi-structured free-text reports remain the standard in the clinical setting.[9] These reports contain various fields, such as the type of tissue sample, identified findings, provisional diagnoses, and patient history. The increasing adoption of WSIs in hospitals[10,11] has resulted in the accumulation of thousands of digital images and diagnoses.

The integration of digital pathology into clinical workflows, coupled with advancements in deep learning, is paving the way for computer-assisted diagnostic (CAD) tools that can learn from clinical data without human intervention.[12,13] Computational pathology, an emerging field, is dedicated to the development of CAD tools designed for the automated analysis of digital pathology images. Among the various techniques applied in this field, convolutional neural networks (CNNs) have taken a central role as the state-of-the-art approach, consistently delivering exceptional performance across a range of computational pathology tasks. Nevertheless, the full potential of digital clinical pathology data remains untapped, and numerous challenges are still unresolved.

First and foremost it is crucial to acknowledge that CNNs demand extensive datasets during training to effectively accommodate the substantial variability encountered in clinical practice.[14] In parallel, fully supervised methods, while delivering superior performance, necessitate the meticulous task of annotating individual pixels,[14] a resource-intensive and time-consuming endeavor in the medical domain.[15] Moreover, the handling of WSIs poses a distinct set of challenges due to their substantial dimensions, often exceeding available memory capacity. Consequently, WSIs are frequently divided into smaller patches to facilitate analysis, albeit with the potential drawback of introducing spatial relationship distortions during subsequent processing. Additionally, WSIs frequently manifest significant disparities in staining patterns, a direct result of the absence of standardized protocols for tissue preparation and image acquisition across different medical facilities.[16] This intrinsic heterogeneity can impact the ability of models trained on such data to generalize effectively.

In recent years, weakly supervised learning approaches have emerged as a potential solution to some of these challenges.[15,17,18] These approaches use global (weak or image-level) annotations instead of local (pixel-wise) annotations, often derived from specific sub-regions within images. While this approach has the advantage of using the reports provided with the WSIs for labeling, it has required medical experts to extract these weak labels from the reports. To address this issue, Marini et al.[2] devised a framework that aims to overcome the limitations that have hindered the full exploitation of digital clinical pathology for training CAD tools. The proposed approach incorporates a natural language processing (NLP) pipeline designed to automatically analyze free-text reports and a computer vision algorithm trained with weak annotations for image classification. The NLP pipeline autonomously extracts semantically meaningful concepts from free-text diagnostic reports, which are then employed as weak labels to train an image classifier.

1.1 Celiac Disease

Celiac disease (CD) is an autoimmune disorder, induced by the intake of gluten present in wheat, barley and rye in genetically predisposed persons.[19,20] With a prevalence of 1-2%[21,22] CD is one of the most commonly known genetic and autoimmune diseases.[19,20]

CD is characterized by gastrointestinal and/or non-gastrointestinal symptoms. Diarrhea, weight loss, flatulence, abdominal pain, deficiency symptoms and fatigue are the most common intestinal symptoms.[19,22,23] Villous atrophy, crypt hyperplasia, strongly positive tissue transglutaminase-2 (tTG) auto-antibodies, intraepithelial T-cell (lymphoma) and in rare cases small intestinal adenocarcinoma, describe the pathological changes in the small intestine.[19,23] In addition CD can lead to iron deficiency anemia, osteoporosis, and various types of cancer and is associated with dermatitis herpetiformis, several neurologic and endocrine diseases, and other autoimmune disorders.[19,24]

According to the current state of research, about 90-95% of celiac diseases can be explained.[19] In these cases, special genetic predispositions are present.
The genes for the human leukocyte antigen (HLA) classes I and II are located in the major histocompatibility complex (MHC) on chromosome 6. Patients with this disease have specific human leukocyte antigens (HLA-DQ2 or HLA-DQ8). These genes encode for glycoproteins that bind peptides, and this HLA-peptide complex is recognised by certain T-cell receptors in the intestinal mucosa.[19–25] In addition, the function of tissue transglutaminase (tTG) is disturbed. Normally, tTG enzymes are responsible for various processes in the body, including the cross-linking of proteins for wound healing and the formation of cell envelopes.[19,23] However, in people with celiac disease, these enzymes can deaminate glutamine residues.[19,20,25] Before the proteins can be degraded, the tTG enzymes react with the gluten-containing peptides due to their high affinity for glutamine and cleave off an amino group, resulting in negatively charged gliadin peptides.[19]

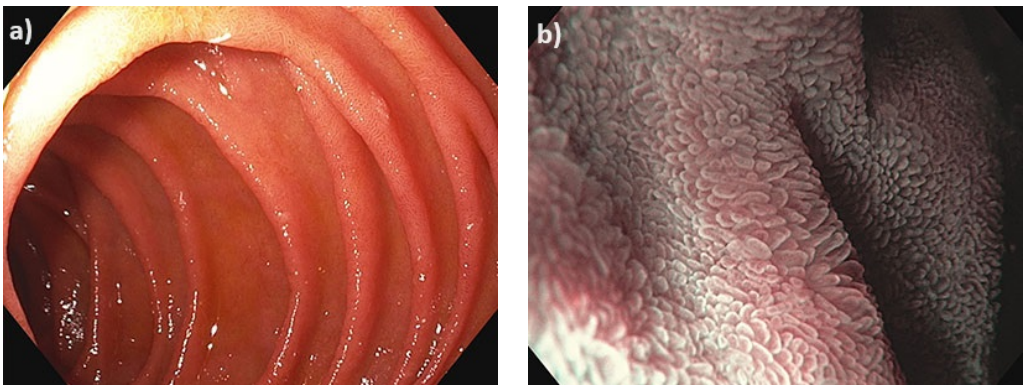

Figure 1. a) Endoscopy of regular duodenal folds b) Endoscopy of regular and erect villi (images taken from Hann et al.[26])

Studies have shown that increased gliadin levels lead to increased release of zonulin, a protein that regulates the permeability of the tight junctions in the intestinal wall.[27,28] Increased zonulin production dissembles the tight junctions and allows larger molecules to enter the intestinal mucosa.[20] This is thought to explain the increased susceptibility of celiac patients to autoimmune diseases.[25,29] Due to the altered intestinal permeability, the gliadin molecules enter the intestinal mucosa.

