NARROW BAND IMAGING VERSUS WHITE-LIGHT: WHAT IS BEST FOR COMPUTER-ASSISTED DIAGNOSIS OF CELIAC DISEASE?

M. Gadermayr[1] *S. Hegenbart*[2] *R. Kwitt*[2] *A. Uhl*[2] *A. Vécsei*[3]

[1] Institute of Imaging and Computer Vision, RWTH Aachen, Germany
[2] Department of Computer Sciences, University of Salzburg, Austria
[3] St. Anna Children's Hospital, Department of Pediatrics, Medical University Vienna, Austria

ABSTRACT

Recent developments of specialized endoscopic hardware with enhanced visualization capabilities such as narrow band imaging have been shown to improve the diagnostic accuracy in clinical practice. The current state-of-the-art in computer-assisted diagnosis of celiac disease (CD) in flexible endoscopy uses data captured under the modified immersion technique with white-light illumination. In this work, the potential benefits of the modified immersion technique using narrow band imaging for automated diagnosis is studied. We provide convincing experimental evidence that the imaging modality has a significant impact on the underlying feature distribution of general purpose image representations. Consequently, the design of systems for automated diagnosis requires the consideration of several factors in this context. We present a large experimental setup studying the most relevant factors for automated diagnosis of CD.

***Index Terms*—** Celiac disease, computer-aided diagnosis, narrow band imaging, classification, endoscopy

1. INTRODUCTION

Celiac disease (CD) is a multisystemic immune-mediated disease, which is associated with considerable morbidity and mortality [1]. In untreated or inappropriately treated CD the inflammation caused by the dysregulated immune response can disrupt the intestinal mucosa thus leading to a total atrophy of the villi, which are finger-like projections of the mucosa. After embarking on a strict gluten-free diet (GFD), which is the CD treatment modality of first choice, the inflammation gradually subsides allowing for mucosal healing. To avoid the most severe complications of CD, an early diagnosis for commencing a strict GFD is of vital importance.

Computer-assisted systems for diagnosis of CD have potential to improve the whole diagnostic work-up, by saving costs, time and manpower and at the same time increase the safety of the procedure. A further motivation for such a system is given by generally high inter-observer variabilities [2, 3]. Biopsy of the small intestine performed during upper endoscopy remains the gold standard for confirmation of CD. Besides standard upper endoscopy, several new endoscopic approaches for diagnosing CD have been evaluated and found their way into clinical practice [4]. The most notable techniques include the modified immersion technique (MIT [5]) under traditional white-light illumination (denoted as WL_{MIT}), as well as MIT under narrow band imaging [6, 7] (denoted as NBI_{MIT}). These specialized endoscopic techniques were specifically designed for improving the visual confirmation of CD during endoscopy. The authors report improvements of the diagnostic accuracy using these techniques in clinical practice. It is generally unclear however, how the altered visualization of duodenal tissue affects systems for automated diagnosis. Hegenbart et al. [8] have presented strong empirical evidence, that data captured under WL_{MIT} is superior for computer-aided diagnosis as compared to conventional white-light endoscopy. Since then WL_{MIT} has been used as the modality of choice in automated CD diagnosis in flexible endoscopy [9] and has become a de-facto standard.

In this work, we evaluate the diagnostic yield of NBI_{MIT} for computer-assisted diagnosis of CD in comparison to WL_{MIT} in the clinically most relevant binary distinction between healthy tissue and celiac-tissue.

We present empirical evidence using a kernel two-sample test that the underlying distributions of general purpose image representations under either technique are significantly different. Consequently, we study relevant scenarios towards a clinical deployment of a fully automated system in the context of multiple imaging modalities such as using opposing and combined techniques for training and prediction. We finally evaluate the diagnostic benefits of each modality in relation to the amount of available training samples (likely the most relevant problem in clinical practice).

2. ENDOSCOPIC IMAGING

The MIT technique is based on rapid instillation of water into the duodenal lumen after evacuation of air by suction. Villi, if present, straighten up in water and appear as tiny finger-like structures. Experimental evidence was gained [5] that the vi-

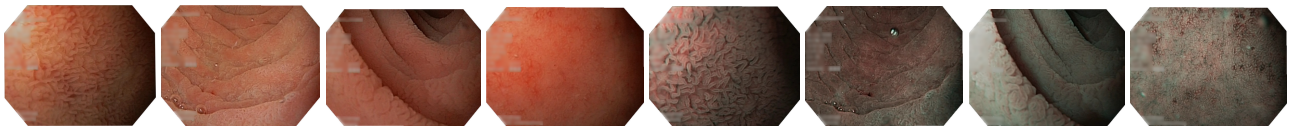

Fig. 1: Comparison of endoscopic markers visualized under WL$_{MIT}$ and NBI$_{MIT}$

sualization of villi using MIT has a higher positive predictive value. Additionally, the water provides several other benefits such as a more homogeneous illumination without specular reflections and bubbles. A major benefit is, that no specialized endoscopic hardware is required to capture data under WL$_{MIT}$.

