Automated Classification of Celiac Disease During Upper Endoscopy: Status Quo and Quo Vadis

M. Gadermayr\textsuperscript{a,*}, G. Wimmer\textsuperscript{b,*}, H. Kogler\textsuperscript{c}, A. Vécsei\textsuperscript{c}, D. Merhof\textsuperscript{a}, A. Uhl\textsuperscript{b}

\textsuperscript{a}Institute of Imaging & Computer Vision, RWTH Aachen University, 52074 Aachen, Germany
\textsuperscript{b}Department of Computer Sciences, University of Salzburg, 5020 Salzburg, Austria
\textsuperscript{c}St. Anna Children’s Hospital, Vienna, Austria

Abstract

A large amount of digital image material is routinely captured during esophagogastroduodenoscopies but, for the most part, is not used for confirming the diagnosis process of celiac disease which is primarily based on histological examination of biopsies. Recently, considerable effort has been undertaken to make use of image material by developing semi- or fully-automated systems to improve the diagnostic workup. Recently, focus was especially laid on developing state-of-the-art deep learning architectures, exploiting the endoscopist’s expert knowledge and on making systems fully automated and thereby completely observer independent. In this work, we summarize recent trends in the field of computer-aided celiac disease diagnosis based on upper endoscopy and discuss about recent progress, remaining challenges, limitations currently prohibiting a deployment in clinical practice and future efforts to tackle them.

Keywords: Celiac disease, decision support, computer-aided diagnosis, deep learning, observer independent, classification

1. Introduction

Celiac disease (CD), or coeliac disease, is a complex autoimmune disorder that affects the small intestine in genetically predisposed individuals of all age groups after the introduction of food containing gluten. Characteristic for the disease is an inflammatory reaction in the mucosa of the small bowel. During the course of CD the mucosa loses its absorptive villi and hyperplasia of the enteric crypts occurs, leading to a diminished ability to absorb nutrients. Recent studies showed that the prevalence of CD in Northern America and Europe ranges between 1:80 and 1:300 [1, 2, 3, 4].

\textsuperscript{*}Shared first authorship, corresponding author

\textit{Email addresses:} michael.gadermayr@lfb.rwth-aachen.de (M. Gadermayr), gwimmer@cosy.sbg.ac.at (G. Wimmer), uhl@cosy.sbg.ac.at (A. Uhl)
The current gold standard for diagnosis of CD is endoscopy biopsies. Histopathological analysis is performed according to the modified Marsh classification [5] which distinguishes between the classes Marsh-0 to Marsh-3, with subclasses Marsh-3a, Marsh-3b, and Marsh-3c, resulting in a total number of six classes. According to this classification scheme, Marsh-0 denotes healthy mucosa (without visible changes of the villous structure) and Marsh-3c designates a complete absence of villi (villous atrophy).

The whole diagnostic work-up of CD, including endoscopy with biopsies, is time-consuming, cost-intensive, and rather invasive. Consequently, to save costs, time and manpower and to increase the safety of the procedure, a less invasive approach avoiding or reducing biopsies would be highly desirable. Studies by Cammarota et al. [6, 7] showed that upper endoscopy using an immersion technique can reliably detect abnormalities of duodenal villi. Automated diagnosis systems could help the physician during endoscopy by applying a continuous automated mucosa analysis that raises alarm in the case areas affected by CD are identified. Alternatively, the endoscopist could make use of the option to manually trigger the classification process for a specific mucosal region of interest.

Due to the possibly patchy distribution of CD [8, 9], areas affected by CD can be in the midst of normal mucosa. So, given the case that the biopsies are taken from areas of healthy mucosa, CD will be missed due to a sampling error. Automated systems could be used to indicate areas that are affected by CD and thus improve the targeting accuracy of biopsy taking.

Previous research also showed that the histological staging of biopsies is subject to a significant degree of intra- and inter-observer variability [10, 11, 12]. It is foreseeable that this variability might worsen especially in pediatric CD. In 2012, the European Society of Pediatric Gastroenterology, Hepatology and Nutrition issued new diagnostic criteria that for the first time allowed a non-invasive diagnosis under certain circumstances: in children and youths presenting with a history compatible with celiac disease and having tissue transglutaminase-IgA-autoantibody levels more than ten times the upper limit of normal and positive endomysium antibodies endoscopy with biopsies is not needed for the confirmation of the diagnosis [13]. A multicenter prospective European study conducted after 2012 provided sound evidence for the validity of this biopsy-skipping approach and showed that more than half the children with celiac disease do not need endoscopy with biopsies any longer for an accurate diagnosis of celiac disease [14]. However, the remaining children with celiac disease, who do not fulfil these criteria, still require endoscopy with biopsies, the analysis of which will be even more difficult because this group of children predictably will comprise more equivocal cases with more subtle alterations of the intestinal mucosa further aggravating the intra- and interobserver variability. Automated diagnosis systems are observer independent and could provide a “second opinion” either for a specific region in the bowel or for the overall diagnosis on a patient and thus improve the accuracy of the histopathological classification. For this purpose, the digital image material, which is routinely captured during esophagogastro-duodenoscopies, can be exploited.
Research on automated diagnosis systems for the classification of CD was performed for two highly different endoscopic modalities:

