**In-vivo multispectral video endoscopy towards in-vivo hyperspectral video endoscopy**

*Martin Hohmann*[^*], 1, 2, 3, R. Kanawade1, 2, F. Klämpfl1, A. Douplik2, 3, J. Mudter4, 5, M. F. Neurath2, 4, and H. Albrecht4

1 Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Institute of Photonic Technologies (LPT), Konrad-Zuse-Straße 3/5, 91052 Erlangen, Germany
2 Erlangen Graduate School in Advanced Optical Technologies (SAOT), Paul-Gordon-Straße 6, 91052 Erlangen, Germany
3 Ryerson University, Department of Physics, 350 Victoria Street Toronto, Ontario, MSB 2K3, USA
4 Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department of Internal Medicine 1, Ulmenweg 18, 91054 Erlangen, Germany
5 Sana Clinic Ostholstein, Department of Gastroenterology, Hospitalstraße 22, 23701 Eutin, Germany

Received 21 January 2016, revised 2 June 2016, accepted 3 June 2016
Published online 13 July 2016

**Key words:** endoscopy, hyperspectral, SVM, AdaBoost, stomach

---

1. **Introduction**

Gastric cancer is the second most frequent cause of cancer related death worldwide [1]. Despite the common usage of esophagogastroduodenoscopy (EGD) for screening programs, it is expected that cancer remains undetected in 20% of the patients [2]. It is even more complicated to find high-grade...
dysplasia (HGD) in Barrett’s Esophagus (BE). In these cases high definition white light endoscopy (HD-WLE) shows a sensitivity of 79%–85% [3, 4].

To achieve higher sensitivity and specificity more sophisticated methods like Narrow Band Imaging (NBI) or Autofluorescence Imaging (AFI) have been developed. However, in a systematic review Curvers et al. [5] could show only a small potential of NBI for detection of HGD. Despite promising early studies, later studies could not yield significantly better results than HD-WLE. Also another review from Subramanian and Raganath [6] reports that there is little evidence for the usage of NBI in the daily clinical routine. AFI-studies provide in general a good sensitivity of 90% but the false positive rate is up to 81% [7]. Hence, it is not suited for clinical usage for detection of malignant BE. Very often the results of both methods vary a lot between different endoscopists and the new imaging methods normally need special training and/or experience for the chosen method.

Newer meta analyses [8] show a strong tendency in favor of chromoendoscopy or virtual chromoendoscopy rather than random biopsies for the detection of high-grade dysplasia in BE. Both methods either dye the mucosa or virtually change the image. Instead of using dyes or spectral estimation techniques [9], we propose multispectral imaging (MSI) as first step towards hyperspectral video endoscopy (HSVE). Therefore, in our study, we used a flexible endoscope and automatic classification of healthy and cancerous areas which allows the usage of the real spectrum instead of an estimation of the spectral back reflection.

One must also recognize that targeted biopsies performed after a thorough inspection with white light endoscopy contribute most of the diagnosis. Therefore, a red flag technology is needed which is made possible by hyper spectral imaging (HSI). Due to the high spectral resolution, it allows a better contrast and due to this the result should be better than with white light imaging. As Swager et al. [10] state in their review: spectroscopic quantitative measurements of tissue need further investigation to demonstrate that it facilitates direct optical diagnosis of early neoplasia or risk stratification based on the presence of field carcinogenesis. It helps to gain more information about the tissue samples which should result in better diagnostic performance compared to current techniques. Moreover, due the possibility of using spatial information of the sample, it is possible to investigate the spatial changes of different wavelengths. These variables can be used for the classification to achieve a higher accuracy than by the sole usage of point spectroscopy.

HSI is an emerging field. It combines machine vision with spectroscopy [11, 12]. It enables the acquisition of two-dimensional images with the spectral information for each pixel. Considerable progress has been achieved in many different areas [13] and in the last years HSI became more and more popular in different fields of diagnostic application [14]. HSI is also used for the detection of cancer as one of the main applications of HSI for medical purposes [14]. This method has proven to be successful for the finding of cancer in many different parts of the human body: e.g. cervix [15], breast [16, 17], colon [18–22] or esophagus [23, 24]. Also new approaches spectrum [16, 24].

