Optimized Selection of Weak Methods for the Classification of Endoscopic Images Using an Ensemble Classifier

M. Häfner*, S. Hegenbart†, M. Liedlgruber†, A. Uhl†, A. Vécsei‡ and F. Wrba§
*Department for Internal Medicine, St. Elisabeth Hospital, Vienna, Austria
†Multimedia Signal Processing and Security Lab (WaveLab),
Department of Computer Sciences, University of Salzburg, Austria
Email: {shegen, mliedl, uhl}@cosy.sbg.ac.at
‡Endoscopy Unit, St. Anna Children's Hospital, Vienna, Austria
§Department of Clinical Pathology, Medical University of Vienna, Austria

***Abstract*—In the past we developed an ensemble classifier in order to improve the accuracy in terms of the classification of endoscopic images. However, since we have a variety of feature extraction methods for the construction of a weak method set at hand, the number of different possible weak method combinations for the ensemble is quite huge.**

In order to address this issue we propose two different methods which aim at determining a set of weak methods which delivers an optimal overall classification rate. While the first algorithm determines optimal weak method candidates by a rating based on the candidate set, the second algorithm aims at constructing a set of statistically significant methods in order to increase the diversity of the ensemble.

Based on previously developed methods, we evaluate the proposed methods by comparing the overall rates achieved by the respective combinations to the overall rate achieved by the best possible combination. We show that the proposed selection algorithms are able to find the best combination of methods or at a least competitive one.

I. INTRODUCTION

Today the gold standard for colon examination is colonoscopy which is performed by using a colonoscope. Modern colonoscopes are able to take pictures from inside the colon which allows to obtain images for a computer-assisted analysis with the goal of detecting tumorous lesions. To get highly detailed images a magnifying endoscope is used [1]. Such an endoscope represents a major advance in colonoscopy as it provides images which are up to 150-fold magnified, thus uncovering the fine surface structure of the mucosa as well as small lesions.

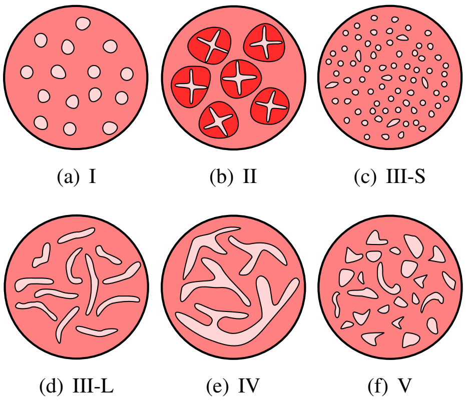
In the past we already developed a variety of different methods, which yield different kinds of features. However, some of these methods delivered rather unstable classification results across the different image classes. A common way to improve the accuracy of a medical image classification system in such a case is to employ an ensemble classifier. In the past we therefore developed an ensemble classifier to overcome this limitation [2], [3]. While we were able to show that a classifier ensemble is indeed able to yield more robust classification results, up to now choosing the optimal combination of weak methods has been done in a heuristic way (i.e. manually evaluating different combinations).

In this work we propose two different methods which aim at finding optimal combinations of weak methods to alleviate the need for a manual optimization.

The remaining part of this paper is organized as follows: in Section II we give a brief introduction to the medical background of this work. This is followed by a presentation of the proposed weak method selection algorithms in Section III. In Section IV we explain the experimental setup used and discuss the results obtained. Section V concludes this paper.

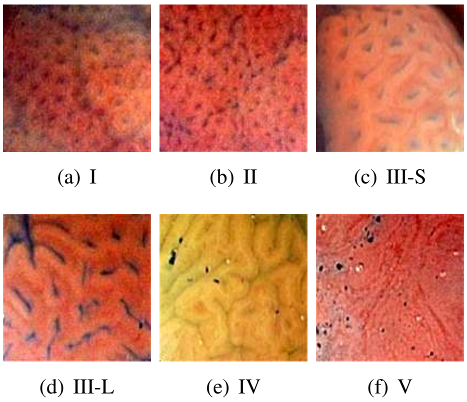
II. MEDICAL BACKGROUND

Due to the fact that colonic polyps have a rather high prevalence and are known to either develop into cancer or to be precursors of colon cancer, an early detection of such pathologies can lower the mortality rate drastically. Hence, automated classification systems targeted at the assessment of the malignant potential of colonic polyps aim at avoiding random and, probably, unnecessary biopsies. As a consequence such systems could potentially help to save time, lower the


(a) I (b) II (c) III-S

(d) III-L (e) IV (f) V

Fig. 1. A schematic illustration of the different pit pattern types according to Kudo et al.



