0

Predicting Pathology in Medical Decision Support Systems in Endoscopy of the Gastrointestinal Tract

Michael Liedlgruber and Andreas Uhl
Multimedia Signal Processing and Security Lab (WaveLab),
Department of Computer Sciences, University of Salzburg
A-5020 Salzburg
Austria

1. Introduction

Since medical endoscopy is a minimally invasive and relatively painless procedure, allowing to inspect the inner cavities of the human body, endoscopes play an important role in modern medicine. In medical practice different organs such as the respiratory tract, the urinary tract, and the female reproductive system are regularly inspected by using an endoscope. Another important field in medical endoscopy, which this chapter focuses at, is the inspection of the gastrointestinal tract (GI tract).

Based on endoscopy of the GI tract, physicians are able to detect severe diseases already in early development stages and therefore the mortality rate for many diseases, especially different types of cancers, has been lowered drastically throughout the last years. Some examples of conditions which are known to be pre-malignant or to increase the risk of cancer in the GI tract are adenomas, Barrett's esophagus, Crohn's disease, celiac disease, GI bleeding, and a Helicobacter pylori infection. The parts of the human GI tract, which are most commonly inspected with an endoscope, are shown in Figure 1.

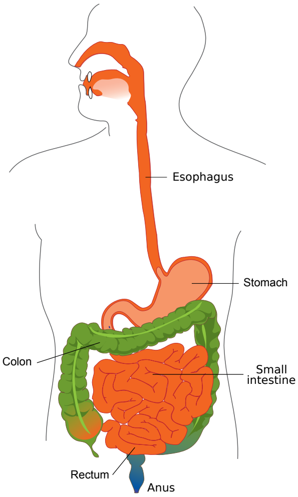
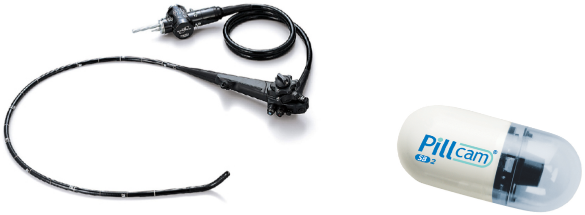


Fig. 1. A schematic illustration of the human GI tract.

In general the area of applications for endoscopes is wide. Besides medical procedures, endoscopes are also used to inspect airplane turbines, pipes in buildings or industrial machinery, car engines, tanks in ships, and for veterinary endoscopy. However, throughout this work the terms "endoscope" and "endoscopy" always denote the medical device and procedure, respectively.

1.1 Technological Advances in Endoscopy

The first time the term endoscope was used was in 1806, when Philipp Bozzini developed the first kind of endoscope which he called "Lichtleiter". By using this device he already made the first attempts to examine the inner cavities of the human body. But endoscopes as we know them today significantly differ from the one Bozzini developed. In the early days of endoscopy the devices were lit by external light sources (a candle in the case of by Bozzini's apparatus) and not flexible. Thus, these devices were somewhat limited in terms of their usability.



(a) Flexible endoscope

(b) WCE capsule

Fig. 2. (a) A flexible endoscope (Image courtesy of Olympus) and (b) an example of a WCE capsule (Image courtesy of Given Imaging).

Endoscopy, as we know it today, is performed using a flexible endoscope (Figure 2(a)), sometimes also referred to as videoscope. This type of endoscope has been introduced in the mid 1960s. While the first endoscopes used fiber optics and an eyepiece lens to visualize the inner cavities of the human body, modern endoscopes are very compact devices, including a light source, and a CCD or CMOS chip for taking pictures. But the basic concept did not change very much since those days. In addition to the digital imaging chip, modern endoscopes contain a light source at the distal tip and are equipped with an accessory channel, which allows the entry of medical instruments for example to take tissue samples, perform cleansing of poorly prepared areas, perform polypectomies, and perform endoscopic resections without any invasive surgery involved.

More recent advances in endoscopy are zoom-endoscopy and chromoendoscopy. Zoom-endoscopy allows to zoom in at regions of interest, using a magnification factor of up to 150. Such devices offer a significant advance since smaller and finer details in the region to be examined get uncovered (Hurlstone et al., 2004). Another possibility to obtain images with a higher level of detail are high definition (HD) endoscopes, which also provide images of higher resolutions and therefore allow to detect subtle changes in the mucosa. Chromoendoscopy aims at enhancing superficial patterns on a mucosal layer by topically applying color dyes. An alternative to this rather time-consuming procedure is to use Narrow Band Imaging (NBI), which allows to enhance the contrast of vascular patterns on the mucosal surface (Emura et al., 2008). Since NBI is based on a rotating filter in front of the light source (narrowing the spectrum of the visible light to bands of blue and green) this technique is not dependent on applying color dyes. Other systems similar to NBI, like FICE (Fujinon Intelligent Chromoendoscopy) or I-scan, use computer algorithms to post-process endoscopic images. Systems like NBI, FICE, or I-scan are referred to as "virtual chromoendoscopy".

Another recent advance in endoscopy is confocal laser endomicroscopy (CLE) (Kiesslich and Neurath, 2007). This procedure allows to inspect the mucosal surface in a highly detailed manner. This is achieved by a laser-based endomicroscope which scans the surface of the

mucosa and even allows to inspect sub-surface features up to a depth of 250 microns by adjusting the focal point of the laser. The resulting images have a resolution corresponding to a magnification factor of 1000, making "smart" biopsies possible, thus avoiding random and possibly unnecessary biopsies. It has already been shown that the diagnostic accuracy of CLE is comparable to histology (Buchner et al., 2010).
Since the small intestine is very long and convoluted a traditional flexible endoscope is only of limited use. A recently developed technique to overcome this limitation and to make endoscopic procedures more safe, less invasive, and more comfortable for the patient, is wireless capsule endoscopy (WCE) (Coimbra et al., 2007). To perform WCE the patient swallows a small capsule (Figure 2(b)), containing a light source, lens, camera, radio transmitter, and batteries. Propelled by peristalsis, the capsule then travels through the digestive system for about eight hours and automatically takes more than 50 000 images. These are transmitted wirelessly to a recorder worn outside the body. Throughout the last years WCE has already proven to be valuable tool to detect the cause of gastrointestinal bleeding within the small bowel (Eliakim, 2004). Recently also other areas of interest for WCE within the GI tract have emerged, such as the colon (Fireman and Kopelman, 2007) or the esophagus (Eliakim et al., 2004). Although WCE currently lacks the ability to treat lesions, obtain biopsy samples, and clean poorly prepared areas, this new technique has already proven to be an effective diagnostic modality for detecting small bowel tumors and small bowel lesions (Cobrin et al., 2006), and may also become an important tool to detect other abnormalities in the GI tract (El-Matary, 2008).