In hollow organs like colon or esophagus the use of HSI allows only studies performed on ex vivo tissue [18–21, 23, 24] or histological samples [22]. To our knowledge, no in vivo clinical studies on the esophagus of humans could be done despite different approaches [11, 23, 25] with flexible endoscopes. This is caused by the lack of flexibility and the rigid design of the used endoscope. In order to overcome this problem, we modified only the light source of a certified flexible endoscope from Olympus with the goal of providing us with a simple, stable and durable setup which can be used for a long time in clinical routine without the need of calibration. In general, however, alterations of the endoscope have to be limited because a certification for its usage in patients is difficult to obtain if parts are altered which are in direct contact with the patient.

In this study, we show first results of an endoscope, capable of fulfilling the aforementioned requirements for clinical usage. We are presenting the automated finding of adeno carcinomas in the stomach. This focus is chosen due the fact that more patients are available for adeno carcinomas and the detection is more simple. Therefore, the effect of mucus, inflammation and the difficulty of generating precise margins of the carcinomas can be studied more precisely. Different classifiers are used due to the fact that for supervised learning a precise margin for the training data is beneficial. But especially for early cancer of the stomach or esophagus the finding of the exact tumor-margin often remains difficult [26]. For this reason Support Vector Machine (SVM) [27] with linear and Gaussian Kernel, AdaBoost (AB) [28], RobustBoost (RB) [29] and Random Forest-walk (RFW) [30] are used to test if these methods can handle mislabeling when applied to hyper or multi spectral imaging.

2. Material and methods

2.1 Patients

A total number of 14 patients are investigated by the Department of Internal Medicine 1, Friedrich-
Alexander-Universität Erlangen-Nürnberg. All these 14 patients had histopathologically confirmed carcinoma in the stomach. Eight of these patients had previously undergone therapy, mostly chemotherapy. In this study, the patients with and without chemotherapy are mixed because the number of patients at the clinic is low. The age of the patients ranges between 50 and 85 years. Eleven patients are male and three patients are female. During the endoscopic procedure the patients are kept under analgosedation to minimize anxiety and discomfort. The HSVE procedure is performed before the regular endoscopic procedure with HD-WLE because the diagnosis was known before. From each patient multiple images of the suspected area are taken at different angles and distances to get as much training data as possible for further imaging. This is done in order to ensure that most of the imaging situations occurring during a regular endoscopy session are covered. Additionally, four to six biopsies are taken from each lesion. The biopsies are histologically analyzed. The patients received complete information about this study before providing their consent for the participation and the research is carried out in accordance with the Declaration of Helsinki. The study has been approved by the Institutional Review Board (IRB) of the Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

2.2 MSI setup

The setup is a modified version of an earlier presented setup [23]. The setup has to be robust and stable because it is used in a daily routine at the endoscopy unit. For this reason a very simple setup design is chosen which can withstand bumps or other shocks happening during the routine work in the clinics. Furthermore, parts which may come into contact with a patient, have to be certified. These factors limit the possibility of modifications to the endoscope.

Therefore, the multispectral endoscopy setup is a standard endoscopy system, consisting of an Olympus endoscope GIF 100 (Olympus Corporation, Tokyo, Japan), an Olympus video processor CV-140 (Olympus Corporation, Tokyo, Japan), a modified light source Olympus CLV-U40 (Olympus Corporation, Tokyo, Japan) and an external light source (Lumencor spectra 7-LCR-XA, Beaverton, OR, USA). The schematic is shown in Figure 1(A). The external light source is coupled into the light source unit CLV-U40. The video signal from the endoscope is used as a trigger for the external light source. The system is controlled by a personal computer (PC). The data input, data output, graphical user interface (GUI) and the control of the external light source is done by a Matlab (The MathWorks, Inc., Natick, MA, USA) program. Figure 1(B) presents the setup and how it is used in the clinical environment.