(a) I (b) II (c) III-S

(d) III-L (e) IV (f) V

Fig. 2. Example images for the different pit pattern types.

cost for colonoscopy procedures, and reduce the risk of complications during such procedures.

One classification scheme, commonly used to to distinguish between the different types of polyps, is the pit pattern classification scheme, originally reported by Kudo et al. [4]. Based on the visual pattern of the mucosal surface this system allows to differentiate between normal mucosa, hyperplastic lesions (non-neoplastic), adenomas (a pre-malignant condition), and malignant cancer. A schematic illustration of the pit pattern classification and example images are given in figures 1 and 2, respectively.

While the pit pattern classification distinguishes six different types of patterns, in this work we focus on a 3-class classification according to [5]. This classification groups the six different pit pattern types into normal lesions (pit pattern types I and II), non-invasive lesions (pit pattern types III-S, III-L, and IV), and invasive lesions (pit pattern type V). This classification scheme is of particular interest since normal mucosa needs not to be removed, non-invasive lesions must be removed endoscopically, and invasive lesions must not be removed endoscopically.

III. PROPOSED METHODS

Throughout the literature different strategies dealing with the optimal selection of weak methods for an ensemble classifier exist [6]. In order to improve classification accuracies by an ensemble, the methods combined should yield a high overall rate and have a high diversity. In the following we describe our proposed algorithms which aim at determining optimal combinations of weak methods out of all possible combinations, considering M, which denotes the set of weak methods available. Hence, in this work we are not aiming at the highest possible ensemble classification rate, but at finding the best performing weak method combination out of all possible combinations in a fast and efficient way.

A. Ensemble based on rating

This method is based on the computation of a rating value for a given candidate set of methods S. After initializing S to contain the best performing method $B \in M$ only, new methods are added to S in an iterative manner. If there exists a method which has not been tested yet for inclusion and which is able to improve the rating R of S, it is added to S.

The whole process of testing all methods not contained in S (denoted by T) and adding new methods is repeated N times, where N corresponds to the number of weak methods available in M.

Algorithm 1 Selection based on combination rating

```
1:  $M \leftarrow$ set of available methods
2:  $N \leftarrow |M|$
3:  $B \leftarrow$ best performing method in $M$
4:  $S \leftarrow \{B\}, R \leftarrow 0$
5:  for $iter = 1$ to $N$ do
6:      $T \leftarrow M \backslash S$
7:      for all $t \in T$ do
8:          $T \leftarrow T \backslash t$
9:          $S \leftarrow S \cup t$
10:         $R_{new} \leftarrow 0$
11:         for all $i \in I$ do
12:             $C \leftarrow 0$
13:             for all $s \in S$ do
14:                 if $i$ is correctly classified by $s$ then
15:                     $C \leftarrow C + 1$
16:                 end if
17:             end for
18:             if $C < |S|/2$ then
19:                 $R_{new} \leftarrow R_{new} - 1$
20:             else if $C > |S|/2$ then
21:                 $R_{new} \leftarrow R_{new} + 1$
22:             end if
23:         end for
24:         if $R_{new} < R$ then
25:             $S \leftarrow S \backslash t$
26:         else
27:             $R \leftarrow R_{new}$
28:         end if
29:     end for
30: end for
```

The rating of the candidate set S is based on the idea that more than half of the methods in the candidate set should classify a given image correctly. If this is the case, the newly computed rating R_{new} for S is increased by one. If exactly half of the methods classify a given image correctly, the rating remains untouched. Otherwise R_{new} is decreased by one. After testing this condition for each image in the image set used, the newly computed rating is adapted if R_{new} exceeds the previously computed rating for S. Otherwise the just added method is removed from S (see also Algorithm 1, lines 11–28).

After all iterations have been finished, the final set of methods is contained in S.