Narrow band imaging [6] has been reported to improve the diagnostic accuracy in various fields of endoscopy [7]. This sophisticated endoscopic imaging technology utilizes specific blue (440 to 460 nm) and green (540 to 560 nm) wavelengths for illumination to enhance the contrast of vascular patterns on the mucosal surface. It is employed to specifically delineate the outline of the residual villous structures (if present) due to a better visualization of villous height and shape compared with traditional white-light endoscopy.

2.1. Staging of Celiac Disease and Endoscopic Markers

The mucosal alterations caused by CD are classified following Oberhuber et al. [10] in six classes. Vécsei et al. [11] pointed out problems with a lack of visual distinction in appearance between classes of type Marsh-3 and suggest to use a more simplified, visually focused systems for computer-assisted diagnosis. The relevant classes for automated diagnosis in the binary classification scenario are comprised of type Marsh-0 (healthy) and the Marsh-3 types (ranging from 3A to 3C).

The most prevalent endoscopic markers for CD include scalloped folds, mosaic patterns of the mucosa and a nodular mucosa [12]. All of these features become more pronounced and more easily detectable when using WL$_{MIT}$ as compared to conventional imaging. Narrow band imaging is employed (if available) to specifically delineate the outline of the residual villous structures (if present). Narrow band imaging allows a better assessment of the villous height and shape compared to the conventional white-light endoscopy. Figure 1 provides a comparison of the visualized appearance of the most prevalent markers under WL$_{MIT}$ and NBI$_{MIT}$.

2.2. Endoscopic Image Data

Experimentation is performed on images captured during standard upper endoscopy at the St. Anna Children's hospital in children with indications for CD. The images show specific markers for the manifestation of CD and were acquired by the endoscopist during the clinical treatment. Due to the high amount of endoscopic image degradations, sub-images with a dimension of 128×128 pixels were manually extracted by a domain expert to guarantee a controlled experimental environment. This configuration allows for a fair comparison between modalities, even though the reported accuracies might not be realistic for a clinically deployed system.

Table 1: Distribution of image data

	Images		**Patients**	
	Marsh-0	**Marsh-3**	**Marsh-0**	**Marsh-3**
NBI	415	306	88	42
WL	327	286	85	41

To particularly study the effects of each modality on the textural appearance of duodenal tissue, we use gray-scale image data throughout experimentation. Table 1 illustrates the distribution of images and patients used for experimentation. We assure a fair comparison by under-sampling the training data (when necessary) during cross-validation as explained in Section 3.

3. EXPERIMENTS

The classification pipeline used throughout experimentation utilizes four image representations. Two methods (based on Local Binary Patterns), which were previously reported as competitive in automated diagnosis of CD [11] were utilized in addition to state-of-the-art global mid-level image representations which are obtained by pooling local image descriptors. Local Binary Pattern (LBP) [13] represent textures as the joint distribution of underlying micro structures, modeled via intensity differences in a pixel neighborhood. We employ a standard eight-neighborhood with a radius of two pixels. We also use a variation of LBP, utilizing a ternary decision with thresholding instead of relying on a binary sign function known as Local Ternary Patterns (LTP [14]). Again the standard eight-neighborhood was used with a radius of two-pixels and threshold of five. Improved Fisher Vectors (IFV [15]) are based on estimated Gaussian mixtures of locally pooled SIFT (Scale-invariant Feature Transform [16]) descriptors. The improved version based on a non-linear Hellinger's kernel and l^2 normalization with 16 clusters is used. Finally the Vector of Locally Aggregated Descriptors (VLAD [17]) image rep-