- "Manual" esophagogastroduodenoscopies (EGDs): During this manual endoscopy, the physician has control of the camera and focuses on informative regions of the small intestine where images as well as biopsies are taken. Research in this field only made use of image patches that were manually extracted from actively captured image data [15, 16, 17, 18, 19, 20, 21, 22, 23].

- Wireless capsule endoscopy: Although the image data shows the same content, the arising challenges are different, as there is no way to control the camera in case of capsule endoscopy. Previous works using wireless capsule endoscopy were performed by Ciaccio et al. [24, 25, 26, 27, 28, 29, 30]. For a survey on quantitative analysis of capsule endoscopy we refer to [31].

In this work we only consider manual EGD. Recent literature [32, 33, 34, 35, 36] in this field generally aims at a two-classes classification scheme. Specifically, mucosa showing no visible changes (Marsh-0) should be distinguished from mucosa showing visible markers of villous atrophy (Marsh-3A, Marsh-3B and Marsh-3C). For classification, mostly patches with a size of 128 $\times$ 128 pixels were manually extracted [32, 33, 34, 35, 36]. Pipelines generally consisted of a feature extraction and a separate classification stage. For image representation, either general purpose features such as local binary patterns [16, 37, 38, 39, 32] or Fisher vectors [40, 41] or domain specific approaches [21, 42] were employed. For classification, mainly support vector machines or k-nearest neighbor classifiers were applied. To measure classification performance, either cross-validation was applied or a fixed separation into training and test set was employed.

### 1.1. Contribution

In this work, we provide a review of recent trends on computer aided CD diagnosis based on endoscopic images obtained during EGDs. We focus on the last three years, i.e. we summarize published work since the survey of Hegenbart et al. [43], showing a significant change in methods as well as in the considered application scenarios. Hand-crafted features were completely replaced by state-of-the-art deep neural network architectures (Sect. 2). Apart from that, focus was laid on making the approaches completely automated by getting rid of the restriction that the patches need to be manually extracted (Sect. 3). Furthermore, diagnosis systems were not only applied to single patches (as done previously) but also to the combined data of patients enabling a more accurate decision on patient level. Finally, effort was undertaken to exploit the endoscopist’s expert knowledge to increase the system’s accuracy (Sect. 4). We put the recent work in a larger context and discuss about progress, remaining challenges and about limitations that are currently prohibiting a deployment in clinical practice and about future efforts to tackle them.
2. Deep Learning for the diagnosis of CD

Convolutional neural networks (CNN) are gaining more and more interest in computer vision. CNNs widely replaced handcrafted image representations exhibiting the previous state-of-the-art. Also in terms of the automated diagnosis of celiac disease, CNNs replaced and clearly outperformed handcrafted image representations [34, 35, 36].

Generally thousands or millions of images are used and required as training data corpus to achieve well generalizing CNNs. In endoscopic image classification, however, the available amount of data usable as training corpus is quite limited (the CD image data set employed in all three publications applying CNNs for the automated diagnosis of celiac disease [34, 35, 36] consist of 1661 image patches of size 128 × 128 pixels).

Consequently, all three publications about CNNs for the diagnosis of CD worked on solutions for the problem of limited training data. The solutions in those works were to make use of smaller networks with relatively few parameters [34], to use pretrained or partly pretrained networks [35] and to increase the amount of training data by applying data augmentation techniques [36].

In [34], we investigated if deep architectures are too complex for the classification of celiac disease and if simpler and more shallow architectures could be better suited in such a scenario due to the small amount of training data and the low number of categories. To evaluate the optimal filter sizes and assess how “deep” networks need to be, four different architectures ranging from very shallow (only one convolutional block) to deep (four convolutional blocks) and three different filter sizes (large, medium or small filters) per network architecture were employed for the classification of CD. The experiments showed that deep architectures utilizing large or medium sized filters are best suited for the classification of CD, despite of the higher number of parameters that need to be optimized.

In [35], we raised the question if networks trained on huge general purpose data sets can be employed for the classification of CD or if they need to be adapted for this task. For this, three “training” strategies were applied to train CNNs:

1. CNNs pretrained on the ImageNet data set were used as fixed feature extractors without any training on the CD data set.
2. Only the highly class specific fully connected layers of the pretrained networks were adapted by training the nets on the CD data set. The less class specific convolutional layers remained unchanged.
3. All layers (convolutional and full-connected layers) were adapted.