B. Ensemble based on statistical significance

Our second method does not rely on the computation of a ranking value. Instead this method iteratively adds new methods to S which produce a significantly different output as compared to the previously added method (denoted by C). In addition, only methods exceeding a certain level of classification accuracy are potentially added to S. The idea behind this is to increase the diversity of a weak method set

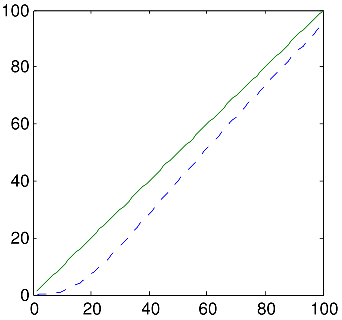


Fig. 3. Comparison of O_B (solid line) and the respective overall accuracy lower bound L for adding new methods (dashed line).

by combining methods which produce significantly different output but yet perform rather comparable in terms of the classification accuracies.

After initializing S to contain the best performing method $B \in M$ only, new methods are added to S in an iterative manner. First, D is initialized as the set of methods which produce a significantly different output as compared to C (see Algorithm 2, lines 14–19). If this set remains empty the algorithm is aborted since there are obviously no more methods which satisfy the requirement of being significantly different to C. Otherwise, the algorithm searches for the method $m \in D$ exhibiting the highest overall classification accuracy O_D. If O_D is higher or equal to a lower bound L, the respective method is added to S. Otherwise the algorithm is aborted since there is obviously no more method in D which is able to meet the lower bound requirement (see Algorithm 2, lines 23–26).

The lower bound L is introduced in order to avoid adding methods to S which might have a negative influence on the ensemble. Hence, we compute L based on the overall rate of B, denoted by O_B. The higher the value of O_B the smaller the tolerance for adding methods with respect to the overall accuracy should get, and vice versa. The value for L is obtained by a non-linear computation to have more control over the tolerance in different accuracy ranges. We therefore impose the constraint on L to be 1/1 (100%), 1/5 (20%), and 1/20 (5%) below the value of O_B for an overall accuracy of 1%, 50%, and 100%, respectively. In order to be able to compute L for all possible accuracy values, we fit the following polynomial of degree 3

$$G(x) = ax^3 + bx^2 + cx + d \tag{1}$$

to the denominators of the constraint values $\{1, 5, 20\}$ at locations $\{1, 50, 100\}$. The final value for L is obtained by taking the multiplicative inverse of $G(x)$ subtracted from one and multiplied by O_B, which can be expressed formally as

$$L = O_B \left(1 - \frac{1}{G(O_B)}\right). \tag{2}$$

A comparison between O_B and the respective overall accuracy lower bound for adding a new method is shown in Fig. 3.

Algorithm 2 Selection based on McNemar's test statistic

```
 1: $a \leftarrow 5.992158558 \cdot 10^{-6}$
 2: $b \leftarrow 1.300914834895 \cdot 10^{-3}$
 3: $c \leftarrow 0$
 4: $d \leftarrow 0.998693093006543$
 5: $G(x) := ax^3 + bx^2 + cx + d$
 6: $M \leftarrow$ set of available methods
 7: $B \leftarrow$ best performing method in $M$
 8: $O_B \leftarrow$ overall rate achieved by $B$
 9: $C \leftarrow B$
10: $S \leftarrow S \cup C$
11: $L \leftarrow O_B \left(1 - \frac{1}{G(O_B)}\right)$
12: $T \leftarrow M \backslash C$
13: while true do
14:     $D \leftarrow \{\}$
15:     for all $t \in T$ do
16:         if McNemar's test$(t, C) =$ true then
17:             $D \leftarrow D \cup t$
18:         end if
19:     end for
20:     if $D = \{\}$ then
21:         exit
22:     end if
23:     $O_D \leftarrow$ highest overall rate $\forall m \in D$
24:     if $O_D < L$ then
25:         exit
26:     end if
27:     $C \leftarrow$ first method yielding $O_D$
28:     $T \leftarrow T \backslash C$
29:     $S \leftarrow S \cup C$
30: end while
```