The modified light source unit is capable of using six wavelength bands. They reach from 400 nm to 650 nm (Table 1). First, white light imaging is performed. To do so, the six wavelength bands are used at the same time. The multispectral imaging mode is implemented as a spectral scanning system. It allows taking a single multispectral image and afterwards continues imaging under normal white light illumination. The number of multispectral images taken varies, depending on how many motion artefacts occurred and how many different angles are possible. The reso-

---

**Figure 1** (A) Schematic setup of the endoscopic imaging system consisting of a light source unit, an endoscope image processor, an external light source and a PC. (B) Photo of the setup: On the upper rack under the screen the video processor can be seen. The modified light source unit is placed below. The light source unit is increased in its size to allow major modifications of the optical path. On the lower rack, the controlling PC and the endoscope (black box) are placed.

**Table 1** Centre wavelength and the half width full maximum of the used wavelength bands.

<table>
<thead>
<tr>
<th>Wavelength in nm</th>
<th>438 ± 12</th>
<th>475 ± 17</th>
<th>512 ± 12</th>
<th>542 ± 13</th>
<th>575 ± 12</th>
<th>628 ± 20</th>
</tr>
</thead>
</table>

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
lution is approximately 350 \times 370 pixels and it is lim-
ited due to the normal resolution of the Olympus en-
doscope. The multispectral images are taken in a se-
quence. The whole imaging process takes 2 seconds (s)
including pre- and post-processing. Therefrom, 0.45 s
are needed to acquire the full multispectral image.
This time is measured during the imaging process with
the commands “tic” and “toc” from Matlab.

The field of view is 120 degrees and the image is
sharp with a distance of more than 3 mm between
the sample and the tip of the endoscope. The endo-
scope has a fixed focus with a depth of focus from
3 mm to 100 mm. In realistic images the stomach has
a distance from 0.5 cm to 5 cm. Thus, the pixel size
at the image plane varies from 0.05 mm to 0.5 mm.
The main limiting factor is the SNR. The noise level
is around ± 5 intensity values from 256 possible in-
tensity values and it is constant independent of the
signal intensity. Thus, in the optimal case with per-
fected illumination, the SNR is 20. For realistic cases, it
is around 10 but it can also be lower. However, the
signal which is measured is a modulation on top of
the intensity values. Hence, effectively the SNR is
even lower. Another distortion is the barrel distor-
tion which is typical to endoscopes. Spectral distor-
tions like keystone and smile are minor. There is
a distance from 0.5 cm to 5 cm. Thus, the pixel size
at the image plane varies from 0.05 mm to 0.5 mm.
The main limiting factor is the SNR. The noise level
is around ± 5 intensity values from 256 possible in-
tensity values and it is constant independent of the
signal intensity. Thus, in the optimal case with per-
fected illumination, the SNR is 20. For realistic cases, it
is around 10 but it can also be lower. However, the
signal which is measured is a modulation on top of
the intensity values. Hence, effectively the SNR is
even lower. Another distortion is the barrel distor-
tion which is typical to endoscopes. Spectral distor-
tions like keystone and smile are minor. There is
one pixel wide shift of the image for blue and red.

Two methods allow us to test and prove the sta-
bility of the system. First, before the measurement of
each patient, the reflection is calibrated against a re-
flector made of barium sulphate. Therefore, the tip of
the endoscope is placed in an endoscope calibration
chamber coated with barium sulfate (Fraunhofer).
Afterwards, the light of one wavelength band is
switched on with low intensity and gradually increased
until the endoscope measures the threshold intensity.
The input parameter for the light source is stored and
the process is repeated for the other wavelength
bands. Finally, measured intensity at the threshold
value is used as calibration value. Second, after its
usage for half a year in the clinical environment, only
minor alterations of the optical path were found.

2.3 Pre-processing

Before data analysis, the tumor margin in each im-

age has to be found. Before delineating this margin
the histopathological diagnosis was again confirmed.
The margin was drawn by a medical expert accord-
ing to endoscopic differentiation criteria between

dysplasia and normal tissue. However, a certain
amount of the tissue is expected to be labeled
wrongly due to the fact that even a medical expert
cannot be absolutely perfect at finding the margin
[26]. Also in in-vivo situations biopsies cannot be ta-
ken at every point for compensation.