1.2 Computer-aided Pathology Prediction

Due to the digital imaging chips used in modern endoscopes such devices are also regularly used to take digital pictures and record video sequences. These abilities created the whole new field of computer-aided decision support systems (CADSSs) in medical endoscopy. The aim of such systems is to predict pathologies and thus to assist a medical expert in improving the accuracy of medical diagnosis (Doi, 2007). Hence, the development of such systems is motivated by the following key aspects:

- **Saving time and reducing cost**
  CADSSs usually help to identify regions which may be of particular interest for a medical expert. Since, in general, such systems are designed to detect abnormal changes within the GI tract (e.g. neoplastic or metaplastic changes) they are also able to help avoiding possibly unnecessary biopsies, allowing to perform targeted biopsies. In the course of such targeted biopsies the time needed for an endoscopy procedure may be lowered drastically. As a consequence the procedure gets more comfortable for the patient. Furthermore, since the number of necessary biopsies to be taken can be reduced, the time needed for the subsequent histopathologic examination can be reduced too.

  Reducing the time needed for an endoscopy and the subsequent histopathologic examination also results in a reduction of costs associated with such procedures.

  Considering the fact that re-investigating a video recorded during an endoscopy session may consume as much time as the endoscopic procedure itself, there is a potential to save time and costs if CADSSs are able to identify parts of such videos which might be of interest for a medical expert. Especially in case of WCE such systems are of particular use, since, due to the vast amount of images generated per WCE session, inspection of such videos is time consuming and therefore expensive in terms of the time raised by a medical expert (Swain, 2003).

- **Enhancing accuracy of diagnosis**
  Endoscopy is a tedious procedure, demanding a constantly high level of concentration by an endoscopist. This is mainly due to the fact that missing an abnormality during endoscopy may be hazardous. Especially lesions which are rather small or only noticeable for a short fraction of time may be missed easily. This particularly applies to WCE, where a lesion may show up in a single frame only (out of more than 50 000 frames in total!). This is due to the rather low frame rate of currently available capsules which usually is about two frames per second.

  The advantage of CADSSs is that computers always exhibit the same level of "concentration" – no matter how long an endoscopic procedure lasts or how many of such procedures a computer has to analyze in series. Also situations with bad light conditions, poor image quality, or poor contrast usually pose a problem for a medical expert, increasing the risk of missed lesions. If designed properly, a computer-based systems may be able to cope well with such circumstances. The constant level of "concentration" and the resistance against poor image conditions may help to avoid missing lesions, leading to an enhanced diagnostic accuracy of an endoscopic procedure.

  But, as indicated in (Church, 2008), sometimes there is a more simple reason for missed lesions: an abnormality may be simply get missed since it is misinterpreted and therefore not recognized as being abnormal. The likelihood of missing a lesion for this reason heavily depends on the expertise of the medical expert who performs the endoscopy, leading to a lowered inter-observer agreement level. CADSSs, on the other hand, do not suffer from this problem. If the detection of abnormalities is designed in a robust way and there is sufficient training data available, a CADSS is expected to always deliver roughly the same level of accuracy.
- **Training of experts to new endoscopic imaging modalities**
  As already pointed out in Section 1.1, there have been many advances in endoscopy within the past few years. While the new imaging modalities have the potential to greatly increase the efficiency of endoscopy, medical experts need to get trained on these new techniques. In order to assess the skills of a medical expert on new techniques, CADSSs can be used as an expert training tool to predict pathology, verify the detection or prediction performance of a medical expert, and serve as an educational resource.

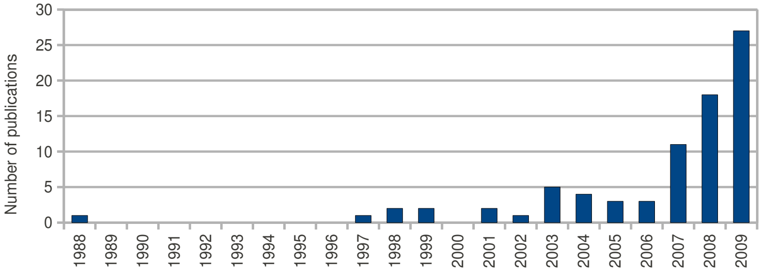


Fig. 3. Number of publications between 1988 and 2009 found on PubMed and ScienceDirect when searching for publications targeting at computer-aided decision support for medical endoscopy in the GI tract (search was conducted in 2010, hence, publication from 2010 are not considered).

To highlight the relevance of CADSSs we conducted an exhaustive search for publications dealing with this topic (on PubMed [1] and on ScienceDirect [2]), which yielded the search results presented in Figure 3. The results show that there is a rising interest in this research topic, starting about one decade ago.

2. Fundamentals of Pathology Prediction

Basically there exist two different approaches for CADSSs aiming at pathology prediction: medical image classification (MIC) and content-based image retrieval (CBIR) (Müller et al., 2004). While these two concepts share some basic building blocks, there are also fundamental differences. In the following section we therefore explain the parts each of these concepts consists of. In addition, we discuss the similarities as well as the differences between CBIR and MIC.

2.1 Building Blocks of CBIR and MIC

CBIR systems as well as MIC systems both basically consist of two steps. The first step is to obtain training data along with ground truth information, which represents the knowledge of the respective system. Based on the training data, the prediction is carried out.

2.1.1 Obtaining Training Data

No matter, whether we are dealing with a MIC or CBIR system, in order to be able to perform a prediction of pathologies, in both cases the first step is the generation of training data along with known ground truth. The respective steps are illustrated in Figure 4.

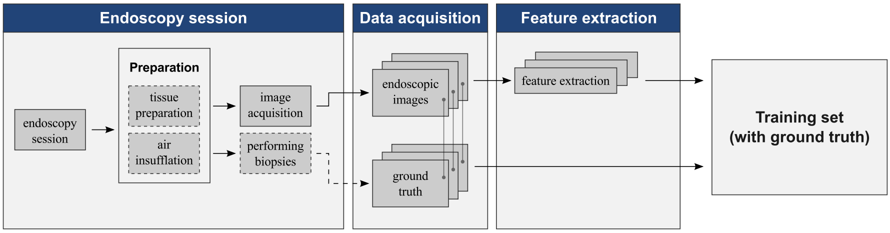


Fig. 4. Schematic illustration of creating training data with ground truth.