To assess whether the differences between the outcomes of two methods are statistically significant we employ McNemar's test due to its simplicity [7]. It must be pointed out however that alternative statistical tests may be used too. For two methods m_1 and m_2 this test statistic keeps track of the number of images which are misclassified by method m_1 but classified correctly by method m_2 (denoted by n_{01}) and vice versa (denoted by n_{10}). The test statistic, which is approximately Chi Square distributed (with one degree of freedom), is then computed as

$$Z = \frac{(|n_{01} - n_{10}| - 0.5)^2}{n_{01} + n_{10}}. \tag{3}$$

From T the p-value can be computed as

$$p = 1 - F_{\chi_1^2}(Z) \tag{4}$$

where $F_{\chi_1^2}$ denotes the cumulative distribution function of the Chi Square distribution with one degree of freedom. The null-hypothesis H_0 for McNemar's test is that the outcomes of m_1 and m_2 lead to equal error rates. Given a fixed significance level α, there is evidence that the methods m_1 and m_2 produce significantly different results if $p < \alpha$. As a consequence we can reject the null-hypothesis H_0. For this work we chose a significance level of $\alpha = 0.05$. This implies that, if m_1 and m_2 are significantly different, there is a confidence level of 95% that the differences between the outcomes of the methods are not caused by random variation.

TABLE I
THE DETAILED GROUND TRUTH INFORMATION FOR THE PIT PATTERN IMAGE DATABASE USED THROUGHOUT OUR EXPERIMENTS.

	Normal	Non-invasive	Invasive	Total
Original set	72	212	43	**327**
Extended set	198	420	98	**716**
Patients	14	27	6	**47**

IV. EXPERIMENTAL SETUP & RESULTS

A. Experimental Setup

1) Image Database Used: The images used throughout our experiments are based on 327 endoscopic color images (either of size 624×533 pixels or 586×502 pixels) acquired between the years 2005 and 2009 at the Department of Gastroenterology and Hepatology (Medical University of Vienna) using a zoom-colonoscope (Olympus Evis Exera CF-Q160ZI/L) with a magnification factor of 150. In order to acquire the images 40 patients underwent colonoscopy. To obtain a larger set of images we extracted subimages with a size of 256×256 pixels from the original images, which resulted in an extended image set containing 716 images in total.

Lesions found during colonoscopy have been examined after application of dye-spraying with indigocarmine, as routinely performed in colonoscopy. Biopsies or mucosal resection have been performed in order to get a histopathological diagnosis. Biopsies have been taken from type I, II, and type V lesions, as those lesions need not to be removed or cannot be removed endoscopically. Type III and IV lesions have been removed endoscopically. Out of all images from the extended set, histopathological classification resulted in 198 non-neoplastic and 518 neoplastic cases.

As already indicated earlier we carry out our experiments based on a 3-class classification according to [5]. Hence, we aim at distinguishing between normal mucosa, non-invasive lesions, and invasive lesions. The detailed ground truth used for our experiments is shown in Table I. Since different types of lesions may develop inside the colon of a single patient such a patient may appear in more than one class. Hence, the total number of patients is slightly higher (47) as compared to the number of patients who underwent colonoscopy (40).

2) Weak Methods Set Used: In order to base our ensemble on a set of diverse feature types, the set of methods used (Q) consists of methods published in earlier work. Q contains methods which extract statistical features from the wavelet domain (WT-LDB, WPC [8], WT-BBC [9], WT-GMRF with and without custom neighborhoods [10]), operate in the spatial domain (GMRF [10]), or investigate local texture properties within images (JC-MB-LBP [11] and LCVP [12] – both with and without a multi-scale extension). But we also include methods which are based on shape-based features (DELAUNAY [13] and EDGEFEATURES [14]).

For this work, Q contains 26 different variations of these methods, which have been tested in order to obtain the respective method outcomes. To estimate the accuracy of each method we employed the Leave-One-Patient-Out Cross-Validation (LOPO-CV) protocol. In this scenario the images from one patient are considered as validation data, while the remaining images are used to train the underlying classifier (the k-NN or the Bayes classifier). This is repeated for each patient contained in the image database.

TABLE II
THE CLASSIFICATION RATES (%) OBTAINED FOR THE BEST POSSIBLE WEAK METHOD COMBINATIONS $S_i^{(B)}$.