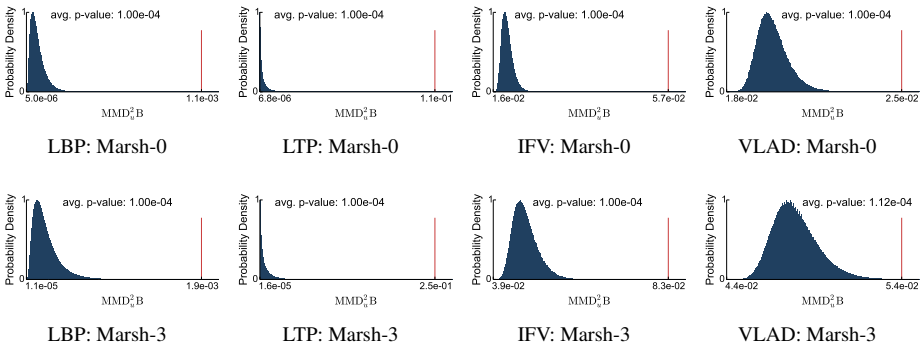

Fig. 2: Empirical distribution of the MMD test-statistic under the null hypothesis (NBI_{MIT} versus WL_{MIT})

resentation with 64 clusters is employed. We rely on in-house MATLAB implementations for LBP and LTP and use the implementations of IFV and VLAD as provided by *VLFeat*.

We follow a standard evaluation protocol based on iterated two-fold cross-validations based on random sampling. In each iteration 80 % of data is utilized for training and 20 % for evaluation. The applied sampling strategy guarantees a partition on a patient basis [18] as well as balanced class distributions (using under-sampling if necessary) to avoid biased evaluations. The results are reported as the mean of 50 iterations with the corresponding standard deviations plotted as error bars. A linear support vector machine (C-SVM, based on LIBLINEAR [19]) was utilized for classification. The cost factor C has been separately evaluated for each iteration in an inner cross validation ($\text{C} \in \{2^0, 2^1, \ldots, 2^{12}\}$).

The empirical error of the SVM is used as criterion for comparison. We consider methods with the highest classification accuracy as most suited for computer-assisted diagnosis. Due to the nature of the data (different underlying distributions), statements of statistical significance cannot be given.

3.1. Experimental Results

Underlying Feature Distributions: The results of the first experiment indicate that it is imperative to take the used imaging modality into consideration when designing a system for automated diagnosis. By means of a kernel (RBF) two-sample test (Maximum Mean Discrepancy, MMD [20]) we present strong evidence that the underlying feature distributions of all considered image representations are significantly affected by the imaging modality. The null hypothesis (equal underlying distributions under both modalities) could be rejected with very high confidence in all cases.

Figure 2 presents the empirical distribution of the test-statistic under the null hypothesis. To avoid a potential bias caused by using images from a single patient in both samples, we perform multiple MMD-tests based on a strategy guaranteeing that a patient is only drawn for either sample from the empirical feature distributions. The minimum test-statistic under the alternative hypothesis (different distributions) of 1,000 iterations (10,000 trials) is indicated by a red line in Fig. 2. The reported p-values are the mean over all iterations.

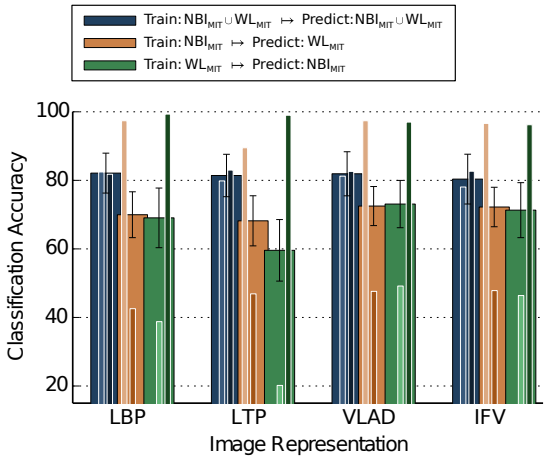

Fig. 3: Results of using opposing and combined modalities for training and prediction

Opposing and Combined Modalities: The development of highly specialized endoscopic imaging hardware promotes a potential diversity in equipment used in clinical practice. It is consequently of high interest for a wide clinical deployment of a computer-assisted system for diagnosis of CD to support both imaging modalities, either implicitly or explicitly, an issue we consider in this experiment.

The practical consequences of using opposing as well as combined endoscopic modalities for training and prediction in an automated system are studied in detail. Classification models are trained on i) data from both modalities combined and used to predict images under both modalities ii) either modality and used to predict images under the other (not trained) modality.