It must be noted however, that some of these steps are not possible in case of capsule endoscopy. Since these devices currently offer no possibility to obtain tissue samples, a histopathologic examination to establish a ground truth is not possible. Hence, in case of WCE, the ground truth is usually obtained by a visual inspection of the images recorded by a medical expert. Apart from that, air insufflation is not possible either. These steps are therefore denoted by dashed lines in Figure 4.

- **Preparation steps (not available in case of WCE)**
  Before actually being able to capture images and taking tissue samples, certain preparation steps may be necessary. One such step is air insufflation, which is performed to expand the wall of the part of the GI tract under inspection. This makes regions of

[1] PubMed located at `http://www.ncbi.nlm.nih.gov/pubmed`
[2] ScienceDirect located at `http://www.sciencedirect.com`

the mucosal wall visible which otherwise would be hidden behind folds and therefore increase the probability of missed lesions. Another step, performed on a regular basis, is the preparation of the tissue under investigation. In order to cleanse poorly prepared areas water is used to flush away e.g. stool residuals inside the colon or undigested food.

Another possibility for tissue preparation is the topical application of color dyes in order to enhance the superficial structure of the mucosa in the region under investigation (e.g. vascular patterns). If a NBI-enabled endoscope is used this step can be omitted since NBI allows to enable and disable the structural enhancement with a simple tip of a button, hence, no color dyes are needed.

- **Capturing images and obtaining ground truth information**
  Images are taken with the tip of a button on the endoscope or, in case of WCE, continuously. When performing endoscopy using a flexible endoscope, also a biopsy is taken in the region covered by the image just taken. After the endoscopy session, the resulting tissue sample is examined by a pathologist under a microscope. In the course of this examination the pathologist determines the histopathologic type for the respective biopsy (e.g. assessing whether there is a malignant potential or not). This classification by the medical expert serves as ground truth for the training data.

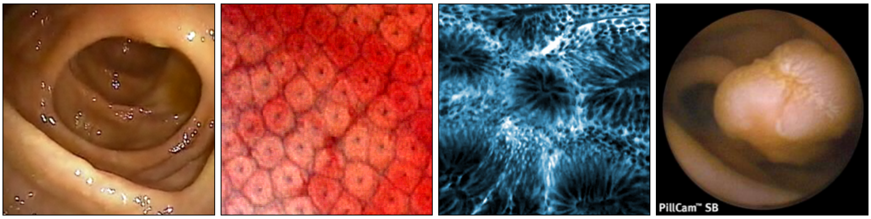
  Due to the limitations of WCE in terms of taking biopsies, the respective ground truth information is usually obtained by visual means.
- **Feature extraction**
  Since images represent rather high dimensional data, suitable features have to be used to reduce the amount of data to be stored in the training set.

  While there is a wide range of feature types used throughout CADSS-related work found in literature, these can be roughly grouped into two categories:

  - **Low-level features**
    Feature types falling into this category are based on the raw information contained within images. Examples for such features are histogram features (e.g. simple color histograms and co-occurrence matrices), features capturing the textural contained within images (e.g. Local Binary Patterns and Markov Random Fields), or statistical features (e.g. computed from histograms).

    But also features based on some sort of frequency-domain transform are used quite often throughout literature. Examples are the wavelet transform or the Fourier transform (e.g. features based on raw coefficients or statistics computed from the respective coefficients, e.g. Häfner et al. (2010a)).
  - **High-level features**
    Approaches found throughout literature are also quite often based on features describing the content of images in an abstract way. Features belonging to this category are for example shape-based features, describing shapes obtained by different edge-detection methods (e.g. Häfner et al. (2010b)).

  But the final choice on the types of features to be used usually depends on the type of pathologies the respective system should be able to deal with and eventual constraints posed to the system (e.g. time constraints). But also the endoscopic imaging modality (e.g. WCE, flexible endoscopy, CLE) the system is designed for plays an important role, as different imaging techniques may result in considerably different types of images. To emphasize on this, Figure 5 show examples taken with different types of endoscopes.



(a) (b) (c) (d)

Fig. 5. Images acquired by using different endoscopic techniques (a) endoscopy (Kelsey, 2005), (b) zoom-endoscopy, (c) confocal laser endomicroscopy (Kiesslich, 2007), and (d) WCE (Copyright © 2005-2010, Given Imaging. All Rights Reserved).

2.1.2 The Pathology Prediction

While obtaining training data with ground truth is identical, no matter whether a CADSS is based on CBIR or MIC, the actual prediction differs in some aspects. The different steps required for pathology prediction are outlined in Figure 6 (similar to Figure 4, the steps not available in case of WCE are denoted by dashed lines).

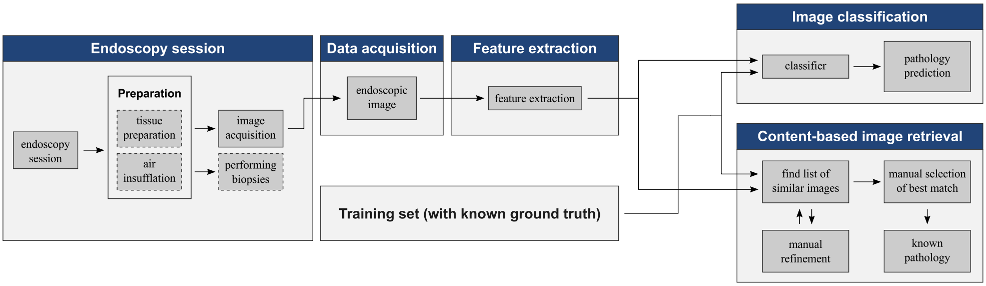


Fig. 6. Schematic illustration of performing pathology prediction for an image with unknown pathology (using either classification methods or image retrieval methods).

As we notice from this illustration, the steps to generate an image to be classified are very similar to the steps already outlined for obtaining training data. However, while biopsies will most likely be performed too, the histopathologic classification remains unknown to the system (since the histopathologic examination takes place after the endoscopy session, the classification result is not available during the endoscopy session anyways).

After obtaining an image, the same type of features as used to generate the training set is extracted from the image. Then, along with the training set composed earlier, the pathology is predicted either using MIC or CBIR. The steps for the respective system types can be outlined as follows:

- **Classifier-based approaches**
  The basic idea behind classifier-based approaches is to provide a second opinion to a medical expert. Given an unknown image, this is achieved by predicting the pathology in a fully autonomous manner.

In case of MIC the first step is to train a classifier using the training set available. Then the classifier is used to predict the pathology based on the features extracted from the image with unknown pathology.