Set	$\lvert\mathbf{S_i^{(B)}}\rvert$	Overall	Normal	Non-Invasive	Invasive
$\mathbf{S_1^{(B)}}$	4	75.56	64.14	98.57	0.00
$\mathbf{S_2^{(B)}}$	3	80.59	82.83	98.33	0.00
$\mathbf{S_3^{(B)}}$	1	80.03	70.20	91.90	48.98
$\mathbf{S_4^{(B)}}$	3	76.96	77.27	94.76	0.00
$\mathbf{S_5^{(B)}}$	4	75.84	63.64	99.29	0.00
$\mathbf{S_6^{(B)}}$	1	80.03	70.20	91.90	48.98
$\mathbf{S_7^{(B)}}$	1	76.26	75.76	88.81	23.47
$\mathbf{S_8^{(B)}}$	1	80.03	70.20	91.90	48.98
$\mathbf{S_9^{(B)}}$	1	80.03	70.20	91.90	48.98
$\mathbf{S_{10}^{(B)}}$	4	72.21	56.57	96.43	0.00

The overall classification accuracies achieved by the weak methods on the image database used in this work range from 60.34% (WPC) to 80.03% (EDGEFEATURES).

3) Ensemble Classifier Setup: The ensemble classifier, which is described in more detail in [2], [3], then combines method outcomes using majority voting. To emphasize more on the method selection ability of our proposed algorithms the method ranking has been disabled. In addition we chose a remapping parameter of 0.5, which corresponds to a linear remapping and, thus, disables the remapping too.

4) Evaluation Setup for the Proposed Methods: For the evaluation of the proposed selection methods we use the evaluation setup as shown in Fig. 4. Out of the complete set of available methods Q we create 10 subsets of methods M_i ($i = 1, \ldots, 10$), where each M_i contains 10 randomly selected weak methods out of Q. These subsets are used since testing all possible combinations from the 26 available methods in order to obtain $S^{(B)}$ would result in $2^{26} - 1$ tests, which can not be computed in an reasonable amount of time.

In order to find the best performing combination of weak methods (in terms of the overall accuracy of the ensemble) for each M_i we carry out ensemble tests for all possible combinations of the elements within M_i, which results in 10 test runs, each consisting of $2^{10} - 1 = 1023$ ensemble tests, thus resulting in 10230 ensemble tests carried out in total. Finally, we obtain the best possible combination $S_i^{(B)}$ for each test run i on M_i.

We then apply each of our selection algorithms to each M_i, which results in the weak method combinations $S_i^{(1)}$ and $S_i^{(2)}$ in case of algorithm 1 and 2, respectively.

B. Results

Table II shows the results obtained with the best combinations obtained for each test run i. One thing we immediately notice from this table is that half the sets $S_i^{(B)}$ consist of one

