J Digit Imaging (2013) 26:920–931
DOI 10.1007/s10278-013-9576-9

A Supervised Learning Approach for Crohn's Disease Detection Using Higher-Order Image Statistics and a Novel Shape Asymmetry Measure

Dwarikanath Mahapatra · Peter Schueffler · Jeroen A. W. Tielbeek · Joachim M. Buhmann · Franciscus M. Vos

Published online: 8 February 2013
© Society for Imaging Informatics in Medicine 2013

Abstract Increasing incidence of Crohn's disease (CD) in the Western world has made its accurate diagnosis an important medical challenge. The current reference standard for diagnosis, colonoscopy, is time-consuming and invasive while magnetic resonance imaging (MRI) has emerged as the preferred noninvasive procedure over colonoscopy. Current MRI approaches assess rate of contrast enhancement and bowel wall thickness, and rely on extensive manual segmentation for accurate analysis. We propose a supervised learning method for the identification and localization of regions in abdominal magnetic resonance images that have been affected by CD. Low-level features like intensity and texture are used with shape asymmetry information to distinguish between diseased and normal regions. Particular emphasis is laid on a novel entropy-based shape asymmetry method and higher-order statistics like skewness and kurtosis. Multi-scale feature extraction renders the method robust. Experiments on real patient data show that our features achieve a high level of accuracy and perform better than two competing methods.

Keywords Crohn's Disease · Detection · Classifiers · Features

D. Mahapatra · P. Schueffler · J. M. Buhmann
Department of Computer Science, ETH Zurich, Zurich,
Switzerland

J. A. W. Tielbeek · F. M. Vos
Department of Radiology, Academic Medical Center, Amsterdam,
The Netherlands

F. M. Vos
Quantitative Imaging Group, Delft University of Technology,
Delft, The Netherlands

D. Mahapatra (✉)
Department of Computer Science, CAB F61.1, Universitatstrasse 6,
8092 Zurich, Switzerland
e-mail: dmahapatra@gmail.com

Introduction

Inflammatory bowel diseases (IBDs) constitute one of the largest healthcare problems in the Western world afflicting over 1 million European citizens. Out of these, nearly 700,000 suffer from Crohn's disease (CD). Crohn's disease is an autoimmune IBD that may affect any part of the gastrointestinal tract causing abdominal pain, diarrhea, vomiting, or weight loss. It is observed that, in most cases, the disease is localized in the most distant part of the small intestine, the terminal ileum. CD detection is essential to grade its severity, extent, and activity and also determines the subsequent therapeutic strategy. Currently, the reference standard for diagnosis relies on results of colonoscopy and biopsy samples [1]. Apart from therapeutic strategy, colonoscopy also has implications for patient prognosis. Through colonoscopy, a trained physician examines the ulcerations in the bowel and evaluates the severity and extent of inflammatory lesions using a standard scale called Crohn's Disease Endoscopic Index of Severity. However, the procedure is invasive, requires extensive bowel preparation, and gives information only on superficial abnormalities. Therefore, it is beneficial to have a non-invasive approach to detect CD.

Several drawbacks of colonoscopy like invasiveness, procedure-related discomfort, and risk of bowel perforation have led to the exploration of imaging techniques to assess extension and severity of IBDs [2–4]. In Bodily et al. [5], sonography and computed tomography (CT) have been explored as alternatives to colonoscopy. For young patients, exposure to ionizing radiations is a serious limitation of CT. Assessment in sonography is limited due to gas interposition. MRI has the potential to overcome these limitations because of high tissue contrast, lack of ionizing radiations, and lower incidence of adverse events related to intravenous contrast employed in CT [6]. Horsthuis et al. [6] report the diagnostic accuracy of various signs for detection of active inflammation. However, this information is not sufficient to guide therapeutic decisions. Rimola et al. [2] determine that rate of contrast enhancement and bowel wall thickness identifies the presence of endoscopically active disease. But its reliance on explicit segmentation of the bowel wall and extensive manual scoring limits its effectiveness.

Related Work on Disease Classification

There do not exist abundant research on image analysis of abdominal MRI to identify Crohn's Disease, although Bhushan et al. [7] look at dynamic contrast-enhanced (DCE) MRI for identifying colorectal cancer and [8] deal with ulcerative colitis. Atasoy et al. [9] addressed the tasks of localization, annotation, or classification of optical biopsies in colonoscopy. Our method will serve as a diagnostic tool for clinicians to localize the diseased area without the need for explicit segmentation of bowel wall.