Throughout literature dealing with CADSSs targeted at medical endoscopy of the GI tract various different types of classifiers have been used for this task (e.g. k-nearest-neighbors classifier, Support Vector Machines, artificial neural networks, Bayesian classifier, Linear Discriminant Analysis classifiers, or simple clustering-based classifiers) (Duda et al., 2000; Fukunaga, 1990). The choice for the classifier to be used mostly depends on the features used (and the according feature dimensionality). While also time-constraints may play an important role in choosing the classifier (since some classifiers have a rather high complexity and, therefore, a high computational demand), each type of classifier has advantages and disadvantages which must be considered too.

The classification outcome of the classifier then corresponds to the predicted pathology.

Due to the fact that classifier-based approaches are operating without any intervention needed, such approaches also potentially suited for an online-diagnosis (i.e. predicting the pathology during endoscopy already).

- **Systems based on content-based image retrieval**
  CBIR can be regarded as the digital counterpart of current clinical practice, where medical experts compare cases with unknown pathology to cases with an already known and verified diagnosis. Since, however, searching through an existing database, containing a huge number of cases, is a time-consuming task. As a consequence, performing the search for similar cases or images in an automated fashion helps to save time and greatly simplifies the work of medical experts.

  For CBIR-based systems the first step is to choose a suitable metric which determines the similarity between two images (or the respective features). Based on such a metric the system returns a list of the images most similar to the image with unknown pathology, along with the respective similarity scores. The result can then be refined by a medical expert by choosing one or more images which seem to be relevant (commonly termed ”relevance feedback”). Then a new query is started based on the images marked as being relevant.

  Once, the medical expert has found the image which, in his opinion, matches the unknown image best, the pathology of the best match can be used in order to determine the pathology for the image with unknown pathology.

  The fact that in case of CBIR-based systems an intervention by a medical expert is needed restricts such systems from being used to obtain a diagnosis during endoscopy already (i.e. online-diagnosis), making CBIR useful for offline-diagnosis only.

Despite the different prediction strategies in case of MIC and CBIR, in certain cases both methods are very similar. If a MIC-based system employs the k-nearest-neighbors classifier with $k > 1$ (the number of neighbors to consider for the prediction process) MIC is very similar to CBIR. This is because the neighbors determined by the classifier basically correspond to the most similar images compared to the incoming images. These k most similar images, however, can be considered to be the k best matches in context of CBIR. But while in case of MIC the classifier determines the final prediction (based on some sort of voting), in case of CBIR the prediction is based on a manual selection of the best matching image.
One of the main differences between MIC and CBIR is therefore the fact that, while MIC performs the prediction in a fully autonomous fashion, CBIR allows a medical expert to refine

the prediction process. In case of CBIR, it is up to the medical expert to choose the final match (along with the respective known pathology) and therefore make a judgment on the pathology of the unknown image.

In addition both types – MIC and CBIR – significantly differ in terms of the possible application scenario. While MIC-based systems allow to perform a real-time diagnosis, CBIR-based systems are restricted to be used in offline-scenarios only. This restriction is imposed by the interactive nature of CBIR. As a consequence, interacting with a CBIR system during an endoscopy procedure would constrain the endoscopist (in particular if the expert needs to provide relevance feedback to the CBIR system).

2.2 Detection vs. Classification

Despite the differences already highlighted, which exist between MIC-based system and CBIR-based systems, the goals of such CADSSs fall into two categories.

The first branch aims at the detection of abnormalities without making a prediction about the respective pathology. Such systems are for example used to detect polyps or adenomas within the GI tract. If such an abnormality has been detected it is up to the medical expert to decide upon determining the respective pathology. Another field of applications for such systems is to find abnormalities within endoscopy videos, usually containing a huge number of frames. The result is a subset of frames, showing abnormalities which might be of particular interest for a medical expert.

But the frames found might also be used as input for another system, which is specifically designed to assess the malignant potential of a potential abnormality found. Such systems fall into the second branch. In contrast to the detection systems the classification systems are usually not able to locate abnormal pathologies. Moreover, they rely on input which already contains the abnormality candidate to investigate (e.g. images showing polyps). Based on such abnormality candidates, these systems are designed to perform a classification according to some medical classification scheme (e.g. classifying the colonic mucosa as being normal, hyperplastic, neoplastic, or metaplastic). Hence, the outcome of such systems is the prediction of the pathology according to an input image.

Depending on the desired application area, a system performing either detection or classification is needed. For a fully automated pathology prediction for endoscopic imagery or videos, however, a combination of both system branches is needed.

2.3 Confidence in Prediction System

When it comes to computer-based systems, designed to assist a medical expert in the course of obtaining a diagnosis, an important question arises: how trustworthy are such systems from the perspective of medical experts?

Due to its interactive nature CBIR allows to roughly follow the steps from an input image to the final match. Since the medical expert interacts with such a system, the final outcome of the system is at least partially comprehensible. The expert is also able to steer the search into another direction, which probably seems to be more appropriate to the medical expert, based on his medical know-how.

In case of classifier-based systems, however, the actual process of pathology prediction is some sort of a black box. The system is given an image and returns a prediction on the pathology of the image (either a detection or a classification result). As a consequence there is no intermediary medical expert, who is able to complement the system in terms of his experience.

3. Pathology Prediction Issues

In order to be of practical use, the quality, reliability, accuracy, and usability of pathology prediction systems are important issues. The following section therefore contains a discussion of these issues.

3.1 Obtaining Ground Truth Information

As already indicated in Section 2, there basically exist two different ways of obtaining ground truth information: either by visual means or through a histopathologic examination of biopsies.

Obtaining the ground truth for an image by visual inspection by a medical expert has the disadvantage that there is no profound knowledge about the real pathology for that image (i.e. the pathology is not verified histologically). Another disadvantage is the fact that in the course of visual pathology assessment the inter-observer agreement rate is likely to be lower, resulting in different diagnosis results. This is due to the fact that the judgment of different medical experts heavily depends on the respective level of experience.

If samples of suspicious tissue are taken the ground truth can be based on the outcome of a subsequent histopathologic examination. But the biopsy site does not need to perfectly correspond to the respective image taken. This may be due to peristaltic motion or slight movements of the endoscope tip (for example in the course of preparing a biopsy), which are especially noticeable in case of high-magnification endoscopy. Hence, while the histopathologic classification can be considered to be the gold standard, the respective pathology is not necessarily visible on the respective endoscopic image.

However, the choice on the way to obtain the ground truth often is not upon the medical expert. As already mentioned earlier, in case of WCE-based systems there is usually no other option than to rely on visual inspections by one or more medical experts, since taking biopsies is not possible with current capsule endoscopes. In case of flexible endoscopy the ground truth can be gathered histologically since taking biopsies is possible.