Given the negative charge of the gliadin peptides, there is a high affinity for the HLA-DQ2/8 molecules.[30] Since the HLA-DQ2/8 molecules are located on the surfaces of antigen-presenting cells of the small intestine such as macrophages, dendritic cells and B cells, they uptake and present the gliadin to the body's T cells.[25] T-cells recognize specific peptides in conjunction with the HLA molecules on the antigen-presenting cells.[24] Within individuals affected by celiac disease, certain T cells, especially CD4(+) T lymphocytes and CD8(+) T lymphocytes, are sensitive to the 33-amino acid sequence of the gluten-containing peptides.[31] This leads to their activation and clonal expansion of B cells that produce antibodies.[19,20] Cytokines released by activated CD4 T

cells interfere with the adaptive immune response, promote various inflammatory mechanisms and cause damage in the gut.[19,20,25]

Less is known about the activation and mechanism of action of intraepithelial T cells mediated by the innate immune system. However, the expression of the cytokine interleukin-15 appears to play a central role in the regulation of various processes leading to increased numbers of intraepithelial lymphocytes (IELs) as well as the destruction of epithelial cells and damage to the intestinal mucosa.[20,32]

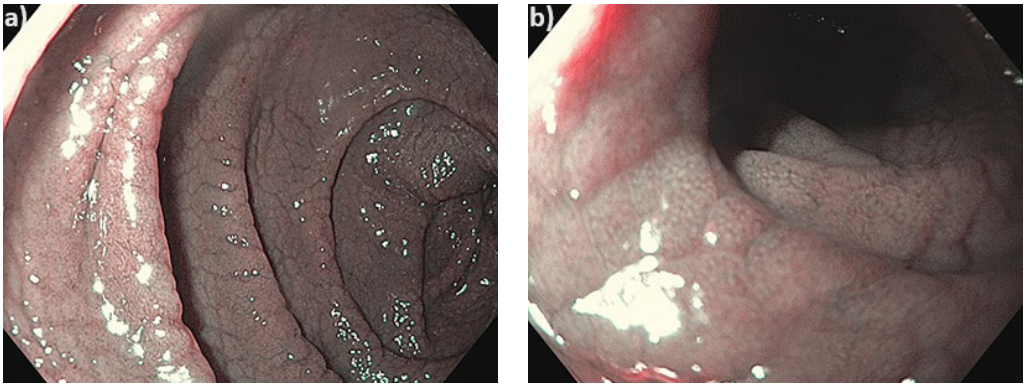

Figure 2. a) Duodenal folds with fissures on the ridges and a nodular mucosa b) On closer inspection complete loss of villi visible (images taken from Hann et al.[26])

For the remaining 5-10%, the origin of the disease is still unknown. It is assumed to be a combination of different gene variants and other environmental influences.

A combination of different methods is always necessary for diagnosis. In suspected cases, serological tests are carried out first. The gold standard here is the tTG-IGA antibody test, which has the highest sensitivity (97%).[20] For adults in cases of strong suspicion despite negative tests and for positive test patients, a biopsy must be performed.[19] At least four biopsies should be taken from the descending duodenal pars,[33] and in doubtful cases (latent or potential celiac disease) also two biopsies from the duodenal bulb.[23,33] In children, the diagnosis can be made without a biopsy if strict criteria are met.[33]
With today's technology, endoscopes equipped with narrow band imaging and near-field focus can be used to visualize endoscopic features of celiac disease in advance, such as loss of folds, fissure formation and nodular appearance of the mucosa (see Fig. 1, 2), which enables more precise tissue sampling.

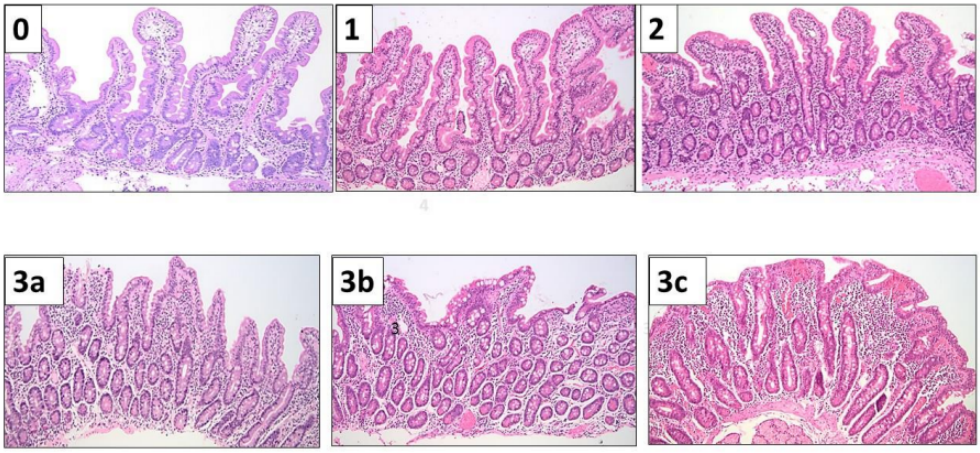

Figure 3. Overview of the Marsh stages for celiac disease classification (image taken from Brown[34])

Pathologists use histopathologic images of these biopsies to determine the stage of symptoms (Marsh stages, see figure 3), which range from 0 (normal) to 3c (complete villous atrophy). Three primary histopathologic changes are seen throughout the course of the disease. Stage 1 shows mild intraepithelial lymphocytosis (IEL). In combination with the presence of crypt hyperplasia, this corresponds to stage 2. Stage 3 is further subdivided into 3a to 3c, the latter depending on the degree of villous atrophy. In 3a, only a slight reduction in villous structure is observed, whereas in 3c, villi are no longer recognizable.

The only known treatment method today is the complete avoidance of foods containing gluten.[19–25] Treatment that reduces the zonulin production is a potential therapy that is part of current research.[20]
With complete abstinence, symptoms such as diarrhea and abdominal pain can be cured within a few weeks.[23] The complete regeneration of the intestine, on the other hand, takes up to two years.[35] Even after complete healing, gluten-containing products must continue to be avoided for life. Any increase in gluten can cause the same symptoms to arise again.