Cheng et al. [10] propose the use of biologically inspired features for classification of different glaucoma types. The biological features are based on saliency maps predicting interesting regions for humans in pictures. Saliency-inspired features have been used in various applications like object tracking [11, 12], registration [13–15], and segmentation [16–19]. Automatic localization of multiple anatomical structures in medical images provides important semantic information for potential benefits in various clinical applications.

Pauly et al. [20] propose supervised regression techniques to detect and localize different parts in whole-body magnetic resonance (MR) sequences. They use 3D local binary patterns along with random ferns to achieve high segmentation accuracy. Kelm et al. [21] propose a random–regression-based method for detecting and grading coronary stenoses in CT angiography data. Their motivation is to provide an automated system that can rule out clinically relevant stenosis in the coronary arteries and serve as a second reader in the absence of an expert physician. Brain imaging data have seen a lot of applications of machine learning methods for classification, particularly Alzheimer's disease [22–24]. Chest images have also generated a lot of interest with methods proposed for localization of chest pathologies in Avni et al. [25], pediatric tuberculosis in Irving et al. [26], and diffuse lung disease in Xu et al. [27].

Scope of Our Work

This paper proposes a semi-automated method to detect CD-afflicted regions from input abdominal MR images. Thus, we do not need an explicit segmentation of the bowel wall which turns out to be quite difficult based on MR imagery. Our method will serve as a tool to assist clinicians, reduce reliance on colonoscopy, and help in rapid diagnosis of CD. Many studies highlight the importance of low-level features in disease identification, e.g., intensity, texture, symmetricity in shape, and structure information [28, 29]. Kovalev et al. [30] showed the importance of anisotropy (a measure of feature asymmetry) from texture maps to detect abnormalities in diseased brain images. Anisotropy can be applied to other features for characterizing diseased regions. For example, in Liu et al. [28], reflectional asymmetry based on a pigmentation model was employed to detect skin lesions.

We propose a novel method to calculate shape asymmetry that uses the entropy of orientation angles. Our approach is simple to compute and gives accurate detection results. Intensity, texture, and shape asymmetry features from multiple scales are used to characterize normal and CD-affected regions in the bowel. This paper makes the following novelties: (1) a shape asymmetry measure based on entropy of the distribution of orientation angles is proposed; I2) it is combined with higher-order image statistics to identify diseased regions; and (3) the features are used for identification of CD areas in abdominal MRI, a novel application that has immense potential in assessing IBDs. We describe our method in "Materials and Methods," present results in "Experiments and Results," and conclude with the section on "Conclusions."

Materials and Methods

Overview of Our Method

Our aim is to detect presence of Crohn's Disease in a MRI volume. For that purpose, we have to classify a voxel in the volume. For a 400×400×100 volume, the processing time can be very high if we were to analyze each individual voxel. Since the bowel is visible in only a few slices, we focus on these slices to identify CD affected regions. Furthermore, in these slices, the bowel occupies part of the image. A rectangular region of interest (ROI) is manually defined which covers the regions most likely to contain diseased tissues but is smaller compared with the whole image. Physicians get an approximate idea of the ROI from the results of other tests like colonoscopy and DCE MRI. Since our proposed approach aims to assist expert doctors, this semi-automated approach seems reasonable. We employ a two-stage classification, details of which are described in the later sections. In the first stage (henceforth referred to as Stage 1), the pixels are first classified as either intestine or background. Those pixels that are labeled as intestine are further classified as either diseased or normal in the second classification stage (Stage 2). Figure 1 shows an example slice with the ROI outline shown in red and the diseased region shown in green.

Feature Extraction

Appropriate features are crucial in determining the accuracy rate of disease identification and its subsequent grading. We

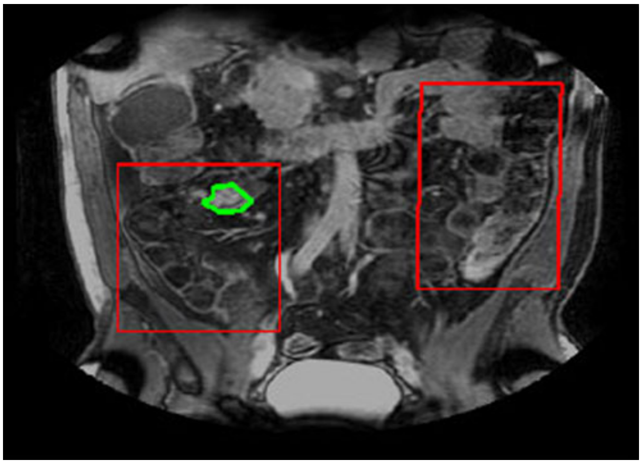


Fig. 1 Illustration showing the original slice image, the ROI in *red* and annotated region in *green*

extract intensity, texture, and shape asymmetry features from three different scales for disease identification. For higher accuracy and robustness, features were extracted from neighborhoods centered around each pixel. Henceforth, we shall refer to these neighborhoods as a patch.