A special case is constituted by CLE since this technique allows in-vivo histologies due to the high level of magnification. As already mentioned in the introduction, it has already been shown that the diagnostic accuracy of CLE is comparable to histology. Hence, the inter-observer agreement is also expected to be similar to the agreement in case of histology.

Considering the CADSS approaches found throughout literature which are based on flexible endoscopy, about 12% of the methods base their experiments on a visually obtained ground truth, while the vast majority of the methods is based on histological findings (about 68%). The remaining approaches are not accompanied by the according information.

Making a recommendation concerning this issue is not easy, since the best way of obtaining the ground truth information very much depends on the endoscopic technique used. While in case of WCE a visual inspection is usually the only way a ground truth can be obtained, in case of CLE a visual ground truth gathering is likely to be sufficient due to its closeness to histology. In case of traditional flexible endoscopy a histological ground truth is highly desired due to its accuracy over visual inspection (in terms of the histopathologic classification).

3.2 Training and Validation Data

When developing a system aiming at pathology prediction, an important aspect is the quality of the imagery used to validate the overall effectiveness of the system.

To assess the accuracy and robustness of a CADSS some sort of image database is needed in order to be able train and validate the system. But the explanatory power of the quality assessment is directly dependent on the quality of the image database set used. The following factors are of particular importance in order to be able to compose a solid image database.

- **Size of training and validation sets**
  An important factor for the quality of an image database is the number of images available for training and validation. A low number of training images may limit the ability of the system to generalize to the prediction problem. As a consequence the probability of a wrong prediction result for new image samples is likely to be higher compared to an image database which contains a sufficiently high number of training samples.

  If the number of validation images is too low it does not matter how many training samples were used, the resulting validation accuracy computed from the data would be rather questionable (due to a rather low informative value).

- **Balance between different image classes**
  In general, image databases used in CADSSs consist of two or more image classes, each denoting a particular pathology. To avoid overfitting the system to a particular pathology the number of images used for the different classes should be balanced.

  In case of an imbalanced training set the CADSS used may have problems to adapt to the image database, resulting in a higher likelihood of assigning unknown images to the image class which contains the most images within the training set (which might result in a wrong prediction). If the validation image set is imbalanced the different accuracy estimations may vary significantly across the classes in terms of their significance. This applies to MIC systems as well as to CBIR systems.

  In case of classifier-based systems the overfitting to the dominant class (in terms of the number of images available) happens in the underlying classifier already. But the likelihood of running into this kind of problem also depends on the classifier used. While there exist classifiers which are more resistant against imbalanced training sets (e.g. the Support Vector Machines classifier), others are not able to cope well with this kind of problem (e.g. the k-NN classifier).

  Depending on the degree of imbalance and the number of results to be returned from the system, a CBIR-based system is also likely to return a list of wrong matches – in favor of the class containing the most training images. But since a CBIR system usually also returns a similarity score along with the retrieval result, wrong matches can be identified more easily.

Another problem is a rather high feature vector dimensionality compared to the number of training images available. In this case a problem called the “curse of dimensionality” (Bellman, 1961) might arise. This problem is caused by the fact that high-dimensional feature vectors are more likely to be sparsely distributed in the feature space if the pool of training images is not sufficiently large. As a consequence the system might loose its ability to generalize to the prediction problem and overfitting might occur.

A possible solution to the “curse of dimensionality” are feature subset selection algorithms, which, based on some selection criterion, reduce a given feature set to a nearly optimal subset (Jain and Zongker, 1997). But, depending on the dimensionality of the original feature set, the desired target dimensionality, and the selection algorithm chosen, the computational burden imposed to the training process of the system might get considerably high (which, however, has no impact on the computational performance during prediction).

3.3 Pathology Prediction Accuracy

Before a CADSS can be used in clinical routine, it must be ensured that certain accuracy requirements are fulfilled.

3.3.1 Commonly Used Validation Protocols

To measure the accuracy of a system usually different strategies can be employed. While the best choice is to use separate training and validation image sets to assess the accuracy of a pathology prediction system (as implicitly assumed in Section 3.2), this is sometimes not possible due to the limited number of images available in the image database used.

Anyhow, in such cases different possibilities exist to be able to assess the system accuracy of a pathology prediction system (Duda et al., 2000). In literature, dealing with CADSSs in medical endoscopy of the GI tract, the most commonly used approaches for system validation are Leave-One-Out cross-validation, Leave-One-Patient-Out cross-validation, and k-fold cross-validation. A schematic illustration of these validation methods is provided in Figure 7.

- **Leave-One-Out cross-validation (LOO-CV)**
  In case of LOO-CV one image out of the used image set is considered to be an unknown sample. The remaining images are used to train the classifier or, in case of CBIR, simply serve as the image database available for retrieval queries. Then the predication step is carried out for the image left out (the unknown sample). These steps are repeated for each image in the image database, resulting in an estimate of the prediction accuracy of the system.
- **Leave-One-Patient-Out cross-validation (LOPO-CV)**
  LOPO-CV is rather similar to LOO-CV. But in contrast to LOPO-CV a more tight restriction is posed to the training (or CBIR database composition) in the cross-validation process. In addition to the image left out for training, all images originating from the same patient are left out too. The prediction and final estimation of the prediction accuracy are then carried out the same way as in case of LOO-CV.

  Sometimes a slightly relaxed variation of LOPO-CV is employed. Instead of restricting the training to images originating from a different patient than the image to be classified originates from, only images from the same pathology (e.g. polyp) must not be used during the training phase. Hence, when training the system, images from the patient currently subject of classification may be included in the training set as well (as long as these do not show the same abnormality). In case of polyps, this is sometimes referred to as Leave-One-Polyp-Out cross-validation.
- **k-fold cross-validation (KF-CV)**
  This type of cross-validation is basically a generalized version of LOO-CV. For KF-CV the set of available images is partitioned into k different subsets. Then, while the samples from one subset are used for the prediction, the samples from the remaining subsets are used to train the classifier in case of MIC-based systems (or compose the training image database for CBIR). This is repeated for each partition available. As a consequence, the training only considers images from other subsets than the one currently under classification. For k equals the number of images in the image database, KF-CV reduces to LOO-CV. If the partitions are chosen to be disjoint in terms of the patients, KF-CV equals LOPO-CV.