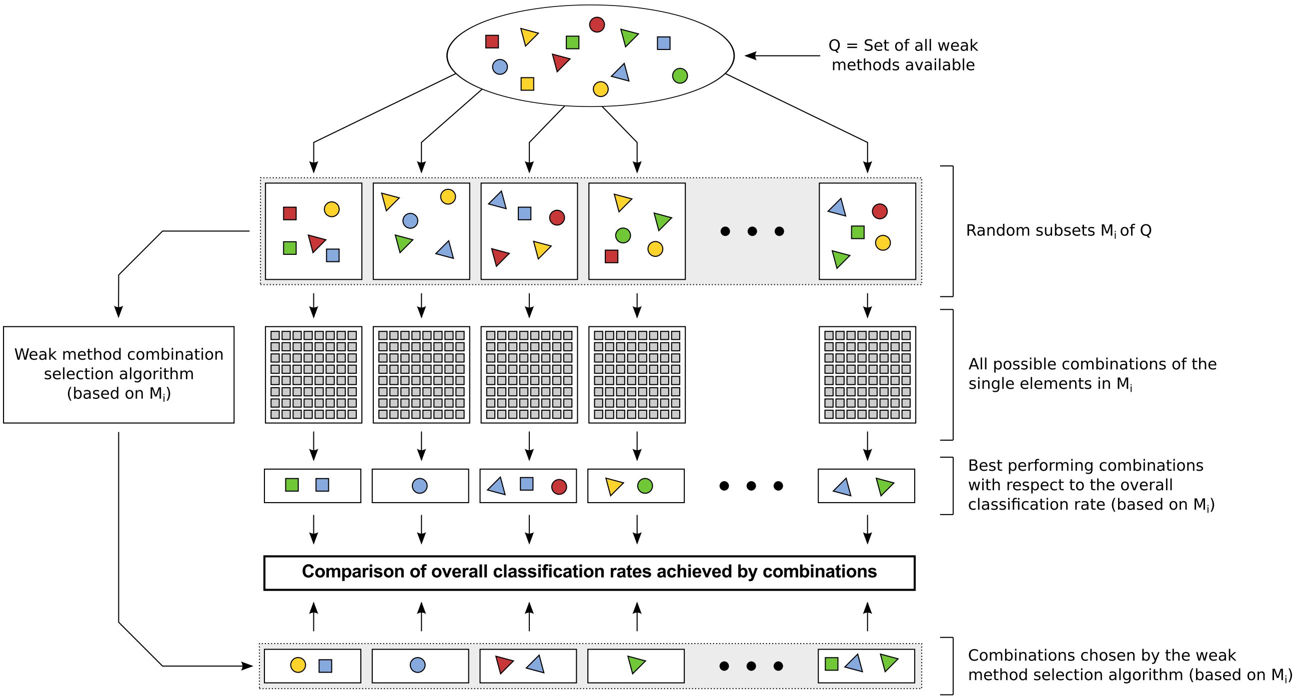


Fig. 4. Evaluation setup used to validate the selection methods proposed in this work.

TABLE III
THE CLASSIFICATION RATES (%) OBTAINED FOR THE WEAK METHOD COMBINATIONS $\mathbf{S_i^{(1)}}$ AND $\mathbf{S_i^{(2)}}$ AS CHOSEN BY OUR ALGORITHMS.

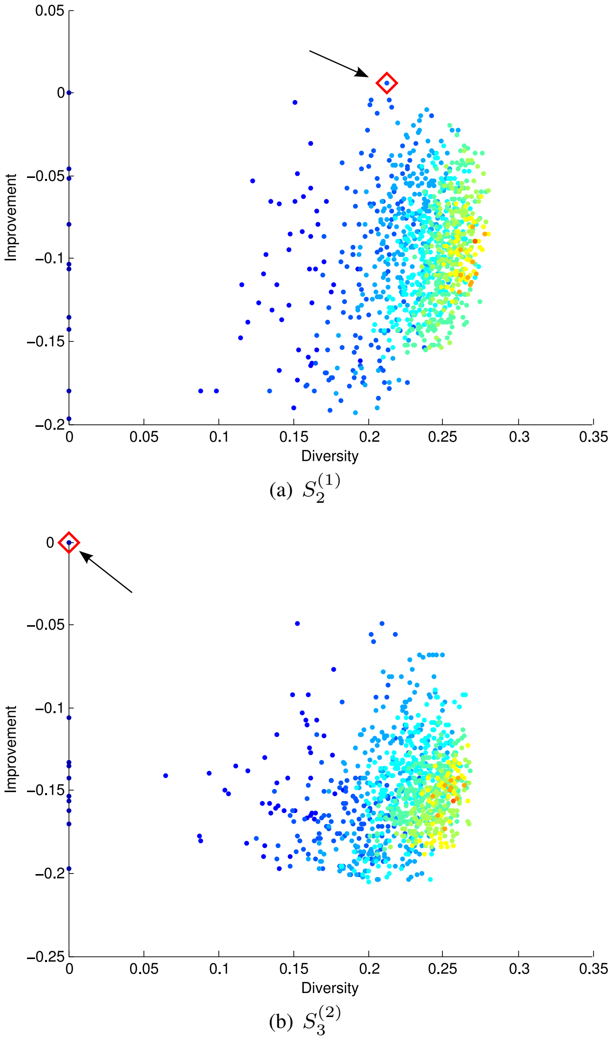
Set	$\lvert\mathbf{S_i^{(\cdot)}}\rvert$	Overall	Normal	Non-Invasive	Invasive
		Algorithm 1			
$\mathbf{S_1^{(1)}}$	3	74.86	69.70	94.76	0.00
$\mathbf{S_2^{(1)}}$	**3**	**80.59**	**82.83**	**98.33**	**0.00**
$\mathbf{S_3^{(1)}}$	3	74.02	59.60	98.10	0.00
$\mathbf{S_4^{(1)}}$	**3**	**76.96**	**77.27**	**94.76**	**0.00**
$\mathbf{S_5^{(1)}}$	3	67.32	31.31	100.00	0.00
$\mathbf{S_6^{(1)}}$	3	77.79	76.26	95.95	3.06
$\mathbf{S_7^{(1)}}$	3	69.41	44.95	97.14	0.00
$\mathbf{S_8^{(1)}}$	3	78.07	72.73	98.81	0.00
$\mathbf{S_9^{(1)}}$	3	77.51	76.77	95.95	0.00
$\mathbf{S_{10}^{(1)}}$	3	67.18	38.89	96.19	0.00
		Algorithm 2			
$\mathbf{S_1^{(2)}}$	2	74.02	63.13	96.43	0.00
$\mathbf{S_2^{(2)}}$	2	79.47	63.43	98.81	28.57
$\mathbf{S_3^{(2)}}$	**1**	**80.03**	**70.20**	**91.90**	**48.98**
$\mathbf{S_4^{(2)}}$	2	74.02	63.13	96.43	0.00
$\mathbf{S_5^{(2)}}$	2	74.30	57.58	99.52	0.00
$\mathbf{S_6^{(2)}}$	2	79.47	63.64	98.81	28.57
$\mathbf{S_7^{(2)}}$	2	70.53	48.99	95.24	8.16
$\mathbf{S_8^{(2)}}$	2	77.09	59.60	99.29	17.35
$\mathbf{S_9^{(2)}}$	2	79.47	63.64	98.81	28.57
$\mathbf{S_{10}^{(2)}}$	7	65.60	26.26	99.29	0.00