Intensity and Texture Features

To identify the exact area afflicted with CD, radiologists rely on the results of different tests like colonoscopy and biopsy, as well as different imaging protocols like MR-T1, MR-T2, and DCE-MRI. A simple visual examination of T1 MRI does not provide sufficient information to identify the diseased part. We propose to investigate features that are not discernible by the human eye but may provide discriminating features for our task. Psychophysical experiments have established that the human visual system is sensitive only to image features of the first- and second-order (mean and variance) [31]. It is common in MR images to have regions that do not form distinct spatial patterns but differ in their third order statistics, e.g., boundaries of some malignant tumors are diffuse and invisible to the naked eye [32]. However, with computational tools at our disposal, we can analyze higher-order statistics to determine their efficacy in disease classification. For every image patch (neighborhood of a pixel), we calculate the mean, variance, skewness, and kurtosis for intensity and texture features. Let us denote each patch as *Si*, the intensity of its *j*th pixel as *Si*(*j*), the mean intensity by *Si,* and the variance by σ_i^2. The skewness (third-order statistic) of the intensity distribution is given by

$$Sk = \left[\frac{1}{N}\sum\nolimits_{i=1}^{n}\left(S_i - \overline{S}_l\right)^3\right] \times \frac{1}{\sigma_i^3} \tag{1}$$

where the term within the square brackets is the third-order moment. The skewness is a measure of symmetry of a distribution. Negative skew indicates that the bulk of the values lie to the right of the mean, and positive skew indicates the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean (e.g., a Gaussian distribution).

Kurtosis (the fourth-order statistic) is a measure of the "peakedness" of the probability distribution and along with skewness describes its shape. It is given by

$$Ku = \left[\frac{1}{N}\sum\nolimits_{i=1}^{n}\left(S_i - \overline{S}_l\right)^4\right] \times \frac{1}{\sigma_i^4} \tag{2}$$

where the term within the square brackets is the fourth-order moment. A high kurtosis distribution has a sharper peak and longer, fatter tails, while a low kurtosis distribution has a more rounded peak and shorter, thinner tails.

Texture Maps Texture can be modeled as patterns which are distinguished by a high concentration of localized spatial frequencies. 2-D Gabor filter banks, which have optimal joint localization in the spatial and spatial-frequency domains, have been used for texture representation in Manjunath and Ma, and Liu and H. Wechsler [33, 34]. Due to its multiscale and multi-orientation structure, Gabor filter banks conform to the receptive fields profiles of simple cortical cells [35] and are able to capture rich visual properties such as spatial localization, orientation selection, and spatial frequency characteristics. Since Gabor filters incorporate Gaussian smoothing, they are robust to noise. Because of these desirable properties, we select the Gabor filter bank to characterize texture in our images. The Gabor filter bank can be represented as

$$g_{\gamma,\omega}(x,y) = a^{\gamma} g(a^{\gamma}(x\cos(\omega\psi) + y\sin(\omega\psi))a^{\gamma}(-x\sin(\omega\psi) + y\cos(\omega\psi))) \tag{3}$$

where γ=0,... $\Gamma-1$, ω=0,...$\Omega-1$. The mother function *g* is Gaussian defined as:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right) exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi jWx\right] \tag{4}$$

Γ=6 is the total number of orientations, Ω=2 is the total number of scales. The rotation factor $\psi=\pi/\Omega$ and the scaling factor $a=(Uh/Ul)^{1\ \Gamma\ -1}$. *Uh* and *Ul* are parameters that determine the frequency range of the filter bank, and *W* is a shifting parameter in the frequency domain. Texture maps are obtained using oriented Gabor filters along six directions (0°, 30°, 60°, 90°, 120°, and 150°).

Shape Asymmetry

In Liu et al. [28], the authors use reflectional asymmetry to identify skin lesions. Also, in Liu et al. [28], global point signatures (GPSs) were defined at each pixel to characterize color and orientation information, and the asymmetry descriptor is calculated based on bin differences about a principal axis of a histogram. The principal axis is chosen from one of 180 bins, and some of the bins are translated to ensure that there are 90 bins on each side of the principal axis. We propose a novel method for quantifying the shape asymmetry which is based on the entropy of the orientation distributions. Entropy-based measures have been used in computer vision studies to detect salient regions [12] and tracking [11]. This novel measure is simpler to calculate and also functions as a good descriptor.