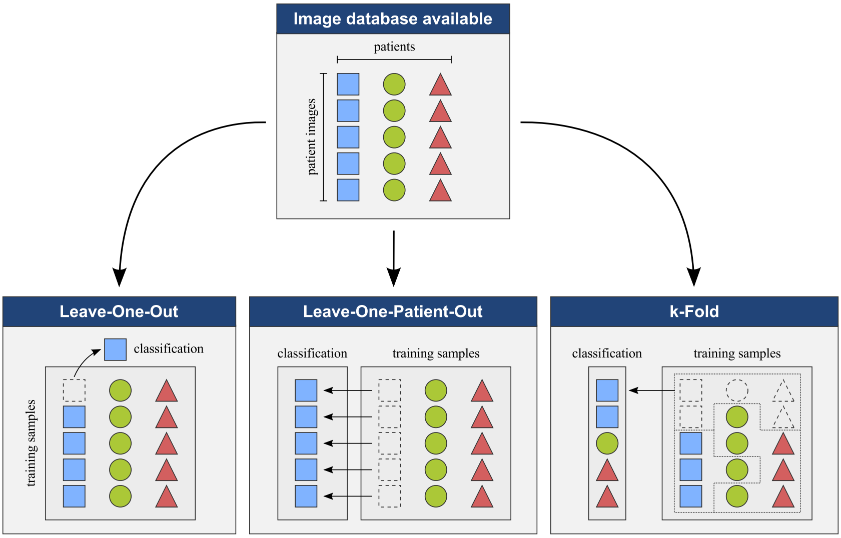


Fig. 7. Schematic illustration of the different cross-validation methods described in the text.

Despite the fact that cross-validation is a convenient way of dealing with a limited number of images available, these methods also have one common drawback, at least in case of classifier-based systems. Since the training set changes with each image, patient, or subset left out during cross-validation, the classifier used must be retrained. Depending on the classifier, this may result in a considerable increase of the computational demand. But at least in case of KF-CV this can be controlled by adjusting the number of partitions accordingly. However, while in case of KF-CV a higher number of partitions means a reduced loss of training data, this also increases the computational demand for training (since training has to be performed more often in this case).

Another problem, especially noticeable in case of LOPO-CV and KF-CV, is the reduced training data available under certain circumstances (this applies to MIC-based systems as well as to CBIR systems). For LOPO-CV this problem emerges in case of a high number of images per patient. As a consequence, leaving out one patient from training may result in a considerable drop in terms of the training images available. Since a rather low number of partitions used in case of KF-CV results in a high number of samples within each partition, the training set gets considerably smaller if one such partition is left out during training. As a consequence, the system might be unable to generalize to the prediction problem stated. The partitions should also be chosen in such a way the the number of images available for each image class is well balanced throughout the partitions used for training.

For LOO-CV the training set reduction is usually not noticeable. However, when using this cross-validation protocol, the increase of the computational demand for training a classifier in MIC systems can get especially noticeable. In addition, the fact that the training set may contain images similar to the image for which the pathology should be predicted for, the accuracy outcome may have a substantially lowered informative value (since there is no restriction imposed on training images, neither in terms of the patient nor in terms of pathologies of single patients).

Considering the different properties of the cross-validation protocols just discussed, the best choice would be to use LOPO-CV. This way, compared to LOO-CV, the danger of overfitting reduced to a great extent. While KF-CV must also be favored compared to LOO-CV, the partition of the image set must be performed carefully in order to have roughly the same number of images from each image class within each partition. In addition the partitions should be chosen such that they are disjoint to a maximum possible extent in terms of patients (or pathologies of single patients).

3.3.2 Explanatory Power of Results

While it is important for a pathology prediction system to exhibit a certain level of accuracy to be of practical use, it is also important to be able to compare the accuracies of a system to accuracies reported for other approaches from literature.

The first issue to be addressed is therefore the choice of the right measures, which must have enough explanatory power in order to make the results reported useful. Throughout literature, dealing with automated pathology prediction in endoscopy of the GI tract, the use of different measures can be observed. In case of MIC-based systems usually one or more of the following measures are used:

- **Overall accuracy**
  The overall accuracy of a system is computed by dividing the total number of classified correctly by the total number of images available in the image database used. While the overall accuracy gives an overview of the overall performance of a system, there is no evidence about the classification performance for the different image classes.

- **Specificity and sensitivity**
  These measures are only applicable if the prediction should be carried out for two image classes only (e.g. is a polyp visible or not on a given image? shows a given image neoplastic tissue or not?). The specificity, also known as true negative rate, corresponds to the number of images taken from healthy patients, which have been predicted as belonging to the image class containing the images from healthy patients. The sensitivity, also known as true positive rate, on the other hand, denotes how many images, showing abnormal pathologies, have actually been classified as belonging to the class of sick patients.

  Usually both measures are given together, as providing only one of both measures may lead to a misinterpretation of the results. If for example a naive classifier is used (which classifies all images as belonging to the class of healthy patients) the respective specificity would equal 100%, but the sensitivity would equal 0%. Hence, in this case, providing the specificity only could lead to the conclusion that the classification performance is perfect. However, when providing both measures together it is obvious that the classifier performs rather poor.

- **Class accuracies**
  Computing accuracies for each image class is necessary in cases where the image databases consists of more than two classes (e.g. distinction between normal tissue, hyperplastic tissue, neoplasms, and metaplastic changes). Hence, the accuracy for a single class is determined by dividing the number of images correctly classified as belonging to the class by the total number of images contained in the class.

- **Area under curve of a ROC plot**
  In ROC plots (Receiver Operating Characteristics) the sensitivity of a system is plotted

against the false positive rate (1-specificity). When computing the sensitivity and the specificity for different choices of parameters inherent to the respective system, the result is a plot. This plot is of particular interest, especially for medical experts, since an investigation of the plot allows to get an idea of the overall system performance. In addition, by computing the area under the curve (AUC) (Hanley and McNeil, 1982), the overall performance of a pathology prediction system can be expressed with a single value.

Compared to MIC-based systems, CBIR-based systems are usually judged by precision and recall (but sometimes also measures typically used in case of MIC-based systems are given). Precision is computed as the number of relevant images in the retrieved list of images divided by the total number of retrieved images, where relevant images are those which should be retrieved since they are of particular interest. Recall, on the other hand, is computed as the number of relevant images contained in the retrieval result divided by the total number of relevant images (in the context of a classifier the recall rate is equal to the sensitivity). Similar to ROC plots plotting precision against recall leads to the PR plot, which allows to get an idea of the overall retrieval performance of the prediction system.

To compare a pathology prediction system against others in terms of accuracy it is not sufficient to just compare the methods in terms of the different measures available. Even if two approaches deliver different values for some accuracy measure there is no evidence given that the observed differences are of any significance. Hence, to assess whether two different approaches differ significantly some sort of statistical test has to be employed.

An example for a tool to decide on statistical significance is McNemar's test statistic (Everitt, 1977). This statistic keeps track whether two different approaches classify an image correctly or not. Based on this information, McNemar's test allows to assess the statistical significance of differences observed between two methods. In addition, using McNemar's test, a p-value can be computed, which is used frequently in medical publications and, hence, of particular interest for medical experts.