1.2 Related Work

Celiac disease classification has been the focus of various research efforts, although until recently, the number of publications and research conducted in this field has been limited. Notable among the early approaches were those of Sali et al.,[36] who achieved promising results using deep residual networks on a dataset from a Virginia hospital (ACC 0.85-0.91 and AUC > 0.96 on patch level and perfect results on slide level). However, this dataset lacked "normal" and Marsh class 2 images and was relatively small, comprising only 160 WSI samples. Sharma et al.[37] also explored CD classification using multiple instance learning, achieving an ACC of 0.862. From 2021 onwards, considerable advancements have been achieved, and the literature now showcases more results. Denholm et al.[38] employed Multiple Instance Learning in 2022 and achieved a remarkable ACC of 0.965 and an AUC of 0.996 on three combined datasets containing in total over 2000 WSIs. Additionally, Faust et al.[39] utilized support vector machines in 2023 to target Marsh 3a or higher classifications, achieving an ACC of 0.9855 by focusing solely on lamina propria patches while disregarding other regions. However, only one multi-scale approach has been documented for CD classification. Shrivastava et al.[40] extracted 1000x1000 and 2000x2000 patches, concatenated their feature vectors, and achieved a F1-score of 0.92 for CD classification based on this composite vector.

Multi-scale approaches have shown promising results in various experiments on histopathology image classification. Schmitz et al.[41] improved a ResNet18-based U-net[42] for three different histopathology image datasets (PAIP 2019,[43] BACH 2018,[44] CAMELYON 2016/MM subset[45]) by concatenating feature vectors from up to three different scales. Similarly, Alqahtani et al.[46] investigated three fusion methods for different magnifications. In the first experiment, they added the features of up to three different magnification feature vectors. In another experiment, the maximum value was selected from different feature vectors, and the channel weighting was determined based on the maximum weighting across all channels. Finally, two different methods of splice fusion were investigated. In cases with two scale features, the splicing fusion method first splits the channel weights in each scale along a particular coordinate axis and then maps the result to the final channel weight through a subsequent convolutional layer. While all methods improved the results compared to the single magnification approach, the summation and splicing fusion achieved the best results.

One major challenge in this field is the scarcity of publicly accessible datasets, with most studies relying on private datasets. Furthermore, the lack of published frameworks and variations in evaluation metrics hinder direct comparisons between projects, making it difficult to determine whether the achieved results are generalizable or specific to the utilized dataset.

The variation in magnification levels of the utilized images across projects is another significant discrepancy, with some projects failing to specify the magnification at all. For instance, Denholm et al.[38] worked with 10x magnification, Faust et al.[39] with 40x, and prior publications, like Wei et al.,[47] utilized 20x magnification.

1.3 Research Objectives

The primary objective of this research was to develop an automated pipeline capable of accurately classifying CD-related signs present in histopathology WSIs with automatically generated labels, to reduce the workload of medical experts in the preparation of the reports and the analysis of WSIs.

To achieve this goal and to address the existing research gap, our study conducted a comprehensive investigation to 1) determine the optimal magnification level for CD classification on WSIs and 2) evaluate the adoption of self-supervised algorithms to pre-train the backbones.

Moreover, we explored the potential benefits of incorporating a multi-scale approach within the CD classification pipeline. By integrating multiple magnification levels, we seeked to assess whether the fusion of features from several scales can improve the overall classification performance compared to relying only on a single magnification level. To achieve this, various fusion techniques such as concatenation, addition and multiplication of features, and attention-based aggregation were explored.

2. MATERIAL AND METHODS

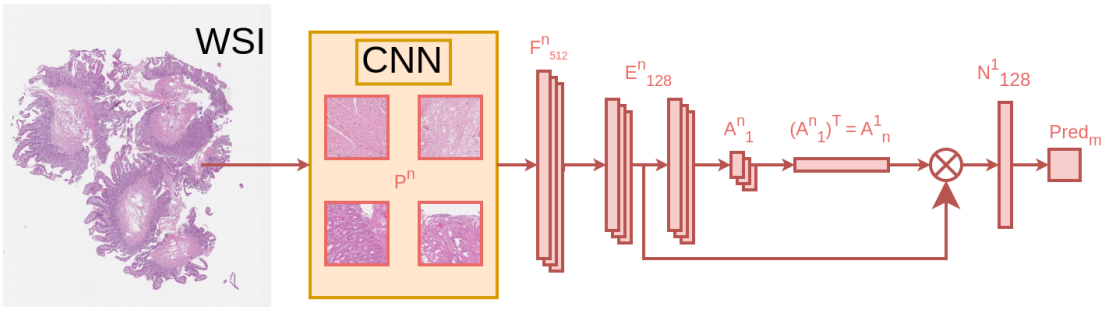

Figure 4. Single-scale pipeline: WSIs undergo patch extraction (P) followed by a pre-trained (ImageNet/MoCo) CNN feature extraction (F). The features are compressed into a flexible embedding vector (E) with lengths of 64, 128, or 256; here, the vector length is 128. This vector is further reduced to the attention layer (A). The transformation of A and multiplication with E results in a normalized vector N, which is employed for the prediction.

2.1 Dataset

The study utilized WSIs from the Radboud Medical University Center (Radboudumc, Nijmegen, The Netherlands) for training, validation, and testing purposes. The private dataset consists of approximately 1000 duodenum biopsies, accompanied by doctors' reports. To automate the labeling process, Marini et al.[2] developed a framework that generates labels (normal, unspecific duodenitis, and celiac disease) from the reports. This work focused on a binary classification task, specifically determining the presence or absence of symptoms related to celiac disease. For the test set, 50 images were randomly selected. To ensure the independence of data, precautions were taken to avoid including biopsies from the same patient in multiple training, validation, or test sets. After conducting a thorough analysis of the initial results, systematic errors in predicting the test set were identified. Certain images consistently yielded incorrect predictions, raising suspicions about the accuracy of the labels used. Further investigation of the clinical reports associated with these specific images revealed that the model for automatically generating labels had inaccurately labeled several images. In some reports, the British spelling for celiac disease (coeliac disease) was used, and all of these cases were labeled as "celiac positive", while the report said "no signs for coeliac disease". Consequently, the decision was made to let a medical expert manually label the test set to ensure the availability of reliable ground truth. Unfortunately, four images had to be excluded from the test set due to the unavailability of their corresponding reports.

Table 1. Distribution of labels in the Radboud dataset and the corresponding extracted test sets

	Radboudumc	automatic testset	manual testset
Celiac	325	13	12
Normal	592	37	34
Total	917	50	46

2.2 Preprocessing

Dividing the images into smaller tiles is a common approach to overcome the challenge of handling WSIs that have dimensions as large as 200,000 x 100,000 pixels. This division allows for simpler processing because directly processing the entire WSI using a GPU is usually not feasible due to its size.