Each patch is divided into 18 sectors of a circle centered at the geometric center of the patch. For each sector, we calculate orientation angles of the pixels and determine the entropy of the angle distribution. Entropy measures the information in the respective distributions. A uniform distribution has high entropy while the entropy is low for a peaked distribution. If the orientation angles are distributed over many angle ranges, there is greater asymmetry in shape, leading to a higher entropy value. On the other hand, low entropy values indicate that most of the pixel orientation angles are distributed along a few directions leading to lower shape asymmetry. The shape asymmetry measure for a sector r is given by

$$Sh^{\mathrm{r}}_{\mathrm{Asymmetry}} = -\sum\nolimits_{\theta} p^r_\theta \log p^r_\theta \qquad (5)$$

p^r_θ denotes the angle distributions in sector *r*. Figure 2a shows an illustration of a circle divided into 18 sectors. Figure 2b shows a patch around a diseased pixel, and Fig. 2c shows a map giving the value of the corresponding orientation angles. Figure 2d shows a patch around a normal pixel with the corresponding angle map shown in Fig. 2e. The angle values are in the range [180°, 180°]. Figure 2f shows the plot of entropy values for each of the 18 sectors for the two patches shown in Fig. 2b and d. It is interesting to note that the orientation profile for the normal patch is quite regular as compared with the diseased patch. This is indicative of the fact that the orientation in diseased regions becomes distorted due to ulcerations or other abnormalities. Thus, they lose the regularity observed in healthy tissues. This is corroborated by the plot in Fig. 2f where the diseased patches show higher entropy, indicating greater randomness.

The above set of features give a 46-dimension feature vector (intensity, 4; texture, 6×4×2=48; and shape, 18) for a single patch. In order to capture information over multiple scales, we extract similar features over neighborhood of different sizes. Thus, for each annotated pixel, features were extracted over three neighborhoods of sizes 25×25, 30×30, and 35×35. The final feature vector is of length 46×3=138. Although the Gabor texture maps are inherently multi-scale representations, intensity and shape measures are not derived from multiple scales. Thus, to include intensity and shape features from multiple scales, we consider three neighborhoods. Although the computation cost is slightly more, it leads to higher classification accuracy. A brief discussion is found in the section on "Importance of Multi-scale Feature Extraction" on the importance of such feature extraction.

Features for Comparison

We compare the effectiveness of our method with two other methods. The first is a wavelet–transform-based method (dual tree complex wavelet transform (DTCWT)) in Berks et al. [29]. The second is a shape–asymmetry-based method (Asy), where asymmetry is calculated similar to Liu et al.'s [28]. Instead of GPS, orientation angles are used.

DTCWT Wavelet transforms have been used extensively in image processing and analysis to provide a rich description of local structure. The DTCWT has particular advantages because it provides a directionally selective representation with approximately shift-invariant coefficient magnitudes and local phase transformations [36]. DTCWT combines outputs of two discrete transforms (differing in phase by 90° to form the real and imaginary parts of complex coefficients. For 2-D images, it produces six directional sub-bands oriented at ±15°,±45°, and ±75° at a series of scales separated by a factor of 2. A feature vector is constructed by sampling DTCWT coefficients from six oriented sub-bands from a $w \times w$ neighborhood centered on the pixel. This gives an 18-dimension feature vector (i.e., six sub-bands from the three neighborhoods stated above). Details of the method can be found in Berks et al. [29].

Asy Liu et al. [28] define a measure of *reflectional* asymmetry based on the pigmentation model of skin lesions. We use a similar approach for an alternative shape asymmetry metric where, instead of the GPS, we calculate shape asymmetry based on orientation angle information. For every patch, the distribution of orientation angles is calculated such that the magnitude of angles lies between 180° and 180° and the number of bins equals 40 with an equal number of bins for positive and negative magnitudes. The shape asymmetry is determined as

$$Asy = \sum\nolimits_{i=1}^{20} (h_i - h_{-i}) \qquad (6)$$

Here $h-i$ denotes the histogram count of *i*th bin of negative magnitudes, and *hi* denotes the histogram count of the corresponding bin of positive magnitudes. Note that we do not take the absolute value of differences. A net positive value of *Asy* indicates that the orientation angles have a net positive leaning, while a negative *Asy* indicates a net negative leaning. Thus, the sign and magnitude of *Asy* indicates an "asymmetry vector."