3.4 Runtime Performance

An important issue, which depends on the target application area, is the computational demand of the CADSS. If a pathology prediction system is to be used offline (i.e. the system is used after the endoscopy procedure) the requirements in terms of the runtime performance are not as tight as this is the case with systems to be used in a real-time environment. This issue is of special importance in case of MIC-based systems since CBIR-based systems are usually designed to be used offline anyways. Nevertheless, even in case of CBIR the system should be designed and optimized such that the prediction can be obtained in a reasonable amount of time. If a query of a CBIR system lasts for several minutes or even hours, the benefit of using the system would diminish or get lost completely.

But if we strive for a MIC system applicable in real-time environments certain time constraints must be met in order to be of practical use. If we assume an endoscopy video to be captured at a frame rate of 25 frames per second, the pathology prediction of the system must be feasible in less than 40 milliseconds for a single frame. If detection of potentially abnormal regions is carried out by the endoscopist, who already provides regions of interest to the MIC system (designed solely for classification, not the detection), the time constraints may be relaxed to a few seconds for a single image. However, if the time constraints are not fulfilled, similar to CBIR, the advantage of using such a system would get lost.

4. Examples from Literature

In the following we compare some recent works found in literature dealing with CADSSs in medical endoscopy of the GI tract. For each prediction type – MIC and CBIR – we present and compare two approaches in terms of issues discussed in the previous sections.

4.1 Classifier-based prediction

In order to compare two classifier-based approaches found in literature we picked two recently published approaches (found in Tischendorf et al. (2010) and Häfner et al. (2010b)). Some facts about these approaches are listed in Table 1.

	Tischendorf et al. (2010)	**Häfner et al. (2010b)**
Endoscope type	Zoom-endoscopy (NBI)	Zoom-endoscopy (Color dyes)
GI tract part	Colon	Colon
Abnormality	Polyps	Polyps
# of classes	2	2
Prediction aim	Neoplastic (Yes/No)	Neoplastic (Yes/No)
Ground truth	Histology-based	Histology-based
Cross-validation	LOO-CV	LOO-CV
# of images available	209	627
# of patients	223	N/A
Overall accuracy	$\approx$ 85%	$\approx$ 93%

Table 1. Some facts about the two examples for MIC-based approaches recently proposed.

While both methods are similar in terms of the GI tract part under investigation, the pathology of interest, the prediction goal (hence, also the same number of classes), the cross-validation protocol used, and the way the ground truth has been obtained, we also notice some differences between these works.

- **Endoscope type**
  Although both methods rely on zoom-endoscopes, the way to highlight mucosal features differs between these works. While in (Tischendorf et al., 2010) NBI has been used, the approach presented in (Häfner et al., 2010b) is based on the application of topical staining (using color dyes), which still is common practice in clinical routine.
- **Number of patients**
  While in (Tischendorf et al., 2010) the experiments are based on images originating from 223 patients, the work presented in (Häfner et al., 2010b) makes no statements about this detail. Nevertheless, as already pointed out earlier, using too many images from one patient may lead to an overfitting. While we know that Tischendorf et al. on average use roughly one image per patient, we are not able to get a picture about the possible degree of overfitting in case of (Häfner et al., 2010b). This is especially problematic since LOO-CV is used to assess the prediction accuracy for the system. Hence, the experiments in (Häfner et al., 2010b) might be overfitted if images from one distinct polyp are used for training and classification.
- **Number of images used**
  Comparing the total number of images between the two approaches clearly shows that

the number of images used in (Häfner et al., 2010b) is roughly three times higher compared to (Tischendorf et al., 2010). Since, as pointed out above, we have no figures about the number of patients in case of (Häfner et al., 2010b), it is not possible to assess whether this is beneficial (since, in case of only a few patients, this would lead to overfitting).

Although, at a first glance, the approach presented in (Häfner et al., 2010b) seems to perform better in terms of the overall prediction accuracy compared to (Tischendorf et al., 2010), a comparison of the overall system accuracies would be meaningless. This is mainly due to the fact that both approaches are based on quite different image databases. In addition, as already pointed out, we have no idea about the degree of possibly happened overfitting in case of (Häfner et al., 2010b).

4.2 CBIR-based prediction

The number of CBIR-based prediction systems found in literature is very limited compared to MIC-based approaches. According to our literature review 4 out of 5 methods found in total, proposed for CBIR-based prediction in case of endoscopy of the GI tract, have been published by the research group around André et al. Nevertheless, we carry out a comparison of two exemplar approaches (found in André et al. (2009) and Münzenmayer et al. (2009)). Some details on these approaches are given in Table 2.

	André et al. (2009)	**Münzenmayer et al. (2009)**
Endoscope type	CLE	Traditional endoscope
GI tract part	Colon	Esophagus
Abnormality	Polyps	Barrett's esophagus
# of classes	2	3
Prediction aim	Neoplastic (Yes/No)	State of epithelium
Ground truth	Histology-based	Histology-based
Cross-validation	LOPO-CV	LOO-CV
# of images available	1036	390
# of patients	52	61
Overall accuracy	$\approx 80\%$	$\approx 81\%$

Table 2. Some facts about the two examples of CBIR-based approaches recently proposed.

While both systems base their experiments on a ground truth obtained by a histopathologic examination and the overall accuracies reported seem to be very similar, a direct comparison of the reported accuracies is not feasible due to some significant differences (different GI tract parts, different abnormalities of interest, and different prediction aims). We nevertheless discuss a few aspects of these approaches:

- **Endoscopy type**
  While both methods rely on flexible endoscopes, in case of (André et al., 2009) a CLE endoscope is used. In contrast to (Münzenmayer et al., 2009) this is an advantage since, as already pointed out in Section 1.1, this allows in-vivo histologies. Nevertheless, André et al. support their finding with a histopathologic examination.
- **Number of images used**
  While the number of patients contained in the underlying image databases is rather high in both approaches, we notice that in case of (André et al., 2009) the number of

total images used is nearly four times higher compared to (Münzenmayer et al., 2009). The implies that the average number of images per patient is approximately 20 and 6 in case of (André et al., 2009) and (Münzenmayer et al., 2009), respectively.

- **Cross-validation used**
  A key difference between the methods compared is the type of cross-validation employed. While in (Münzenmayer et al., 2009) LOO-CV is used, André et al. use LOPO-CV (the images from the patient currently under prediction are left out). Hence, compared to the experiments in (André et al., 2009), the approach validated in (Münzenmayer et al., 2009) suffers from the fact that images from patients currently under prediction might well be contained in the training set too.