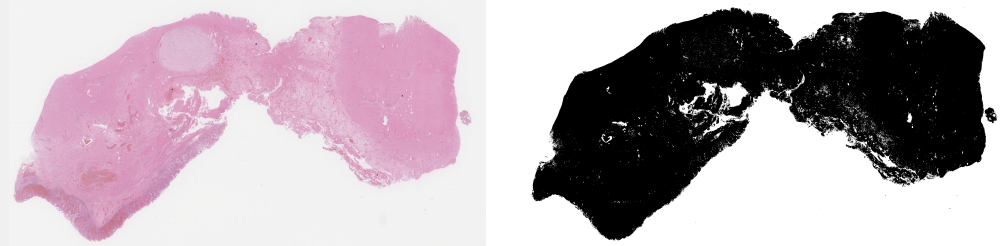

Figure 5. a) Thumbnail (2,414x1,221 pixel) of a WSI (39,660x13,285) b) Mask of the WSI generated with HistoQC

In this work HistoQC[48] was employed to generate black-and-white JPEG masks of the tissue, indicating whether a pixel corresponds to tissue (black) or background (white) (see Fig. 6). The ResNet34/50 architectures, which are pre-trained deep learning models specifically designed for 224x224 pixel images and known for their outstanding performance in image classification tasks, were chosen as the backbones for our system. Therefore every WSI is split into 224x224 pixel tiles. Each tile was then analyzed to determine the proportion of black and white pixels, indicating whether the tile contained sufficient information or consisted mostly of background or fat. Tiles with less than 50% black pixels were discarded and not utilized for training.

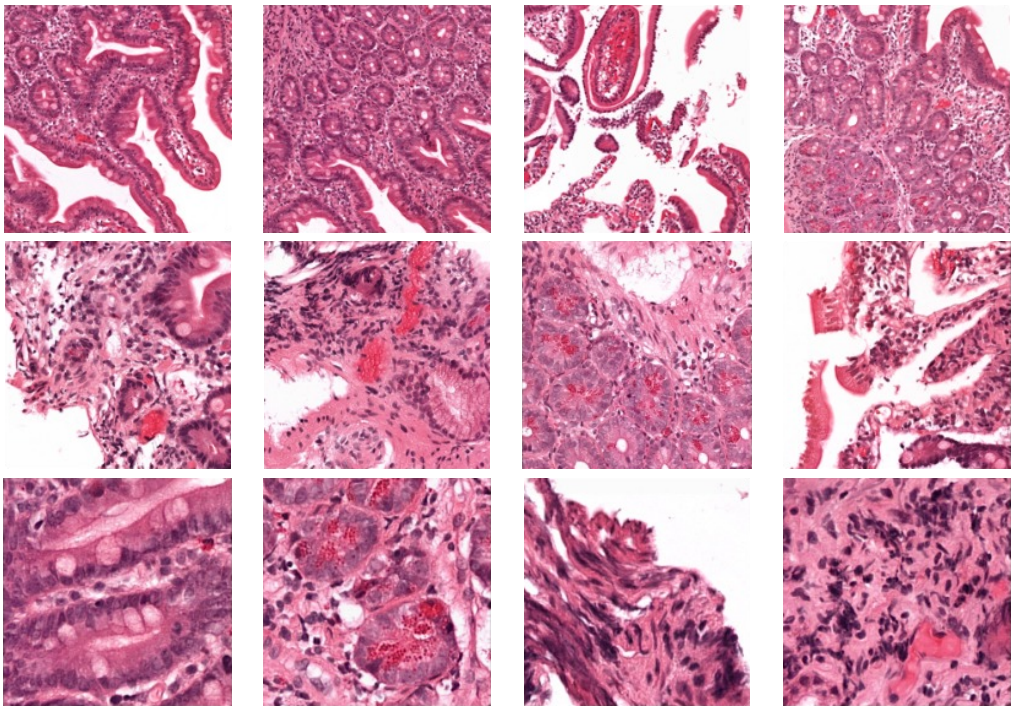

Figure 6. Random selection of tiles from a WSI of the Radboud dataset (1. row Magnification 5x, 2. row 10x, 3. row 20x)

2.3 Multiple Instance Learning

By employing a tiling approach, the data organization shifts from a single, large WSI to a collection of smaller patches, forming a bag of instances. However, since the labels are assigned at the WSI level rather than the

instance level, a single label is associated with the entire bag of instances. To tackle this challenge, the multiple instance learning (MIL) framework was utilized.

MIL is a weakly supervised framework specifically designed for scenarios where data is represented as a bag of instances without instance-level labels.[15,17] The fundamental assumption of MIL is that a bag is considered positive if at least one instance within it is positive, and it is considered negative if none of the instances within are positive. In this work, the MIL CNN proposed by Marini et al.[2] was adopted. This framework generates predictions for each individual instance, enabling the determination of whether the bag should be classified as positive or negative.

2.3.1 Transfer Learning

Transfer learning is a powerful approach that capitalizes on pre-trained models, initially trained on extensive datasets such as ImageNet[49] or IG-1B-Targeted.[50] The central objective is to leverage these models, which have already acquired feature representations from vast image datasets. This obviates the need to build models from scratch. Typically, the classifier layers of the pre-trained network can be fine-tuned, or specific layers can be unfrozen, allowing the model to learn pertinent features from a new dataset. This process involves retraining the model using the target dataset, enabling it to perform a novel classification task effectively. Consequently, transfer learning significantly expedites and enhances the process of adapting models to new tasks with reduced data and computational requirements.[51]

Given its consistent achievements in diverse image classification challenges,[52–54] the resNet34/50[42] CNNs were adopted as the foundational architectures for this framework. To leverage the benefits of transfer learning, the framework incorporated the pre-trained weights obtained from pre-training on the ImageNet dataset.

2.4 Self-Supervision

A self-supervised feature extractor was employed to address the limitation of using a pre-trained model on ImageNet, aiming to achieve better results by learning more specific and meaningful representations from histopathology duodenum images. The proposed self-supervised feature extractor was based on the MoCov2 framework. It uses contrastive learning with InfoNCE loss to train an encoder. By contrasting positive pairs (similar samples) against negative pairs (dissimilar samples) in a dictionary, the model learns meaningful representations without requiring annotations. The encoder is updated using momentum, and the queue-based dictionary allows efficient updates. This approach was demonstrated to surpass the performance of commonly used ImageNet-based pre-trained models across a diverse range of tasks.

2.5 Multi-Scale

Besides employing conventional magnifications, this research explored the utilization of multiple magnification approaches (see Fig. 7) due to the multifaceted manifestation of Celiac Disease symptoms at both cellular and tissue structure levels. For multi-scale approaches usually both feature vectors that are generated with several magnifications of its patches, are concatenated and used to predict the labels. In addition of concatenating feature vectors, methodologies involving addition, multiplication and aggregation through an additional attention layer were investigated.