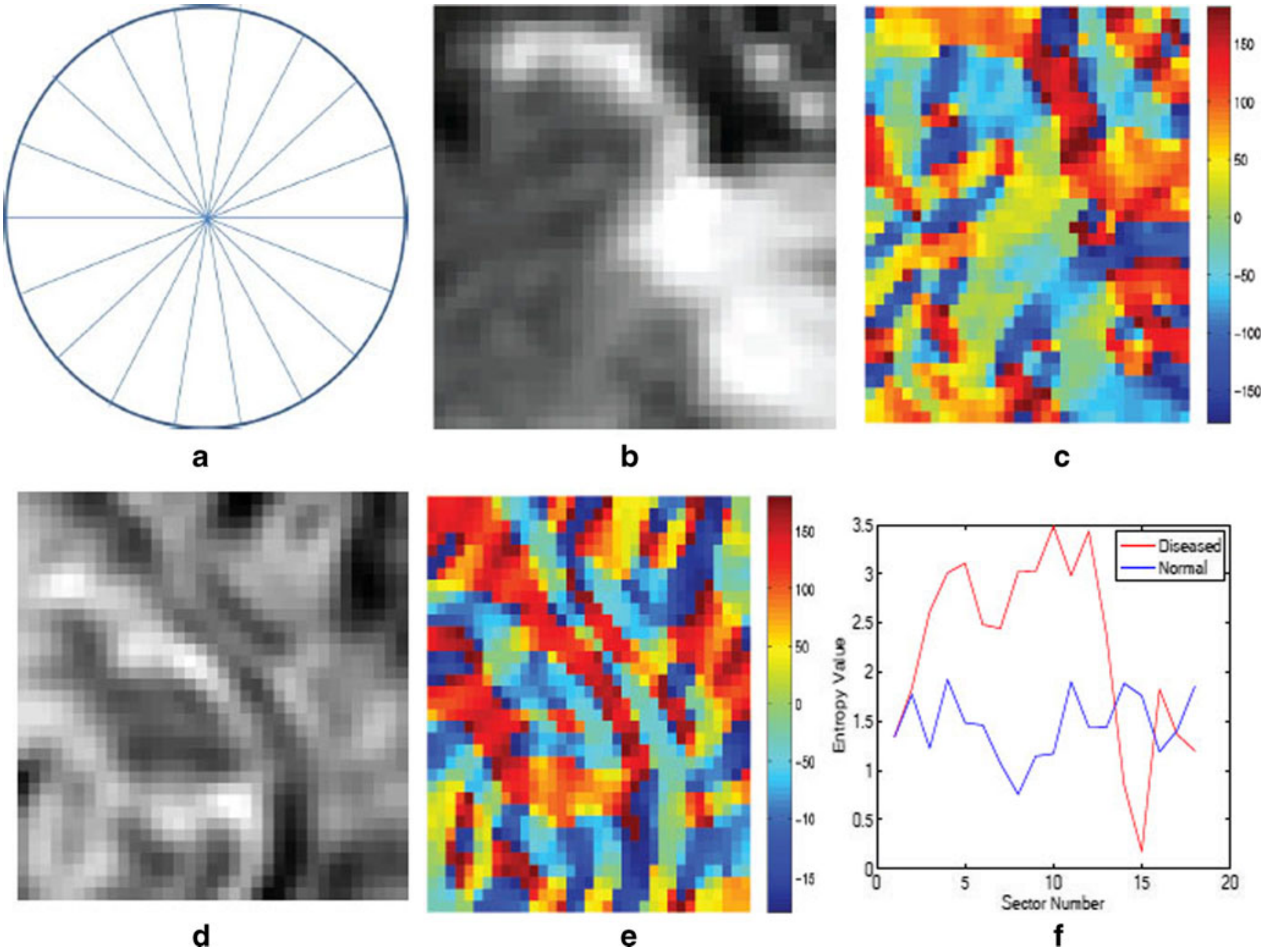


Fig. 2 **a** Illustration of sectors for entropy, **b**, **c** diseased patch, and corresponding map of orientation angles; **d**, **e** normal patch and corresponding map of orientation angles; **f** plot of entropy values for the two patches. The *color bars* show the angle magnitude in degrees