To guarantee a fair comparison we restrict the number of

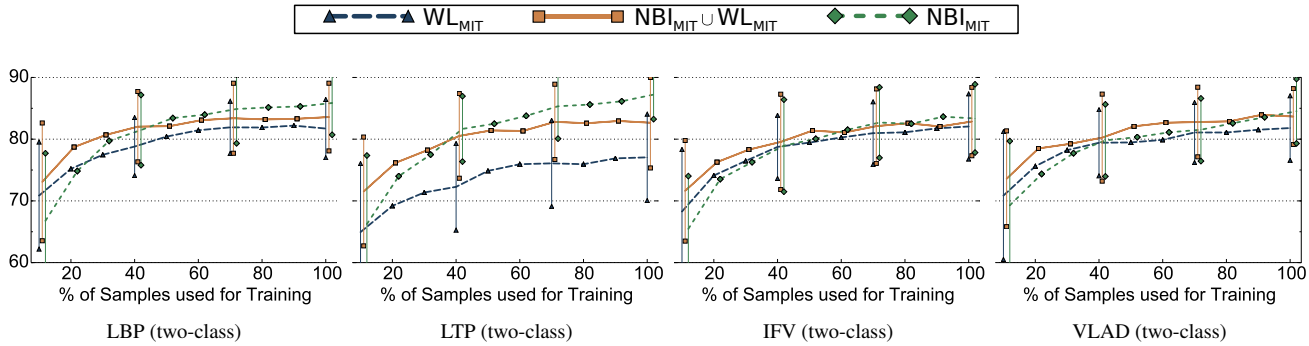

Fig. 4: Classification accuracy as function of the amount of training samples

training samples in the combined scenario to be equal to the number of samples used in either modality.

Figure 3 presents the results of this experiment. The narrow sub-bars indicate the sensitivity (left) and the specificity (right). The results of this experiment confirm the statistical findings of the previous experiment. Classification models trained on either modality are unable to predict images from the other modality reliably. It can be observed that a classifier trained on data from WL_{MIT} leads to a high number of false negatives if used to predict NBI_{MIT} data. Due to the enhanced contrast of vascular patterns under NBI_{MIT} as compared to WL_{MIT}, the visualized textures in mild cases of CD are misinterpreted as normal villi structures. In parallel a classifier trained on NBI_{MIT} data leads to a high number of false positives caused by a less pronounced visualization of endoscopic markers in WL_{MIT}. The first experiment (Figure 2) suggests that feature distributions of IFV and VLAD are less distinct between modalities. This is reflected by slightly better accuracies for training and prediction under opposing modalities as compared to LBP-based methods. A potential clinically relevant finding of this experiment is that a system trained on images drawn from both endoscopic modalities generalizes reasonably well without the need for domain adaptation techniques.

Best Suited Modality: We use the final experiment (Figure 4) to answer the question for the best suited imaging modality in automated-diagnosis of CD. A very common limitation in practice is an insufficient amount of data available for training. In a clinical scenario however, data from both modalities are potentially available (endoscopes supporting NBI_{MIT} generally support WL_{MIT} as well). We consequently also study the potential benefits of providing a larger training corpus, comprised of data from both modalities in automated systems. In each step, classification models are trained based on an increasing fraction of the available training samples of either modality (the amount of data used for evaluation is constant), as well as a combination of both modalities with twice the amount of training samples. We present the classification accuracy as function of the amount of used training samples (as a percentage of all available images for training. The experimental evaluation shows that the benefits of a larger training corpus holds until approximately 40 percent of the available training data is used (approximately 115 images overall). Beyond that point, data under NBI_{MIT} is more beneficial to the training of the classification models. Interestingly classifiers trained on WL_{MIT} data are consistently outperformed by systems based on the larger, combined training corpus. Data captured under NBI_{MIT} is generally better suited for automated-diagnosis as compared to WL_{MIT} data in all scenarios. This is especially pronounced in case of LBP-based methods. The results indicate that the image representations based on local pooling are generally less influenced by the chosen image modality as compared to LBP-based features.

4. CONCLUSION

We presented an experimental evaluation of potential diagnostic benefits using data captured under NBI_{MIT} as compared to WL_{MIT} in computer-assisted diagnosis of CD. We gained convincing empirical evidence that the used imaging modality has a significant impact on the underlying feature distribution of general purpose image representations. These findings were confirmed in our experiments based on data under opposing modalities used for training and prediction. A relevant finding of our experiments is that systems trained on images drawn from both modalities generalize well without requiring additional domain adaptation techniques. We specifically assessed the best suited modality (WL_{MIT} versus NBI_{MIT}) in relation to the amount of available samples for training. We showed that in a potential scenario with insufficient amounts of data for training, combining both modalities to form a larger training corpus improves the accuracy of automated diagnosis. Empirical evidence suggest that systems for automated diagnosis benefit most from data captured under NBI_{MIT}.