2.6 Hyperparameters and Experiments

We conducted the initial experiment using the entire dataset, employing a grid search across magnifications (5x, 10, 20x), backbones (resNet34/50), embedding vector sizes (64, 128, 256), pre-training (MoCov2/ImageNet), and for the multi-scale approach different aggregations were tested (concatenation, summation, multiplication). Subsequently, we repeated the grid search on 10%, 20%, 40%, 60%, and 80% of the dataset to assess the efficiency on smaller datasets, address data scarcity challenges, and identify the most resource-efficient approach.
Further parameters utilized in this study were obtained based on the results of the grid search conducted by Marini et al.[2] Their investigation on cancer classification using histopathology images suggests that the Adam optimizer with a learning rate and weight decay set to 10^{-3} exhibited the best performance.

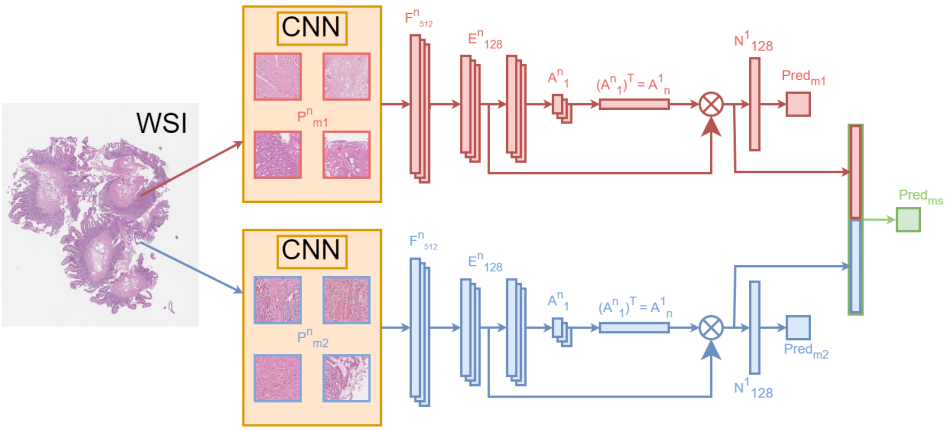

Figure 7. Multi-scale pipeline: Magnification m1 (red) and m2 (blue) undergo patch extraction (P) followed by a pre-trained (ImageNet/MoCo) CNN feature extraction (F). Similar to the single scale pipeline the features are compressed into a flexible embedding vector (E) and a transformation and multiplication normalizes the vectors, which are used for Predictions (Pred) on both scales. An aggregation (concat, sum ..) of the normalized vectors generates the multi-scale branch (green).

2.7 Evaluation

In the model evaluation, three distinct evaluation metrics were applied at the WSI level. During the training phase, after each epoch, the loss of the validation set was utilized to assess whether the current model outperforms the best model obtained from previous epochs. This approach ensures the models' continual improvement throughout the training process.

For the testing phase, considering the binary nature of the classification task and the significant class imbalance in the dataset, two specific performance metrics were employed. These metrics include the weighted F1-score and the area under the receiver operating characteristic curve (AUC-ROC).

The weighted F1-score is used to strike a balance between precision and recall, which is particularly useful in situations with a class imbalance. It provides a comprehensive evaluation of the models' ability to classify both classes effectively while considering their respective frequencies.

On the other hand, the AUC-ROC metric measures the models discrimination ability by plotting the true positive rate against the false positive rate at various classification thresholds. This metric is advantageous for unbalanced datasets since it is less affected by the prevalence of either class.

During the comparison of various models within each grid search, the primary emphasis was placed on evaluating the validation loss. In instances where models exhibited similar or nearly identical losses, additional metrics such as weighted F1-score and AUC-ROC were considered. This approach facilitated the identification of the best multi-scale and single-scale models for each grid-search.

To further assess the performance of these selected models, an independent test set was employed. The metrics used for comparison in this context were AUC-ROC and weighted F1. By employing this methodology, a comprehensive evaluation of the models was achieved, allowing for a robust comparison and selection of the most effective models across the experiments. To ascertain the significance of differences observed in AUC-ROC and weighted F1-score comparisons, the Wilcoxon test, a non-parametric statistical test, was employed. The Wilcoxon test is particularly suited for comparing paired samples, providing a robust analysis that does not rely on the assumption of normal distribution. This method aids in determining whether the disparities observed in model performance metrics are statistically significant, contributing to a more thorough and reliable assessment of the experimental results.

2.8 Software & Hardware

The entire pipeline was implemented using several Python libraries.
PyTorch 2.0.0 is employed for modeling, training, and testing the CNNs. OpenSlide 1.1.2 was utilized to access the WSIs. For evaluating the performance metrics of the models, Scikit-learn 0.24.1 was used.
All experiments were executed on a Tesla V100 GPU.

3. RESULTS

Table 2 presents the best results for both single-scale (SS) and multi-scale (MS) approaches, employing varying subsets of the dataset (10%, 20%, 40%, 60%, 80%, and 100%). Evaluation involved an extensive grid search across different magnifications (5x, 10x, 20x), embedding vector sizes (64, 128, 256), backbone architectures (resnet34, resnet50), and two different pre-trainings of the backbone (ImageNet and MoCo).

In the case of the single-scale approach using the entire dataset, the grid search identified resNet34 with 10x magnification, an embedding vector size of 128, and the use of the MoCo pre-trained network as the optimal parameters. This configuration resulted in an 0.9806 ± 0.017 AUC and 0.9680 ± 0.010 weighted F1-score. For the multi-scale approach, the best results were obtained with resNet34, employing magnifications of 5x and 10x, an embedding vector size of 256, and the summation of normalized vectors, yielding an AUC of 0.9995 ± 0.001 and a weighted F1-score of 0.9191 ± 0.047.

When the training and validation data were reduced to only 10% of the dataset, the optimal single-scale model featured resNet50 with 5x magnification, an embedding vector size of 128, and sole pre-training using ImageNet. This configuration achieved an AUC of 0.8329 ± 0.048 and a weighted F1-score of 0.9154 ± 0.034 as result. Notably, the multi-scale model outperformed significantly, demonstrating improved AUC (0.9811 ± 0.011) by incorporating 5x and 10x magnifications, resNet50 as the backbone, an embedding vector size of 64, and exclusive ImageNet pre-training. The weighted F1-score also improved, although not significantly (0.8715 ± 0.069).