Although a comparison would not be meaningful we can at least deduce, that the validation accuracy reported in (Münzenmayer et al., 2009) must be taken with caution, since overfitting might have happened due to the validation-protocol used. André et al. bypassed this problem by employing LOPO-CV.

5. Conclusion and Future Outlook

In the previous sections we gave an overview of CADSSs targeted at pathology prediction in medical endoscopy in the GI tract. We showed that especially throughout the past two decades the is a rising interest in this research topic.

Since throughout literature concerned with this topic usually either MIC-based systems or CBIR-based systems are proposed, we highlighted the similarities and differences between such systems. We outlined the different different aspects of such systems and highlighted different issues inherent to the development of such systems. In addition, we also discussed common pitfalls which have to be considered to develop reliable and comparable prediction systems.

Especially for a potential use in clinical practice it is important that the accuracy of a prediction systems is above a certain level. But, as we also discussed in the previous section, with the danger of overfitting in mind, it is even more important that the accuracy of such systems is validated properly. While the best option is using different training and validation set this if often not possible due to a limited number of images. In such a case it is therefore even more important to choose a suitable validation protocol which limits the danger of overfitting.

We are currently not aware of systems used in daily routine. While the reasons for this are most probably manifold, it is apparent that stronger collaborations between developers of prediction systems and medical experts are needed, as this is the basis for real world testing of such systems. In addition, the feedback given by medical experts is valuable in order to develop more reliable and accurate systems. However, in order to allow meaningful collaborations with medical experts the explanatory power of the prediction results is important to allow the expert to quickly assess the quality of the system.

Considering recent technological advances in case of endoscopy devices and the new possibilities offered by these endoscopic modalities for a computer-assisted diagnosis, it is out of question that automated prediction will get more precise by fully taking advantage of these techniques. Especially, CLE is a prospective candidate to replace time-consuming biopsies and foster real-time diagnosis systems.

Acknowledgments

This work is partially funded by the Austrian Science Fund (FWF) under Project No. L366-N15 and by the Austrian National Bank "Jubiläumsfonds" Project No. 12514.

References

André, B., Vercauteren, T., Perchant, A., Buchner, A. M., Wallace, M. B., and Ayache, N. (2009). Endomicroscopic image retrieval and classification using invariant visual features. In *In Proceedings of the 6th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'09)*, pages 346–349, Boston, Massachusetts, USA.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Buchner, A. M., Shahid, M. W., Heckman, M. G., Krishna, M., Ghabril, M., Hasan, M., Crook, J. E., Gomez, V., Raimondo, M., Woodward, T., Wolfsen, H. C., and Wallace, M. B. (2010). Comparison of probe-based confocal laser endomicroscopy with virtual chromoendoscopy for classification of colon polyps. *Gastroenterology*, 138(3):834–842.

Church, J. (2008). Adenoma detection rate and the quality of colonoscopy: the sword has two edges. *Diseases of the Colon & Rectum*, 51(5):520–523.

Cobrin, G. M., Pittman, R. H., and Lewis, B. S. (2006). Increased diagnostic yield of small bowel tumors with capsule endoscopy. *Cancer*, 107(1):22–27.

Coimbra, M., Mackiewicz, M., Fisher, M., Jamieson, C., Scares, J., and Cunha, J. P. S. (2007). Computer vision tools for capsule endoscopy exam analysis. *EURASIP Newsletter*, 18(1):1–19.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4–5):192–211.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley & Sons, 2nd edition.

El-Matary, W. (2008). Wireless capsule endoscopy: Indications, limitations, and future challenges. *Journal of Pediatric Gastroenterology and Nutrition*, 46(1):4–12.

Eliakim, R. (2004). Wireless capsule video endoscopy: three years of experience. *World Journal of Gastroenterology*, 10(9):1238–1239.

Eliakim, R., Yassin, K., Shlomi, I., Suissa, A., and Eisen, G. M. (2004). A novel diagnostic tool for detecting oesophageal pathology: the PillCam oesophageal video capsule. *Alimentary Pharmacology & Therapeutics*, 20:1083–1089.

Emura, F., Saito, Y., and Ikematsu, H. (2008). Narrow-band imaging optical chromocolonoscopy: advantages and limitations. *World Journal of Gastroenterology*, 14(31):4867–4872.

Everitt, B. (1977). *The Analysis of Contingency Tables*. Chapman and Hall.

Fireman, Z. and Kopelman, Y. (2007). The colon–the latest terrain for capsule endoscopy. *Digestive and Liver Disease*, 39(10):895–899.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Morgan Kaufmann, 2nd edition.

Häfner, M., Brunauer, L., Payer, H., Resch, R., Gangl, A., Uhl, A., Vécsei, A., and Wrba, F. (2010a). Computer-aided classification of zoom-endoscopical images using fourier filters. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):958–970.

Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vécsei, A., and Wrba, F. (2010b). Classification of endoscopic images using Delaunay triangulation-based edge features. In *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR'10)*, volume 6112 of *Springer LNCS*, pages 131–140, Povoa de Varzim, Portugal.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hurlstone, D. P., Cross, S. S., Adam, I., Shorthouse, A. J., Brown, S., Sanders, D. S., and Lobo, A. J. (2004). Efficacy of high magnification chromoscopic colonoscopy for the diagnosis of neoplasia in flat and depressed lesions of the colorectum: a prospective analysis. *Gut*, 53(2):284–290.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:153–158.

Kelsey, P. (2005). Colon - Normal Colon. The DAVE Project. Available at `http://daveproject.org/viewfilms.cfm?film_id=300`.

Kiesslich, R. (2007). Colon - Endomicroscopic Imaging of NSAID Associated Colitis. The DAVE Project. Available at `http://daveproject.org/viewfilms.cfm?film_id=561`.

Kiesslich, R. and Neurath, M. F. (2007). Endomicroscopy is born – do we still need the pathologist? *Gastrointestinal Endoscopy*, 66(1):150–153.

Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23.

Münzenmayer, C., Kage, A., Wittenberg, T., and Mühldorfer, S. (2009). Computer-assisted diagnosis for precancerous lesions in the esophagus. *Methods of Information in Medicine*, 48:324–330.

Swain, P. (2003). Wireless capsule endoscopy. *Gut*, 52(4):48–50.

Tischendorf, J. J. W., Gross, S., Winograd, R., Hecker, H., Auer, R., Behrens, A., Trautwein, C., Aach, T., and Stehle, T. (2010). Computer-aided classification of colorectal polyps based on vascular patterns: a pilot study. *Endoscopy*, 42(3):203–207.