The best overall classification rates were achieved by adapting all layers of the networks, although the fully adapted networks were clearly overfitted to the training data. We expect an increased positive effect when more training data become available.
In [36], we analyzed which data augmentation techniques are most suited for the classification of CD. The employed data augmentation techniques aim to increase the number of training images and their variety while still modeling realistic viewing condition. Six different augmentation techniques were applied, ranging from simple and common augmentation techniques like flipping and cropping up to augmentation techniques using the full range of affine or projective transformations (see Fig. 1). It turned out that CNNs clearly benefit from data augmentation and that the best results are achieved either using the full spectrum of affine transformations or the full spectrum of projective transformations.

Overall, evidence was provided that CNNs constitute highly effective methods for the computer aided diagnosis of CD. Using pretrained CNNs as initialization for CNN training as well as data augmentation turned out to be useful techniques for the training of CNNs in the presence of limited training data. In comparisons with handcrafted feature extraction techniques, the CNNs clearly outperformed the previous state-of-the-art.

3. Towards Automated Patient-wise Classification

Recently, state-of-the-art deep neural networks definitely improved the classification accuracies on CD image patches. However, the employed CD image patches needed to be manually extracted from endoscopic images. That means,
a "reliable" patch with a size of 128 × 128 pixels was extracted prior to the automated classification process [15, 38, 18, 44, 45, 46, 36, 35]. Consequently, the decision support system was not completely automated and thereby observer independent as the manual patch extraction significantly influences the final output. Final accuracies are therefore expected to highly suffer in case that an unexperienced endoscopist extracts inappropriate patches showing some kind of degradations [38] or an inappropriate scale [18].

To bypass the issue that patch extraction is not automated, we developed different strategies to select reliable regions of interest based on a quality measure. Specifically, we proposed a measure consisting of the weighted sum of a set of five individual metrics measuring illumination, contrast, blur-level, noise-level and reflections [47]. Apart from the selection of the best patches of each patient based on image quality measurements, we also considered the fusion of each patient’s available data in order to obtain one final decision per patient.

The experimental results showed that all five individual quality measures are clearly outperformed by the weighted combination of all five quality metrics. The patch selection approach based on the combination of the five quality metrics achieved accuracies up to 85%, which is approximately 5% below the manual patch extraction (accuracies differ for the individual feature extraction techniques). Furthermore, we showed that the results can be improved by fusing the information of the automatically extracted patches from an image. A feature-level fusion outperformed straight-forward decision-level fusion finally exhibiting accuracies similar to a manual patch selection. It was also shown that weighting is extremely important (the weights were optimized based on a second distinct data set) as equally weighting shows weak performances. We also identified that training needs to be applied on manually extracted patches in order to obtain optimum accuracy.
Two years later, we proposed an approach making use of CNN features to obtain fully-automated classification on patient level [48]. For this purpose, we used one version of the VGG-f network that was trained on the ImageNet database and one version that was furthermore trained on a CD database using manually selected patches (Fig. 2). We made use of the VGG-net features and fitted a discriminator first to distinguish between manually extracted patches. Based on this model, we trained a further probabilistic model to obtain the probability of mis-classification. To achieve one final decision for an individual patient, the decisions and the respective confidences extracted from all available patches (automatically extracted in a regular grid) were aggregated in a histogram and finally evaluated based on a basic nearest neighbor classifier. Although in case of performing fusion on image-level (i.e. all patches per image are fused) the state-of-the-art [47] was not improved, in case of more than one available endoscopic image per patient, we obtained clear improvements. With the best performing approach, mean accuracies of 90% were achieved. It proved to be beneficial to combine features extracted from a CNN trained on the ImageNet data set with the features extracted from a CNN adapted to the CD data set.

Overall, we noticed that the classification accuracy can be increased by using multiple images for the decision process instead of single images. In the experiments we made use of the images actively taken by the endoscopist during the endoscopies. However, it can be expected that classification scores can be increased even further when using the video material harbouring a clearly larger amount of data, albeit of lower quality.

4. Incorporation of Expert Knowledge

It was already shown that human experts are highly accurate in judgment of CD [7, 6]. Besides completely automated and observer independent systems, another field of research is focusing on fusing capabilities of human experts with computer-based techniques. For this purpose, a rather basic yet effective approach was developed in order to merge all available knowledge (see Fig. 3). Specifically, we performed a feature-level fusion by basic concatenation of the image representation with the binary expert decision. We conducted a large experimental study with three medical experts and three different data sets [49] and showed that in the vast majority of cases, both, the computer and the expert were statistically significantly outperformed by the fusion based approach.

In a further study [41], we showed that improvements are not only obtained for classification on patch-level, but also when all available data of each individual patient is fused. We also noticed that trained models are quite robust even when changing the medical expert they were trained for. A "domain change" here leads to moderate decreases in accuracy only, at least when the novel rater is not clearly less accurate.