The most effective single-scale model was trained with 80% of the dataset, achieving a 0.9975 ± 0.002 AUC and a weighted F1-score of 0.9680 ± 0.010. On the other hand, the top-performing multi-scale model was trained with 40% of the dataset, attaining 0.9966 ± 0.004 AUC and a weighted F1-score of 0.9250 ± 0.034.

Table 2. Subset analysis: Best results for each single- (SS) and multi-scale (MS) approach using dataset subsets. Model selection is based on average validation loss, comparison via AUC and weighted F1-score on an independent test set.

Subset	SS AUC	SS F1	MS AUC	MS F1
100%	0.9806 ± 0.017	0.8978 ± 0.045	**0.9995 ± 0.001**	**0.9191 ± 0.047**
80%	**0.9975 ± 0.002**	**0.9680 ± 0.010**	0.9838 ± 0.001	0.9462 ± 0.021
60%	**0.9944 ± 0.005**	0.9147 ± 0.034	0.9887 ± 0.008	**0.9283 ± 0.027**
40%	0.9782 ± 0.018	0.8997 ± 0.05	**0.9966 ± 0.004**	**0.9250 ± 0.034**
20%	0.9946 ± 0.007	0.9192 ± 0.064	**0.9961 ± 0.003**	**0.9210 ± 0.026**
10%	0.9154 ± 0.034	0.8329 ± 0.048	**0.9811 ± 0.011**	**0.8715 ± 0.069**

4. DISCUSSION

The initial results on the complete dataset showcase strong model performance for both single- and multi-scale approaches, indicating their effectiveness in handling the entire dataset. However, to address concerns of overfitting and optimize resource utilization, a systematic exploration of dataset subsets was conducted. As the dataset size decreases, the performance of both pipelines remains robust, with AUCs consistently above 0.97 and F1-scores above 0.89 from 20% of the dataset onward.

Interestingly, the optimal dataset sizes differ for the single- and multi-scale pipelines, with the single-scale pipeline achieving its peak at 80% of the dataset and the multi-scale pipeline at 40%. Despite the relatively large step sizes in dataset reduction, these results indicate potential variations in the models' sensitivity to data size, emphasizing the need for careful pipeline selection based on resource constraints.

Notably, at 10% of the dataset, the multi-scale pipeline exhibits a significant (assessed with Wilcoxon test) improvement in AUC compared to the single-scale pipeline. While the difference in F1-scores is not statistically significant, this finding suggests that the multi-scale approach might be more robust in scenarios with extremely limited data availability.

Table 3. Overview of optimal model parameters for various dataset subsets. Highlighted are the best single- and multi-scale model, as well as the multi-scale model with significantly better results than the corresponding single-scale model

Model	magnification(s)	backbone	E-vector	pre-training	fusion
100% SS	10x	resNet34	256	MoCo	/
100% MS	5 + 10x	resNet34	128	ImageNet	summation
80% SS	**5x**	**resNet50**	**128**	**MoCo**	/
80% MS	5 + 10x	resNet34	128	MoCo	concatenation
60% SS	10x	resNet34	256	ImageNet	/
60% MS	5 + 10x	resNet34	128	MoCo	concatenation
40% SS	5x	resNet34	128	MoCo	/
40% MS	**5 + 10x**	**resNet34**	**256**	**ImageNet**	**concatenation**
20% SS	5x	resNet50	64	MoCo	/
20% MS	5 + 10x	resNet34	128	ImageNet	summation
10% SS	5x	resNet50	128	ImageNet	/
10% MS	**5 + 10x**	**resNet50**	**64**	**ImageNet**	**summation**

The experimental findings reveal several key insights into the parameterization of models for subsets of a given dataset (see Tab. 3). Notably, the combination of 5x and 10x magnifications consistently emerges as the optimal choice for the multi-scale approach, showcasing a harmonious balance between capturing fine details and maintaining overall image coverage. Single-scale models exhibit favorable outcomes with either 5x or 10x magnification, with a slight preference for the former, especially in scenarios involving smaller datasets. However, 20x consistently performed worse. The choice of backbone architecture demonstrates nuanced dependencies on the dataset size; ResNet50 excels in smaller subsets, while ResNet34 gains prominence for subsets comprising 40% or more of the entire dataset, except for the distinctive success of ResNet50 in the 80% single-scale scenario. Embedding vector length exhibits a discernible impact on results, with a tendency towards longer vectors in larger datasets, indicating the need for increased parameterization to accommodate data complexity. Noteworthy is the positive influence of MoCo pre-training on single-scale models, contrasting with its less pronounced impact on multi-scale models. Further investigation into MoCo parameter tuning for multi-scale scenarios could enhance its efficacy. Regarding fusion strategies, concatenation and summation emerge as the optimal choices, with summation potentially serving as a robust alternative to mitigate overfitting or underfitting concerns in datasets of varying sizes.

Table 4. Overview of the time required to train the best single- and multi-scale models and the previous grid-search to determine the best models..

Subset	SS best model	MS best model	Grid-search
100%	28:12 min	35:30 min	31:22:16 h
80%	17:17 min	25:53 min	23:53:17 h
60%	16:40 min	19:11 min	18:44:44 h
40%	07:39 min	16:06 min	12:07:20 h
20%	05:27 min	06:39 min	06:01:26 h
10%	02:57 min	06:49 min	04:01:07 h

The in table 4 provided training times for both single- and multi-scale models, along with the grid search durations, offer valuable insights into the computational efficiency of the different experimental setups. It is essential to acknowledge potential variations in resource allocation, as CPU and GPU resources were shared,

introducing a degree of uncertainty into the exact conditions of each experiment. Nevertheless, these times provide a rough estimate for analysis purposes.

Notably, the grid search duration consistently decreases at a roughly constant rate of approximately 6 hours for every 20% reduction in the dataset size. In terms of training individual models, it is evident that multi-scale models consistently demand more training time. This increase may be attributed to the doubled feature set and the larger model size associated with the multi-scale approach. An intriguing observation arises when comparing the best multi-scale model, trained on 40% of the dataset, to the best single-scale model, trained with 80% of the dataset. Remarkably the training times for these two models are nearly identical. However, the comparison becomes even more noteworthy when considering the time required to find the best models through grid search. While it takes 24 hours for the grid search at 80% of the dataset, the grid search at 40% of the dataset is accomplished in only 12 hours.