Before the analysis, images with motion artefacts
are excluded from the analysis. Also areas of images
with over-saturation (e.g. due to specular reflection)
are excluded from the analysis. Additionally, a mar-

gin of 12 pixels is also excluded to minimize the ef-

eect of partly specular reflective surfaces. The rest of
the data is smoothed with 7 \times 7 pixel wide Gaussian
smoothing. A second set of data is created by calcu-

lating the derivative image. From both images com-
bined the coefficient matrix of the Principal Compo-

nent Analysis (PCA) [31] is calculated from the car-

cinoma data and applied to the test data. By choos-
ing to calculate the PCA only on one part of the
data the detectability of this part is increased [32].
Hence, the detection rate of malignant tissue should
be high in this study. The carcinoma tissue is chosen
because in most cases HSI has a high specificity and
a low sensitivity for gastric and esophageal carcino-

mas [24]. Therefore, the sensitivity is increased by
performing the PCA only on the tumor data. After-

wards, the first principal components accounting for
99% of the variations are used for the classification.
In this study the PCA is used as a further noise re-
duction tool. For all patients except patient 5, this
leads to a usage of nine PCA components. For pa-

tient five ten components are used.

2.4 Data analysis

The data analysis is done for each pixel separately.
For classification an SVM is used with a Gaussian
Kernel and another one with a linear kernel. SVM is
chosen because it proved to be a good classifier for
HSI classification [33–35]. Moreover, AB is used
with a tree learner because normally it shows similar
results as SVM. Additionally, RB is used due to the
fact of being more robust against mislabeled training
data [29, 36]. Hence, it is well suited for the classifi-
cation problem in this study. RFW is used because it
is an advanced version of AB which is also robust
against labelling noise [30]. Furthermore, in some
cases RFW shows when compared to SVM, superior
classification performance for cancer classification
[37] or for the analysis of hyper spectral images [38].

For the training of the classifiers, in the beginning
1% of each patient’s data is chosen randomly for the
training step to reduce the calculation time. Thus, all
classifiers get the same training data. The test data is
generated with a leave-one-out strategy. Therefore,
13 patients are used as training data set and one as
test data set. This results in 70'653 to 82'449 training
data points depending on which patient is left out.
At the same time the left out patient is completely
used for testing except the oversaturated area and a
small margin around the carcinoma boundary. This
is left out due to the fact that the clear boundary of
the carcinoma is not known. Thus, 200'474 to 1'380'109 data points are used for testing.

AB and RB are used with a decision stump tree classifier. The maximum number of splits is set to one, the minimal parent size (minimal amount of observations for nodes) is set to two and the minimal leaf size (minimal size of observations for leaves) is set to one. AB stops after training of 500 weak learners and RB latest after 300 weak learners. For RFW a deep tree learner is used. The maximum number of splits is set to the amount of predictors minus one, the minimal parent size is set to two and the minimal leaf size is set to one. RFW stops after training of 500 weak learners. For both cases of the SVM the settings are on auto and the classifier centers and scales each column of the predictor.

Due to the fact that a certain amount of the data is mislabeled an additional step is introduced to find outliers. For the SVM with a linear kernel and the SVM with the Gaussian kernel a second step is introduced which selects 5% of the training data as outlier. These outliers are discarded for the training process. Before the final analysis outlier rates within the range of 0% and 8% are tested. In contrast to SVM, for RB an expected outlier rate has to be selected. RB is used for an outlier rate of 16%. Hence, in further discussion they will not be separated. At the same time it should be noted that AB and RFW provide nearly identical results (Figure 2).

The sensitivity and specificity are analyzed by comparing the found labeling against expert labeling of a known carcinoma tested with biopsies. The comparison is done for every pixel. Thus, the sensitivity and specificity are influenced by two factors: Whether the carcinoma was found, and if so, whether it was found with exact delineation. Additionally, the accuracy is calculated in two ways:

\[ \text{1. accuracy} = \frac{TN + TP}{TN + TP + FN + FP} \]

\[ \text{2. accuracy}_2 = \frac{\text{Sensitivity} + \text{Specificity}}{2}, \]

where TN is true negative, TP is true positive, FN is false negative and FP is false positive. The sensitivity and specificity are calculated as follows:

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \quad \text{and} \quad \text{Specificity} = \frac{TN}{TN + FP} \]

The first accuracy is chosen because it is the standard way of calculating the accuracy. The second method is used because it has proved to be more robust whenever the number of elements of one class is much higher than the other one. This happens with some of the patients, and in this case the accuracy would be dominated by the class which occurs more often. For the value of accuracy2 there is also the standard deviation calculated to show the variation of the results. This step is introduced to ensure that an algorithm found provides stable results for many patients. Thus, the standard deviation should be low. In all cases for all classifiers the operation point was chosen only based on the training data without any alteration. There was no optimization done for choosing the best operation point.