method only. Our selection algorithms however in most cases constructed bigger sets, as can be seen from Table III. It is also interesting to see that Algorithm 1 always selected 3 weak methods (for each candidate set $\mathbf{S_i^{(1)}}$).

Considering the weak method yielding a highest overall rate of 80.03% (EDGEFEATURES) we notice that in most cases the ensemble performs worse. Only in some cases the ensemble is able to slightly outperform the methods EDGEFEATURES. This can be attributed to the fact that for the evaluation of our selection algorithms we construct random subsets $M_i \in Q$ which potentially just do not include the best performing methods. In addition, the ensemble classifier can potentially be optimized further by adjusting parameters (e.g. the remapping parameter, which has been set to a linear mapping). However, in this work we are just interested in a comparison of the results yielded by the best possible combination $S_i^{(B)}$ and the combinations found by our algorithms, $S_i^{(1)}$ and $S_i^{(2)}$.

When comparing Tables II and III we also notice that for some of the test runs carried out at least one of our method picks the best possible combination (the respective results are shown in bold in Table III. But, all in all, both algorithms perform roughly equally well with respect to the ensemble result of the method combination found. While in some cases Algorithm 1 delivers higher ensemble results, in other cases Algorithm 2 is able to outperform Algorithm 1.

A comparison of the classification performance of the selected combinations against the best possible weak method combinations is possible by the plots shown in Fig. 5. The diversities for the different combinations have been computed as the average of the pair-wise double-fault measure for all pairs within a combination [6]. The colors of the dots denote the number of weak methods contained within each combination (blue denotes a single method, whereas red denotes



(a) $S_2^{(1)}$

(b) $S_3^{(2)}$

Fig. 5. Plots showing the improvements versus the diversities for all possible combinations for M_i and the combination $S_2^{(1)}$ ($i = 2$) selected by Algorithm 1 and $S_3^{(2)}$ ($i = 3$) as selected by Algorithm 2 (selected combinations are denoted by an arrow).

10 methods). The improvement is the difference between the overall rate achieved by the ensemble on each combination and the best performing method within the respective M_i.

From these plots we notice that both selection algorithms pick the best combination possible (shown in Fig. 5 by a red diamond). While Algorithm 1 picks a combination of 3 weak methods, Algorithm 2 picks a combination consisting of just one method – hence, the diversity equals zero.