In the current landscape, where resource optimization is crucial for minimizing carbon footprints and considering the shared nature of CPU and GPU hardware, the efficiency of model training becomes paramount. It is clear from the results that while multi-scale models may demand more training time, careful consideration of the dataset size can lead to resource-efficient solutions. The choice between single- and multi-scale models should factor in both performance and computational efficiency, especially in scenarios where shared hardware resources and environmental concerns play a significant role. This highlights the contemporary need for models that not only deliver robust performance but also contribute to a sustainable and efficient use of computational resources.

5. CONCLUSION

This research paper presented and compared two frameworks for celiac disease classification: a single-scale pipeline and the combination of two single-scale pipelines as multi-scale pipeline. The primary objective of accurately identifying celiac disease symptoms was achieved by both approaches. While both framworks yielded similar results overall, the multi-scale pipeline demonstrated significantly better outcomes with very small datasets. Moreover, the multi-scale approach requires half of the dataset and half of the time compared to the best single-scale result to identify the optimal model.

In essence, the multi-scale framework proves to be extremely data and resource-efficient, achieving significantly better results with less data and requiring fewer resources for optimal performance.

Furthermore, the potential for enhanced results and reduced resource demands through additional research, especially in the context of MoCov2, suggests a pathway for further refinement and optimization. Additional research aimed at exploring the integration of more than two magnifications holds the potential to enhance results even further.

In conclusion, the multi-scale framework emerges as an exceptionally efficient solution, capable of delivering superior results with minimal data and resource demands. This efficiency holds significant promise for alleviating the workload of medical experts in both report preparation and the analysis of WSIs.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825292 (ExaMode, http://www.examode.eu/).

REFERENCES

[1] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., and Yener, B., "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering* **2**, 147–171 (2009).

[2] Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., Van Rijthoven, M., Aswolinskiy, W., et al., "Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations," *NPJ digital medicine* **5**(1), 102 (2022).

[3] Costantini, M., Sciallero, S., Giannini, A., Gatteschi, B., Rinaldi, P., Lanzanova, G., Bonelli, L., Casetti, T., Bertinelli, E., Giuliani, O., et al., "Interobserver agreement in the histologic diagnosis of colorectal polyps: the experience of the multicenter adenoma colorectal study (smac)," *Journal of clinical epidemiology* **56**(3), 209–214 (2003).

[4] Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rueschoff, J. H., and Claassen, M., "Automated gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific reports* **8**(1), 12054 (2018).

[5] Vennalaganti, P., Kanakadandi, V., Goldblum, J. R., Mathur, S. C., Patil, D. T., Offerhaus, G. J., Meijer, S. L., Vieth, M., Odze, R. D., Shreyas, S., et al., "Discordance among pathologists in the united states and europe in diagnosis of low-grade dysplasia for patients with barrett's esophagus," *Gastroenterology* **152**(3), 564–570 (2017).

[6] Märkl, B., Füzesi, L., Huss, R., Bauer, S., and Schaller, T., "Number of pathologists in germany: comparison with european countries, usa, and canada," *Virchows Archiv* **478**, 335–341 (2021).

[7] Pallua, J., Brunner, A., Zelger, B., Schirmer, M., and Haybaeck, J., "The future of pathology is digital," *Pathology-Research and Practice* **216**(9), 153040 (2020).

[8] Hewer, E., "The oncologist's guide to synoptic reporting: a primer," *Oncology* **98**(6), 396–402 (2020).

[9] Hanna, M. G., Reuter, V. E., Ardon, O., Kim, D., Sirintrapun, S. J., Schüffler, P. J., Busam, K. J., Sauter, J. L., Brogi, E., Tan, L. K., et al., "Validation of a digital pathology system including remote review during the covid-19 pandemic," *Modern Pathology* **33**(11), 2115–2127 (2020).

[10] Fraggetta, F., Garozzo, S., Zannoni, G. F., Pantanowitz, L., and Rossi, E. D., "Routine digital pathology workflow: the catania experience," *Journal of pathology informatics* **8**(1), 51 (2017).

[11] Thorstenson, S., Molin, J., and Lundström, C., "Implementation of large-scale routine diagnostics using whole slide imaging in sweden: Digital pathology experiences 2006-2013," *Journal of pathology informatics* **5**(1), 14 (2014).

[12] Van der Laak, J., Litjens, G., and Ciompi, F., "Deep learning in histopathology: the path to the clinic," *Nature medicine* **27**(5), 775–784 (2021).

[13] Bozorgtabar, B., Mahapatra, D., Zlobec, I., Rau, T. T., and Thiran, J.-P., "Computational pathology," (2020).

[14] Madabhushi, A. and Lee, G., "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical image analysis* **33**, 170–175 (2016).

[15] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine* **25**(8), 1301–1309 (2019).

[16] Khalbuss, W. E., Pantanowitz, L., and Parwani, A. V., "Digital imaging in cytopathology," *Pathology research international* **2011** (2011).

[17] Ilse, M., Tomczak, J., and Welling, M., "Attention-based deep multiple instance learning," in [*International conference on machine learning*], 2127–2136, PMLR (2018).

[18] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering* **5**(6), 555–570 (2021).

[19] Rodrigo, L., "Celiac disease," *World journal of gastroenterology: WJG* **12**(41), 6577 (2006).

[20] Caio, G., Volta, U., Sapone, A., Leffler, D. A., De Giorgio, R., Catassi, C., and Fasano, A., "Celiac disease: a comprehensive current review," *BMC medicine* **17**, 1–20 (2019).

[21] Lohi, S., Mustalahti, K., Kaukinen, K., Laurila, K., Collin, P., Rissanen, H., Lohi, O., Bravi, E., Gasparin, M., Reunanen, A., et al., "Increasing prevalence of coeliac disease over time," *Alimentary pharmacology & therapeutics* **26**(9), 1217–1225 (2007).

[22] Tommasini, A., Not, T., Kiren, V., Baldas, V., Santon, D., Trevisiol, C., Berti, I., Neri, E., Gerarduzzi, T., Bruno, I., et al., "Mass screening for coeliac disease using antihuman transglutaminase antibody assay," *Archives of disease in childhood* **89**(6), 512–515 (2004).

[23] Vogelsang, H., "Zöliakie," *Journal für Gastroenterologische und Hepatologische Erkrankungen* **7**(1), 10–14 (2009).

[24] Fasano, A. and Catassi, C., "Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum," *Gastroenterology* **120**(3), 636–651 (2001).

[25] Green, P. H. and Cellier, C., "Celiac disease," *New england journal of medicine* **357**(17), 1731–1743 (2007).