### 3. Results and discussion

#### 3.1 Results

The results are presented in in Table 2. The accuracy varies from 0.55 for RFW to 0.64 for RB and from 0.59 to 0.64 for accuracy2. For all methods except SVM with Gaussian kernel the increase of accuracy is caused by a rise of sensitivity while specificity only varies a little bit.

For nearly all patients RB provides the best results. RB is the best classification algorithm for nine out of 14 patients (Figure 2). For presenting the results of each patient, accuracy2 is chosen, due to the fact that in some patients the amount of carcinoma or healthy tissue proved to be much bigger than that in the remaining patients. Therefore, when using the standard accuracy, the results would not be comparable. At the same time it should be noted that AB and RFW provide nearly identical results (Figure 2). Hence, in further discussion they will not be separated any more.

Even though SVM with a linear kernel achieves the second best results, it has, on the other hand, a

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Accuracy 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.56</td>
<td>0.62</td>
<td>0.55</td>
<td>0.59 ± 0.08</td>
</tr>
<tr>
<td>Forest Walk</td>
<td>0.57</td>
<td>0.62</td>
<td>0.61</td>
<td>0.60 ± 0.11</td>
</tr>
<tr>
<td>RobustBoost</td>
<td>0.66</td>
<td>0.52</td>
<td>0.56</td>
<td>0.59 ± 0.09</td>
</tr>
<tr>
<td>SVM linear kernel</td>
<td>0.66</td>
<td>0.52</td>
<td>0.56</td>
<td>0.59 ± 0.09</td>
</tr>
<tr>
<td>SVM Gaussian kernel</td>
<td>0.56</td>
<td>0.63</td>
<td>0.55</td>
<td>0.59 ± 0.08</td>
</tr>
</tbody>
</table>

Table 2 Classification results of the five tested classifiers for 14 patients with leave one out strategy. The accuracy2 is presented with its standard deviation.
higher standard deviation. Therefore, it is not suited for later usage, despite achieving better results than RB for four out of the 14 patients. The comparison of the results can be done because the deviation of the mean is a factor \( \sqrt{14} \approx 3.7 \) smaller than the standard deviation presented in Table 2.

SVM with Gaussian kernel, AB and RFW always provide worse results than RB except for patient six. The reason for this finding might be that SVM with Gaussian kernel, AB and RFW have problems with mislabeling which is introduced by drawing the margin by the clinical expert. A hint that mislabeling might in this case influence the result is found if the mean of 100 subsequent data points is taken for classification instead of choosing 1% of the data randomly. Then the best classifier is SVM with a linear-kernel with an accuracy2 of 62%. The accuracy2 of RB drops in this case to 59%. Thus, RB seems to at least partly compensate the effect of mislabeling and therefore provides the best classification results. Due to the superiority of RB, all further consideration will be carried out with RB.

An overlay between the margin from the medical expert and the margin from the classification for a few typical cases is shown in Figures 3–5. The examples represent a bad (Figure 3), a medium (Figure 4) and a good (Figure 5) classification result. To minimize potential selection bias all images of the test-data are shown. The area classified as carcinoma is presented in green and the margin from the medical expert is shown in red. Moreover, the areas excluded from the analysis due to oversaturation are shown in a less bright color if they are classified as carcinoma or in purple in the other case.

For the bad result (Figure 3) the carcinoma is not found in most cases. Sometimes it also happens that nearly everything is classified as cancerous. For the medium good classification results (Figure 4) the carcinoma is found. However, the margin is not found correctly. In some cases this might lead to over- or underestimation of the cancerous region. For the good example the classification nearly matches perfectly to the margin found by the medical expert.