5. REFERENCES

[1] F. Biagi, G.R. Corazza, F. Biagi, and G.R. Corazza, "Mortality in celiac disease," *Nat Rev Gastroenterol Hepatol.*, vol. 7, no. 3, pp. 158 – 162, 2010.

[2] Federico Biagi, Emanuele Rondonotti, Jonia Campanella, Federica Villa, Paola Ilaria Bianchi, Catherine Klersy, Roberto De Franchis, and Gino Roberto Corazza, "Video capsule endoscopy and histology for small-bowel mucosa evaluation: A comparison performed by blinded observers," *Clinical Gastroenterology and Hepatology*, vol. 4, no. 8, pp. 998–1003, 2006.

[3] Sonia Niveloni, Alcira Fiorini, Ruben Dezi, Silvia Pedreira, Edgardo Smecuol, Horacio Vazquez, Ana Cabanne, Luis A. Boerr, Jorge Valero, Zulema Kogan, Eduardo Maurino, and Julio C. Bai, "Usefulness of videoduodenoscopy and vital dye staining as indicators of mucosal atrophy of celiac disease: assessment of interobserver agreement," *Gastrointestinal Endoscopy*, vol. 47, no. 3, pp. 223–229, 1998.

[4] N. Chand and A. A. Mihas, "Celiac disease: Current concepts in diagnosis and treatment," *J Clin Gastroenterol*, vol. 40, no. 1, pp. 3 – 14, 2006.

[5] A. Gasbarrini, V. Ojetti, L. Cuoco, G. Cammarota, A. Migneco, A. Armuzzi, P. Pola, and G. Gasbarrini, "Lack of endoscopic visualization of intestinal villi with the immersion technique in overt atrophic celiac disease," *Gastrointest Endosc*, vol. 57, pp. 348–351, 2003.

[6] F. Emura, Y. Saito, and Ikematsu H., "Narrow-band imaging optical chromocolonoscopy: advantages and limitations," *World J Gastroenterol*, vol. 14, no. 31, pp. 4867–4872, 2008.

[7] F. Valitutti, S. Oliva, D. Iorfida, M. Aloi, S. Gatti, C. M. Trovato, M. Montuori, A. Tiberti, S. Cucchiara, and G. Di Nardo, "Narrow band imaging combined with water immersion technique in the diagnosis of celiac disease," *Dig and Liver Dis*, vol. 46, no. 12, pp. 1099–1102, 2014.

[8] S. Hegenbart, R. Kwitt, M. Liedlgruber, A. Uhl, and A. Vécsei, "Impact of duodenal image capturing techniques and duodenal regions on the performance of automated diagnosis of celiac disease," in *Proceedings of ISPA'09*, 2009, pp. 718 – 723.

[9] Sebastian Hegenbart, Andreas Uhl, and Andreas Vécsei, "Survey on computer alded decision support for diagnosis of celiac disease," *Computers in Biology and Medicine*, 2015.

[10] G. Oberhuber, G. Granditsch, and H. Vogelsang, "The histopathology of coeliac disease: time for a standardized report scheme for pathologists," *Eur J Gastroenterol Hepatol*, vol. 11, no. 10, pp. 1185 – 1194, 1999.

[11] A Vécsei, G. Amann, S. Hegenbart, M. Liedlgruber, and A. Uhl, "Automated marsh-like classification of celiac disease in children using local texture operators," *Comput Biol Med*, vol. 41, no. 6, pp. 313 – 325, 2011.

[12] William Dickey and Dermot Hughes, "Prevalence of celiac disease and its endoscopic markers among patients having routine upper gastrointestinal endoscopy," *The American Journal of Gastroenterology*, vol. 94, no. 8, pp. 2182 – 2186, 1999.

[13] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recogn*, vol. 29, no. 1, pp. 51–59, 1996.

[14] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Proceedings of AMFG'07*, 2007, pp. 168–182.

[15] F. Perronnin, Yan Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Proceedings of CVPR'10*, 2010, pp. 3384–3391.

[16] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of CVPR'99*, 1999, vol. 2, pp. 1150–1157.

[17] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proceedings of CVPR'10*, 2010, pp. 3304–3311.

[18] S. Hegenbart, A. Uhl, and A. Vécsei, "Systematic assessment of performance prediction techniques in medical image classification. a case study on celiac disease," in *Information Processing in Medical Imaging*, 2011, vol. 7512 of *LNCS*, pp. 498 – 509.

[19] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[20] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, Mar. 2012.