[26] Hann, A., Walter, B. M., and Meining, A., "Endoskopische befunde bei zöliakie (einheimischer sprue)," (2017). https://www.endoscopy-campus.com/bildergalerie/endoskopische-befunde-bei-zoeliakie-einheimischer-sprue/ accessed: 04.01.2024.

[27] Clemente, M. G., De Virgiliis, S., Kang, J., Macatagney, R., Musu, M., Di Pierro, M., Drago, S., Congia, M., and Fasano, A., "Early effects of gliadin on enterocyte intracellular signalling involved in intestinal barrier function," *Gut* **52**(2), 218–223 (2003).

[28] Sander, G. R., Cummins, A. G., and Powell, B. C., "Rapid disruption of intestinal barrier function by gliadin involves altered expression of apical junctional proteins," *FEBS letters* **579**(21), 4851–4855 (2005).

[29] Ventura, A., Magazzù, G., Greco, L., et al., "Duration of exposure to gluten and risk for autoimmune disorders in patients with celiac disease," *Gastroenterology* **117**(2), 297–303 (1999).

[30] Molberg, Ø., McAdam, S. N., and Sollid, L. M., "Role of tissue transglutaminase in celiac disease," *Journal of pediatric gastroenterology and nutrition* **30**(3), 232–240 (2000).

[31] Shan, L., Molberg, Ø., Parrot, I., Hausch, F., Filiz, F., Gray, G. M., Sollid, L. M., and Khosla, C., "Structural basis for gluten intolerance in celiac sprue," *Science* **297**(5590), 2275–2279 (2002).

[32] Maiuri, L., Ciacci, C., Ricciardelli, I., Vacca, L., Raia, V., Auricchio, S., Picard, J., Osman, M., Quaratino, S., and Londei, M., "Association between innate response to gliadin and activation of pathogenic t cells in coeliac disease," *The Lancet* **362**(9377), 30–37 (2003).

[33] Al-Toma, A., Volta, U., Auricchio, R., Castillejo, G., Sanders, D. S., Cellier, C., Mulder, C. J., and Lundin, K. E., "European society for the study of coeliac disease (esscd) guideline for coeliac disease and other gluten-related disorders," *United European gastroenterology journal* **7**(5), 583–613 (2019).

[34] Brown, I., "Histopathology of coeliac disease," (2016). https://www.envoi.com.au/sites/default/files/publications/pathology_of_coeliac_disease.pdf, accessed: 12.10.2023.

[35] Wahab, P. J., Meijer, J. W., and Mulder, C. J., "Histologic follow-up of people with celiac disease on a gluten-free diet: slow and incomplete recovery," *American journal of clinical pathology* **118**(3), 459–463 (2002).

[36] Sali, R., Ehsan, L., Kowsari, K., Khan, M., Moskaluk, C. A., Syed, S., and Brown, D. E., "Celiacnet: Celiac disease severity diagnosis on duodenal histopathological images using deep residual networks," in [*2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*], 962–967, IEEE (2019).

[37] Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C. A., Syed, S., and Brown, D., "Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification," in [*Medical Imaging with Deep Learning*], 682–698, PMLR (2021).

[38] Denholm, J., Schreiber, B., Evans, S., Crook, O., Sharma, A., Watson, J., Bancroft, H., Langman, G., Gilbey, J., Schönlieb, C.-B., et al., "Multiple-instance-learning-based detection of coeliac disease in histological whole-slide images," *Journal of Pathology Informatics* **13**, 100151 (2022).

[39] Faust, O., De Michele, S., Koh, J. E., Jahmunah, V., Lih, O. S., Kamath, A. P., Barua, P. D., Ciaccio, E. J., Lewis, S. K., Green, P. H., et al., "Automated analysis of small intestinal lamina propria to distinguish normal, celiac disease, and non-celiac duodenitis biopsy images," *Computer Methods and Programs in Biomedicine* **230**, 107320 (2023).

[40] Shrivastava, A., Kant, K., Sengupta, S., Kang, S.-J., et al., "Deep learning for visual recognition of environmental enteropathy and celiac disease," in [*2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*], 1–4, IEEE (2019).

[41] Schmitz, R., Madesta, F., Nielsen, M., Krause, J., Steurer, S., Werner, R., and Rösch, T., "Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture," *Medical image analysis* **70**, 101996 (2021).

[42] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[43] Kim, Y. J., Jang, H., Lee, K., Park, S., Min, S.-G., Hong, C., Park, J. H., Lee, K., Kim, J., Hong, W., et al., "Paip 2019: Liver cancer segmentation challenge," *Medical image analysis* **67**, 101854 (2021).

[44] Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al., "Bach: Grand challenge on breast cancer histology images," *Medical image analysis* **56**, 122–139 (2019).

[45] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama* **318**(22), 2199–2210 (2017).

[46] Alqahtani, Y., Mandawkar, U., Sharma, A., Hasan, M. N. S., Kulkarni, M. H., and Sugumar, R., "Breast cancer pathological image classification based on the multiscale cnn squeeze model," *Computational Intelligence and Neuroscience* **2022** (2022).

[47] Wei, J. W., Wei, J. W., Jackson, C. R., Ren, B., Suriawinata, A. A., and Hassanpour, S., "Automated detection of celiac disease on duodenal biopsy slides: A deep learning approach," *Journal of pathology informatics* **10**(1), 7 (2019).

[48] Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., and Madabhushi, A., "Histoqc: an open-source quality control tool for digital pathology slides," *JCO clinical cancer informatics* **3**, 1–7 (2019).

[49] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE conference on computer vision and pattern recognition*], 248–255, Ieee (2009).

[50] Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D., "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546* (2019).

[51] Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., and Ganslandt, T., "Transfer learning for medical image classification: a literature review," *BMC medical imaging* **22**(1), 69 (2022).

[52] Brancati, N., De Pietro, G., Riccio, D., and Frucci, M., "Gigapixel histopathological image analysis using attention-based neural networks," *IEEE Access* **9**, 87552–87562 (2021).

[53] Zhang, L., Wu, Y., Zheng, B., Su, L., Chen, Y., Ma, S., Hu, Q., Zou, X., Yao, L., Yang, Y., et al., "Rapid histology of laryngeal squamous cell carcinoma with deep-learning based stimulated raman scattering microscopy," *Theranostics* **9**(9), 2541 (2019).

[54] Tajane, K., Sheth, S., Satale, R., Tumbare, T., and Panchal, O., "Breast cancer detection using machine learning algorithms," in [*Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, Volume 1*], 347–355, Springer (2022).