Additionally, the ROC curves are shown in Figure 6. It can be seen that they are not completely symmetric for most of the patients. The drop of the false positive rate for increasing true positive rate is lower than the other way around. E.g. for a false positive rate of 0.6 the true positive rate is higher than 0.8 for eight patients while for a true positive rate of 0.4 the false positive rate is smaller than 0.2 for only 6 patients.

Furthermore, it should be examined why certain patients are not correctly classified. Hence, the results are checked with regard to pretreatment, grade of inflammation and the amount of mucus. In general, the presence of mucus reduces the accuracy because the mucus masks the tissue. Table 3 compares the results obtained for patients with a different level of mucus, inflammation and treatment. It clearly
shows that the best results are obtained for patients without or with only minor inflammation, whereas two of the three patients with the worst results have increased amounts of mucus or strong inflammation. For the other results there seems to be no clear dependency. In order to figure out the reason behind it, the accuracy was calculated depending on the amount of mucus, the grade of the inflammation and whether or not the patient has already undergone chemotherapy prior to MSI measurement. Table 4 shows the accuracy value obtained for these cases.

Table 4 exhibits that there seems to be no difference in the detection of carcinoma if the patient has undergone chemotherapy prior to MSI measurement. For a more detailed consideration, this effect is also checked for all those patients who have no or low presence of mucus or inflammation because it is expected that a high amount of mucus or a stronger inflammation might overlay with the effect of the chemotherapy. Table 4 shows that in this case the difference between patients with and without chemotherapy gets bigger. Hence, it might pinpoint that chemotherapy makes the detection of carcinoma more difficult. With the number of patients presented in this study, the effect is too low to make a statistically significant prediction.

While a low amount of mucus does not seem to reduce the accuracy, at the same time a great amount of mucus strongly reduces the accuracy. The patient with medium amount of mucus seems to be an outlier. As far as inflammation is concerned, the effect is similar. In the case of absence of inflammation, low and medium inflammation, there seems to be no difference in the detection abilities of our MSI

**Figure 4** Example of a patient with medium accuracy (patient 5). Overlay of the margin drawn by the medical expert (red) with the classification results (green). The purple area shows the part which is excluded due to oversaturation. Also the less bright green part is excluded due to oversaturation.

**Figure 5** Example of a patient with high accuracy (patient 4). Overlay of the margin drawn by the medical expert (red) with the classification results (green). The purple area shows the part which is excluded due to oversaturation. Also the less bright green part is excluded due to oversaturation.
system, while severe inflammation reduces them strongly. Chemotherapy might also have some effect on data accuracy. The reason for this observed effect might be that the treatment partly reduced the carcinoma. Therefore, the visibility drops and it becomes harder to detect. The reason for the impact of mucus on the measurement is best explained by the coverage of the area under investigation. The white opaque mucus just blocks the light penetrating into deeper clinically relevant tissue. It is also known that a mucus cap can pinpoint towards dysplasia, especially in colonic lesions [39]. Hence, except the values of medium amount of mucus, the results are within the expectations. For inflammation there are two explanations. First, the inflammation disguises the malignant alterations. Second, there is only one patient found with strong inflammation. Therefore, there is no similar patient as training data, if this patient is tested as a leave-one-out sample, and the result might be an outlier.

3.2 Discussion

To our knowledge, there is no in-vivo study for hyper spectral video endoscopy in the upper GI published so far. To be able to perform research on esophagus or colon other groups have used histological or ex-vivo samples [18–24]. The results in our study brings HSI one step closer towards the clinical environment. So far, hyper spectral studies of the upper gastrointestinal (GI) tract are only done by Kiyotoki et al. [24, 40]. They reached a final sensitivity of 78.8% and a final specificity of 92.5% for ex-vivo investigations. Our study