Experiments and Results

The datasets comprised of samples from 26 patients diagnosed with CD (mean age, 36 years; range, 19–72 years; 17 females and 9 males) at the Academic Medical Center (AMC), Amsterdam, The Netherlands. Two radiologists with more than 7 years of experience with dealing with abdominal MR images annotated regions (with consensus) corresponding to diseased, normal, and background (normal nonintestine) on the 26 patients. Patients fasted 4 h before a scan and drank 1,600 ml of mannitol (2.5 %) (Baxter, Utrecht, The Netherlands) 1 h before the scan. Images were acquired with patients in supine position using a 3-T MR imaging unit (Intera, Philips Healthcare, Best, The Netherlands) with a 16-channel torso phased array body coil. The protocol consists of axial and coronal single-shot fast-spin echo (SSFSE) sequences followed by a coronal fat saturated SSFSE sequence and coronal 3D T1-weighted spoiled gradient echo sequence. The spatial resolution of the images was 1.02×1.02×2 mm, and the acquired volume dimension was 400×400×100 pixels.

The number of samples was 6,827 from diseased regions, 5,156 from normal, and 3,725 from background regions. Here, each sample refers to the feature vector from a pixel's multiple neighborhoods. We employ a tenfold cross-validation (leave-one-out) approach for testing our classifier's performance. Each sample class is divided into roughly ten equal parts, of which nine parts are used for training and the other one for testing. The study was approved by the AMC's ethics committee and waived informed consent. The study was in accordance with the rules of the European Community's Seventh Framework Programme. Our method was compared with *DCTWT* and *Asy*. Each of the three methods was evaluated using three different classifiers, random forests (RF), support vector machines (SVM), and a Bayesian classifier (BC).

Classifiers

Random forests [37] have been successful in a variety of domains and compare favorably with other state-of-the-art algorithms [38]. A random forest is an ensemble of decision

Table 1 Quantitative measures for the *Stage* 1 classification for various features and using different classifiers

	Asy			DTCWT			*OurFeatures*		
	SVM	BC	RF	SVM	BC	RF	SVM	BC	RF
Accuracy (%)	80.4 (2.6)	72.0 (2.3)	79.9 (2.2)	82.2 (2.4)	71.3 (2.9)	80.1 (2.5)	86.4 (1.5)	73.2 (4.4)	83.3 (4.1)
Specificity (%)	67.9 (1.8)	41.5 (1.8)	68.0 (1.7)	68.1 (1.6)	42.7 (1.7)	67.6 (1.8)	71.1 (1.8)	49.1 (2.1)	70.6 (2.2)
Sensitivity (%)	86.2 (1.9)	81.5 (1.4)	84.6 (1.8)	93.9 (2.7)	88.3 (2.1)	85.7 (1.9)	96.7 (1.2)	90.1 (6.5)	92.1 (4.1)
Precision (%)	90.1 (1.1)	77.9 (1.8)	89.6 (1.3)	92.1 (2.1)	78.8 (1.7)	89.5 (1.1)	96.3 (1.8)	80.4 (4.2)	90.9 (2.9)

Values indicate mean (standard deviation)

trees where each tree is trained with a different subset of the training data to improve the classifier's generalization ability. Training finds the set of tests that best separate the training data into different classes. Random forests and their variants have been used to detect abnormalities in mammograms [29] and identify coronary artery stenoses [21] and semantic segmentation in CT images [39]. In our experiments, 100 trees were used for the RF classifier.

SVMs construct a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data of any class (so-called functional margin). In general, the larger the margin, the lower the generalization error of the classifier. SVMs have also seen wide application in classification tasks like brain tumor segmentation [40, 41], chest pathologies [25], and glaucoma classification [10], among others. For SVMs, we use the LIBSVM package [42] and define a radial basis function (RBF) as the kernel.

The default naive Bayesian classifier in MATLAB was the third classifier. A Bayesian classifier was chosen to highlight the linearly non-separable nature of the data and the advantages of having a RBF kernel in SVMs. We have two classification stages for all classifiers. For all classifiers, we employ tenfold cross-validation (leave-one-out with ten subsets of the original data) approach.

Classification Results for Stage 1

Table 1 shows the average accuracy and sensitivity of the first classification stage using tenfold cross-validation. Here, each sample is classified as either intestine or background. The highest classification accuracy is obtained using our features, the results of which are shown in the box plots of Fig. 3. In this stage, we desire a high sensitivity or true-positive rate (TPR) even at the expense of low overall accuracy. True positive refers to an intestine sample correctly classified as intestine. We do not want an intestine sample to be incorrectly labeled as background, thus increasing the false-negative rate (FNR). In such a situation, the diseased samples (which are part of intestine in the first stage) get classified as background and, hence, escape the scrutiny of the next stage. This is particularly undesirable in a clinical decision making system.