Making use of expert knowledge during EGD, we showed that computer’s diagnostic performance can be increased significantly. Vice versa, also the expert’s diagnostic performance is improved most of the time depending on the level of
expertise. However, the issue remaining is that such a hybrid system should be trained for each individual expert to obtain the optimum performance. At least the overall performance of the rater should be known beforehand. This prerequisite definitely constitutes an obstacle for deployment in clinical systems.

5. Discussion

In this paper, we summarize recent work on automated diagnosis systems for the diagnosis of CD. The aim of the diagnosis systems is to improve the clinical workflow and especially to increase the final accuracy on CD diagnosis. Although not perfect, we notice generally good classification performance. However, the systems are not yet applied in clinical practice. In the following we will discuss reasons for this lack of application in the routine as well as further trends in this field.

In the introduction we already listed the different application scenarios for automated diagnosis systems. We will now discuss which application scenarios are already covered with the present state of research and which applications scenarios still need future work to be applied in clinical practice.

The first application scenario is the endoscopist’s option to trigger a classification process for specific mucosal regions of interest. The CNN based methods presented in this work as well as previously published methods for the classification of CD are directly applicable for this type of application and the obtained accuracies are high. So we are confident, that this scenario is already covered by the current state-of-the-art machine learning approaches.

The second application scenario is to provide the histopathologist one single “second opinion” based on all digital image material that is captured during an esophagogastroduodenoscopy. Basically all machine learning approaches developed for the first application scenario can be applied for this scenario using the framework for classification on patient level presented in [48]. It was already shown in [48] that classification on patient level achieves good results with CNNs. Furthermore, it was shown in [34, 35, 36] that CNNs outperform hand-crafted image representations. So applying CNNs using the framework for classification on patient level is state-of-the-art and achieves good results for the automated diagnosis of CD. A remaining question is how the quality of the captured image data affects the automated diagnosis systems. We noticed that the quality of the endoscopic image material varies between physicians.
The impact of such a domain shift has not been investigated so far. However, due to experience with modern machine learning approaches, we expect that changes in distribution can be learned if sufficient training data showing a high variability is available. A further relevant open question is, what the physician can do to optimize the final classification accuracy by generating image data which is specifically appropriate for computer aided diagnosis. Finally, diagnosis systems as proposed in literature only provide one single final decision (in this case ‘diseased’ or ‘healthy’) without any reasons for the decision. If medical doctors should rely on such a binary decision, we definitely need to supply further information to make the system more transparent. We could, e.g. show selected “informative” frames showing high quality data in combination with the single decisions as well as confidences. To make such a system transparent, it could also be combined with the third application scenario explained in the following paragraph.

The third application scenario is a continuous automated mucosa analysis during endoscopy. This scenario could help the endoscopist to detect CD affected areas and to find well defined areas to take biopsies. So far, no methodologies were presented to enable a continuous automated mucosa analysis. A challenge regarding this application scenario is given by the facts that large parts of the videos are out-of-focus and that the rapid movements of the endoscope causes motion blur and fast changes between in- and out-of-focus frames. So prior to the application of classification methods, a selection must be performed that identifies the informative parts of the video and ignores all parts of the video with insufficient quality. One problem that complicates this pre-selection is that images showing mucosa affected by villous atrophy looks very similar to out-of-focus images showing (or rather hiding) healthy mucosa [18]. This makes it really hard to distinguish between regions that were captured out-of-focus (and major parts of videos are captured out-of-focus and hence appear blurry) and regions that really show villous atrophy. Although image-based measures were proposed to determine the image quality [47], for a reliable measurement the video stream need to be analyzed in order to make use of the temporal dimension.

Working with video material could definitely be a great opportunity to increase classification accuracies, to determine image quality as well as to increase the amount of training data by including endoscopic sequences or endoscopic frames as training data additionally to the so far collected patches. Noticing that data fusion has a positive impact, we expect a further gain for all application scenarios by considering the third temporal dimension as well.

6. Conclusion & Future Work

In recent work, we investigated the capability of state-of-the-art neural network approaches for diagnosis of CD and proposed pipelines for fully-automated patient-wise diagnosis as well as for integrating expert knowledge into the automated decision process. To go one step further towards clinically deployable
systems, focus need to be on developing transparent pipelines supporting the endoscopist during the whole clinical procedure and not only providing one single final (binary) decision. Computer and physician need to work hand in hand to obtain maximum diagnosis accuracy and acceptance. We are confident that state-of-the-art machine learning technology can be applied and adapted to obtain a software tool leading to a significant progress in the diagnostic workup.

7. Acknowledgments

This work is partially funded by the Austrian Science Fund (FWF) under Project No. 24366 and by the German Research Foundation (DFG) under grant no. ME3737/3-1.

References