Table 3 Comparison of the results of all patients with the amount of mucus, the grade of the inflammation and if chemotherapy was done before.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Chemotherapy</th>
<th>Mucus</th>
<th>Inflammation</th>
<th>Classification result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Medium</td>
<td>Weak</td>
<td>++</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>No mucus</td>
<td>No inflammation</td>
<td>o</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>A lot</td>
<td>Medium</td>
<td>- -</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>No mucus</td>
<td>No inflammation</td>
<td>++</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Some</td>
<td>Weak</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>No mucus</td>
<td>Strong</td>
<td>- -</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Some</td>
<td>No inflammation</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>No mucus</td>
<td>Some</td>
<td>o</td>
</tr>
<tr>
<td>9</td>
<td>Yes</td>
<td>No mucus</td>
<td>No inflammation</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>Yes</td>
<td>No mucus</td>
<td>Medium</td>
<td>+</td>
</tr>
<tr>
<td>11</td>
<td>No</td>
<td>Some</td>
<td>No inflammation</td>
<td>- -</td>
</tr>
<tr>
<td>12</td>
<td>No</td>
<td>Some</td>
<td>Medium</td>
<td>+</td>
</tr>
<tr>
<td>13</td>
<td>Yes</td>
<td>Some</td>
<td>No inflammation</td>
<td>o</td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>Some</td>
<td>Weak</td>
<td>++</td>
</tr>
</tbody>
</table>

Table 4 Accuracy2 for different strength of inflammation, amount of mucus and the presence of chemotherapy for RobustBoost with the standard deviation. In the brackets the amount of patients is shown. The third row (Chemotherapy limited patients) shows the influence of the chemotherapy for patients with no or weak inflammation and no or some mucus.

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Some/Weak</th>
<th>Medium</th>
<th>A lot/strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemotherapy all patients</td>
<td>0.66 ± 0.11 (8)</td>
<td>0.63 ± 0.07 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemotherapy limited patients</td>
<td>0.68 ± 0.12 (4)</td>
<td>0.63 ± 0.04 (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mucus</td>
<td>0.64 ± 0.10 (6)</td>
<td>0.65 ± 0.06 (6)</td>
<td>0.72 (1)</td>
<td>0.51 (1)</td>
</tr>
<tr>
<td>Inflammation</td>
<td>0.65 ± 0.10 (6)</td>
<td>0.68 ± 0.06 (4)</td>
<td>0.62 ± 0.10 (3)</td>
<td>0.53 (1)</td>
</tr>
</tbody>
</table>
showed a sensitivity of 63% and specificity of 64% for the best chosen classifier. This is a large difference when comparing both studies. However, there are also five major differences of our experiment and data evaluation compared to the one from Kiyotoki et al. [24, 40]. The differences and the expected strength of the influences are summarized in Table 5:

1. In the study of Kiyotoki et al. [24] the data of the carcinoma is normalized by dividing it by the mean of the values of five measurements of healthy tissue of each patient. This step leads to two major conclusions: above all, this step needs a priori knowledge. It is assumed that a healthy area of the patient is known. This step would not be suited for using HSI as a red flag technology. Furthermore, this step almost completely omits the inter-patient variations of the data analysis. By dividing all values by the mean of five healthy data points a normalization step is introduced which compensates variations of the patients (inter-patient variation). Therefore, the classifier in the study of Kiyotoki et al. [24] mainly has to compensate the intra-patient variations and not the inter-patient variations. Due to the large differences between patients it is expected that this point has the biggest impact on the difference of the results. If the analysis would be done in this study only with intra-patient variations, the accuracy would be also between 80% and 95%.

2. Due to the usage of ex-vivo samples in the study of Kiyotoki et al. [24] a homogeneous illumination of the sample is possible which simplifies the classification. For in-vivo studies, homogeneous illumination of the esophagus is not possible due to the hollow tube like structure of the esophagus. Areas which are further away from the tip of the endoscope normally appear darker than the ones which are closer. Nevertheless, this point is expected to have only a weak influence because the classification can be done with relative spectra. This means that one wavelength is used as reference wavelength and all others are divided by this wavelength. In this case, parts of the image with a higher distance to the tip of the endoscope have a lower overall signal and due to this the noise is more dominating.

3. Kiyotoki et al. [24] used ex-vivo tissue. Even if the tissue was used right after the resection, it was washed with saline and might have changed in the meantime the spectral reflection of the tissue. However, due to the usage of very fresh tissue by Kiyotoki et al. it is expected that this effect is not of great importance except for the removal of blood.