We also immediately notice that a higher diversity does not automatically imply a higher improvement. Moreover, we see that most combinations degrade the result as compared to the best performing method in M_i (negative improvement). In addition, we see that, while combining more methods improves the diversity, the improvement drops for higher method counts.

V. CONCLUSION

In this work we compared our proposed weak method selection algorithms against the best possible combinations in terms of the overall classification accuracy. We showed that our methods are in most cases able to either select the best performing combination or, at least, a competitive one.

While finding an optimal combination out of a large set of potential method candidates may result in trying all possible combinations, our methods help to speed up this process too since our methods need a rather low number of iterations as compared to evaluating all possible combinations.

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

[1] M. J. Bruno, “Magnification endoscopy, high resolution endoscopy, and chromoscopy; towards a better optical diagnosis.” *Gut*, vol. 52, no. 4, pp. 7–11, Jun. 2003.

[2] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, “Pit pattern classification using multichannel features and multiclassification,” in *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*. Hershey, PA, USA: IGI Global, 2009, pp. 335–350.

[3] A. Vécsei, T. Fuhrmann, M. Liedlgruber, L. Brunauer, H. Payer, and A. Uhl, “Automated classification of duodenal imagery in celiac disease using evolved fourier feature vectors,” *Computer Methods and Programs in Biomedicine*, vol. 95, pp. S68–S78, 2009.

[4] S. Kudo, S. Hirota, T. Nakajima, S. Hosobe, H. Kusaka, T. Kobayashi, M. Himori, and A. Yagyuu, “Colorectal tumours and pit pattern,” *Journal of Clinical Pathology*, vol. 47, pp. 880–885, Oct. 1994.

[5] S. Kato, K.-I. Fu, Y. Sano, T. Fujii, Y. Saito, T. Matsuda, I. Koba, S. Yoshida, and T. Fujimori, “Magnifying colonoscopy as a non-biopsy technique for differential diagnosis of non-neoplastic and neoplastic lesions,” *World Journal of Gastroenterology*, vol. 12, no. 9, pp. 1416–1420, Mar. 2006.

[6] L. I. Kuncheva, *Combining Pattern Classifiers*. Wiley, 2004.

[7] B. Everitt, *The Analysis of Contingency Tables*. Chapman and Hall, 1977.

[8] M. Häfner, M. Liedlgruber, F. Wrba, A. Gangl, A. Vécsei, and A. Uhl, “Pit pattern classification of zoom-endoscopic colon images using wavelet texture features,” in *Proceedings of the International Conference on Advances in Medical Signal and Image Processing (MEDSIP’06)*, W. Sandham, D. Hamilton, and C. James, Eds., Glasgow, Scotland, UK, Jul. 2006, pp. 1–4.

[9] M. Liedlgruber and A. Uhl, “Statistical and structural wavelet packet features for Pit pattern classification in zoom-endoscopic colon images,” in *Proceedings of the 7th WSEAS International Conference on Wavelet Analysis & Multirate Systems (WAMUS’07)*, P. Dondon, V. Mladenov, S. Impedovo, and S. Cepisca, Eds., Arcachon, France, Oct. 2007, pp. 147–152.

[10] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, “Combining Gaussian Markov random fields with the discrete wavelet transform for endoscopic image classification,” in *Proceedings of the 17th International Conference on Digital Signal Processing (DSP’09)*, Santorini, Greece, Jul. 2009, pp. 177–182.

[11] ——, “Pit pattern classification using extended local binary patterns,” in *Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine (ITAB’09)*, Larnaca, Cyprus, Nov. 2009, pp. 1–4.

[12] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, “Color treatment in endoscopic image classification using multi-scale local color vector patterns,” *Medical Image Analysis*, accepted.

[13] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, “Classification of endoscopic images using Delaunay triangulation-based edge features,” in *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR’10)*, ser. Springer LNCS, vol. 6112, Povoa de Varzim, Portugal, Jun. 2010, pp. 131–140.

[14] ——, “Endoscopic image classification using edge-based features,” in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR’10)*, Istanbul, Turkey, Aug. 2010, pp. 2724–2727.