In trying to reduce the FNR, we observe an increase in false-positives, i.e., increasing number of background regions are identified as intestine. This situation does not adversely affect the outcome of the decision system. The background regions classified as intestine are invariably identified as normal in the second classification stage. Those

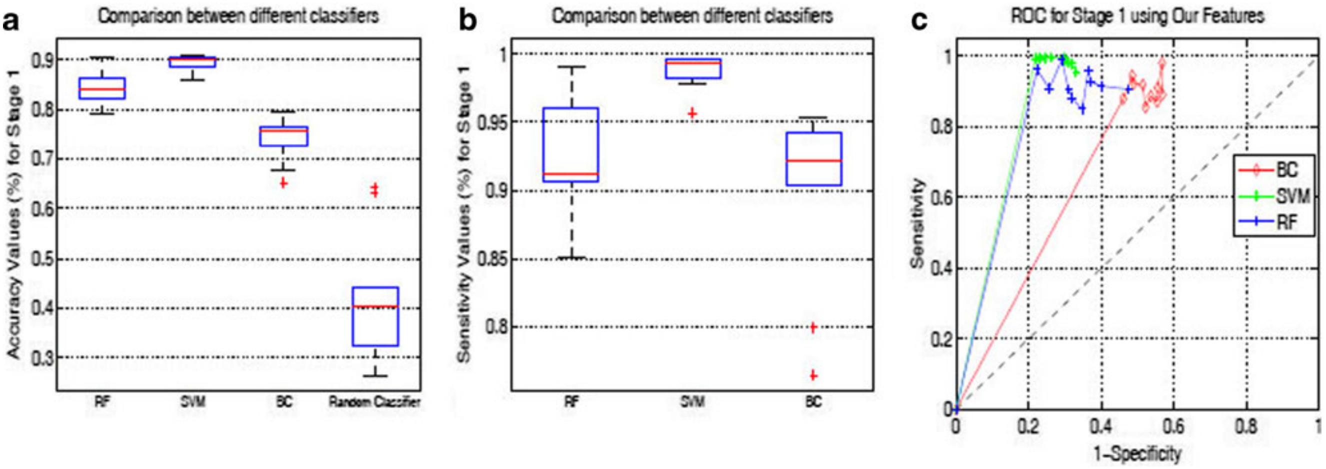


Fig. 3 *Box plots* for *Stage* 1 classification: **a** Accuracy; **b** Sensitivity. **c** ROC curves for three classifiers using *OurFeatures*

Table 2 Quantitative measures for individual and different combination of features using RF classifier

	Int	Tex	Shape	Tex+Int	Shape+Int	Shape+Tex
Accuracy (%)	77.1 (2.3)	81.6 (2.1)	79.1 (2.7)	79.2 (1.3)	79.5 (2.4)	82.3 (1.3)
Sensitivity (%)	79.3 (3.2)	86.9 (2.1)	82.3 (1.9)	83.1 (3.1)	83.8 (2.3)	86.6 (2.8)

Values indicate mean (standard deviation)

background regions identified as diseased in the second stage can be discarded by examination of a clinician.

Classifiers' Performance in Stage 1 A comparison of ROC curves of all three methods using RF classifier in Stage 1 is shown in Fig. 3c. During tenfold cross-validation, sensitivity and specificity values are obtained for each run which are then used to draw the ROC curves. All the three methods give high sensitivity (more than 90 %), but their specificity values are comparatively lower, indicating a large number of false-positives. The overall accuracy (i.e., correct classification percentage of both intestine and background samples) is lower than 86 % in all cases. This indicates a high number of false positives, i.e., many background samples are classified as intestine. This is not a disadvantage since these incorrectly labeled background samples are invariably identified as normal in Stage 2. It is interesting to note that BC has an overall accuracy less than 75 %, because of a high number of false-positives. This is reflected in the specificity values in Table 1 and the ROC curves in Fig. 3c. This gives some interesting insight into the nature of data and performance of classifiers. BC can find a good decision boundary for data which are linearly separable, or which have low overlap amongst the feature values. For different classes which have similar feature values, BC produces lots of errors. The performance of different classifiers suggests that our data have such characteristics. In spite of the low specificity, BC shows an accuracy of 74 % because the number of background samples is about a third of the number of intestine samples. Had the number of samples been equal, then BC's accuracy would have been still lower. All classifiers perform better than a random classifier.

In order to achieve good discrimination, it is important that the samples in the training set represent all possible variations in the real world. However, this is not always possible in practice. Feature selection plays an important role in order to make the best use of the available samples. Intensity, texture, and shape asymmetry provide lot of important discriminating information between the different classes. An additional factor is multi-scale feature extraction which overcomes issues like choosing an appropriate neighborhood and sufficient neighbors to get a stable statistical measure.