4. Currently, in this study only a multispectral device with six wavelength bands is used in comparison to the real HSI device with around 70 wavelength bands used by Kiyotoki et al. [24]. The finer wavelength resolution might increase the contrast slightly [41] which simplifies the classification. Thus, also this point should have only a slight impact on the difference. Moreover, this reduction might even be eliminated by the Wiener estimation [42] or similar spectral reconstruction techniques.

5. From the study of Kiyotoki et al. [40] the near infra-red (IR) wavelength band seems to be best suited for carcinoma detection. In this study, this part of the spectrum cannot be used due to the limitations of the Olympus endoscope which does not support IR imaging without modifications. This would prevent the in-vivo usage. The impact of this point seems to be significant because in the study of Kiyotoki et al. [24] the accuracy rises from below 65% for wavelength below 600 nm to around 75% for wavelengths between 650 to 800 nm. Moreover, the accuracy below found 65% for visible light is similar to the accuracy of 64% reached in this study.

Despite using a multispectral imaging device, the results and the statistical methods should be transferable to hyperspectral endoscopy. First, real hyper spectral endoscopy with a finer wavelength resolution should lead to an increase of the contrast according to Wang et al. [41]. Furthermore, studies comparing the output of RGB and multispectral images against hyperspectral images normally favor hyperspectral imaging [43, 44]. Nevertheless, for both cases the same methodologies can be applied. The second reason is that all methods used in this

---

**Table 5** Differences of the system and the analysis between the studies from Kiyotoki et al. [24, 40] and this study and its estimated influence on the results.

<table>
<thead>
<tr>
<th>Difference this study, Kiyotoki et al. [24, 40]</th>
<th>Kiyotoki et al.</th>
<th>this study</th>
<th>Estimated influence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inter patient variations</td>
<td>No</td>
<td>Yes</td>
<td>Very large</td>
</tr>
<tr>
<td>2. Homogeneous illumination</td>
<td>Yes</td>
<td>No</td>
<td>Medium</td>
</tr>
<tr>
<td>3. In-vivo experiments</td>
<td>Yes</td>
<td>No</td>
<td>Medium</td>
</tr>
<tr>
<td>4. Fine wavelength resolution</td>
<td>Yes</td>
<td>No</td>
<td>Medium</td>
</tr>
<tr>
<td>5. Usage of Visible and IR light</td>
<td>Yes</td>
<td>No</td>
<td>Large</td>
</tr>
</tbody>
</table>

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
study are normally working well with hyperspectral data. Third, with a closer comparison to the study from Kyotoki et al. [24], it can be seen that in their study, the accuracy for the common wavelengths is not more than 65% (compare Figure 5 in their study). This is a similar accuracy to the one found in this study. Therefore, it highlights that in this study there might not be a major drawback due to the usage of multispectral imaging. Finally, the problem of the mislabeled training data set will also be present for the case of hyperspectral endoscopy. Thus, the results from this study should be transferable to hyperspectral endoscopy.

4. Conclusion

In conclusion, this study shows problems which will be faced for applying HSI in real in-vivo situations, especially for applications as red flag technology. First, it seems to be difficult to correctly identify and mark the margin of the carcinoma. This issue mainly arises when the boundaries for classifiers for an in-vivo situation have to be available because a biopsy cannot be taken at every point. Nevertheless, RB seems to partly compensate this problem and provides better classification results than SVM, AB and RFW. Second, the inter-patient variations seem to strongly reduce the accuracy at the current state of the technology. Nevertheless, it should be possible to transfer the results to hyperspectral imaging. In the future, a combination of RB, infra-red imaging, more patients for a more precise statistical evaluation and finer spectral resolution, or alternatively Wiener estimation of the spectra might help to solve these issues.

Acknowledgements The authors gratefully acknowledge the funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the Deutsche Forschungsgemeinschaft (German Research Foundation – DFG) within the framework of the Initiative for Excellence.

The research was also founded by the ELAN-Fond of the Medical Faculty of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (AZ12-08-14-1)

Additional support from Canadian NSERC Individual Discovery Grant, Ryerson Graduate Chang School and Ryerson Physics Department is greatly acknowledged.

References