Importance of Different Features

We also investigate the importance of different features in our classification framework. Here, we report the results of a combination of different features using the RF classifier. Table 2 shows different quantitative measures for Stage 1 classification using only intensity, texture, and shape asymmetry and their different combinations. As expected, the sensitivity and accuracy for the individual features are lower than the values in Table 1. The combination of texture and shape features produces results that are closest to the values in Table 1. However, this does not indicate that intensity information is unimportant. Conducting a *t*test on the values for *Tex+Shape* (Table 2) and *All Features* (Table 1) gives $p<0.032$ which clearly indicates statistically different results. Furthermore, we also conduct *t*tests for features *Tex* versus *Tex-Int*, and *Shape* versus *Shape-Int*. In both cases, we find that $p<0.04$, thus clearly showing that inclusion of intensity statistics improves classification accuracy.

DTCWT Versus Tex DTCWT also calculates texture maps of an image. While *DTCWT* only calculates the mean across different orientations and scales, *Tex* calculates mean, variance, skewness, and kurtosis across orientation and scales.

Table 3 Quantitative measures for the *Stage* 2 classification for various features and using different classifiers

	Asy			DTCWT			*OurFeatures*		
	SVM	BC	RF	SVM	BC	RF	SVM	BC	RF
Accuracy (%)	81.5 (1.3)	59.1 (0.9)	81.7 (1.2)	82.2 (1.4)	58.4 (6.1)	81.9 (1.2)	89.5 (2.6)	62.8 (5.4)	88.9 (1.5)
Specificity (%)	80.8 (3.1)	35.1 (4.8)	83.4 (2.4)	82.8 (1.4)	37.7 (2.7)	84.2 (1.9)	90.2 (1.7)	39.3 (4.1)	90.1 (1.6)
Sensitivity (%)	84.5 (1.9)	60.5 (1.2)	84.9 (1.8)	86.9 (1.7)	61.3 (8.2)	86.1 (1.9)	91.9 (2.6)	64.8 (9.7)	90.4 (1.2)
Precision (%)	82.9 (1.4)	59.7 (1.9)	82.7 (1.5)	85.3 (1.4)	59.7 (5.4)	84.9 (1.4)	90.2 (2.0)	63.3 (4.3)	88.9 (1.3)

Values indicate mean (standard deviation)

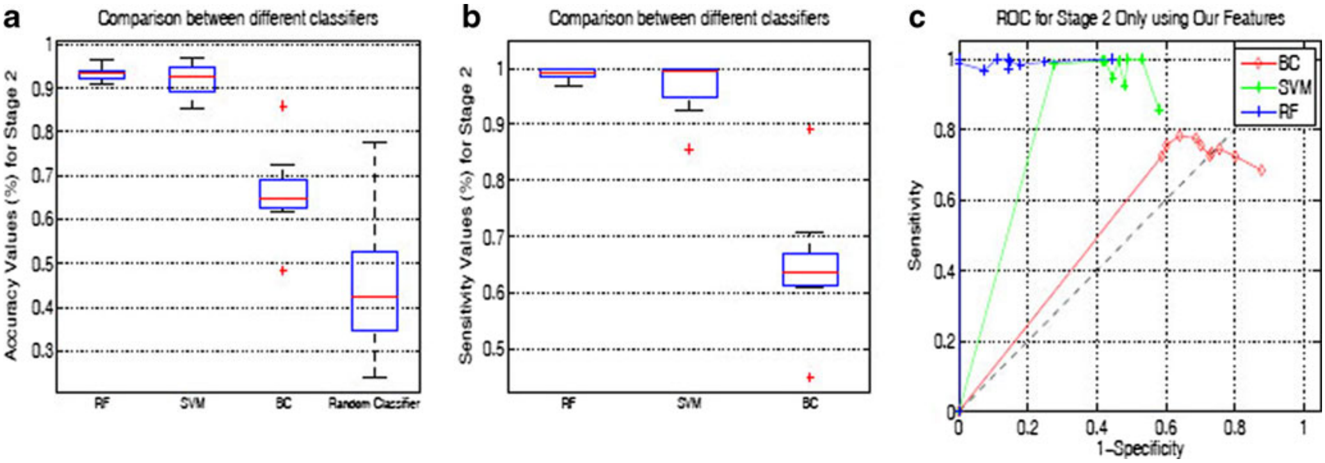
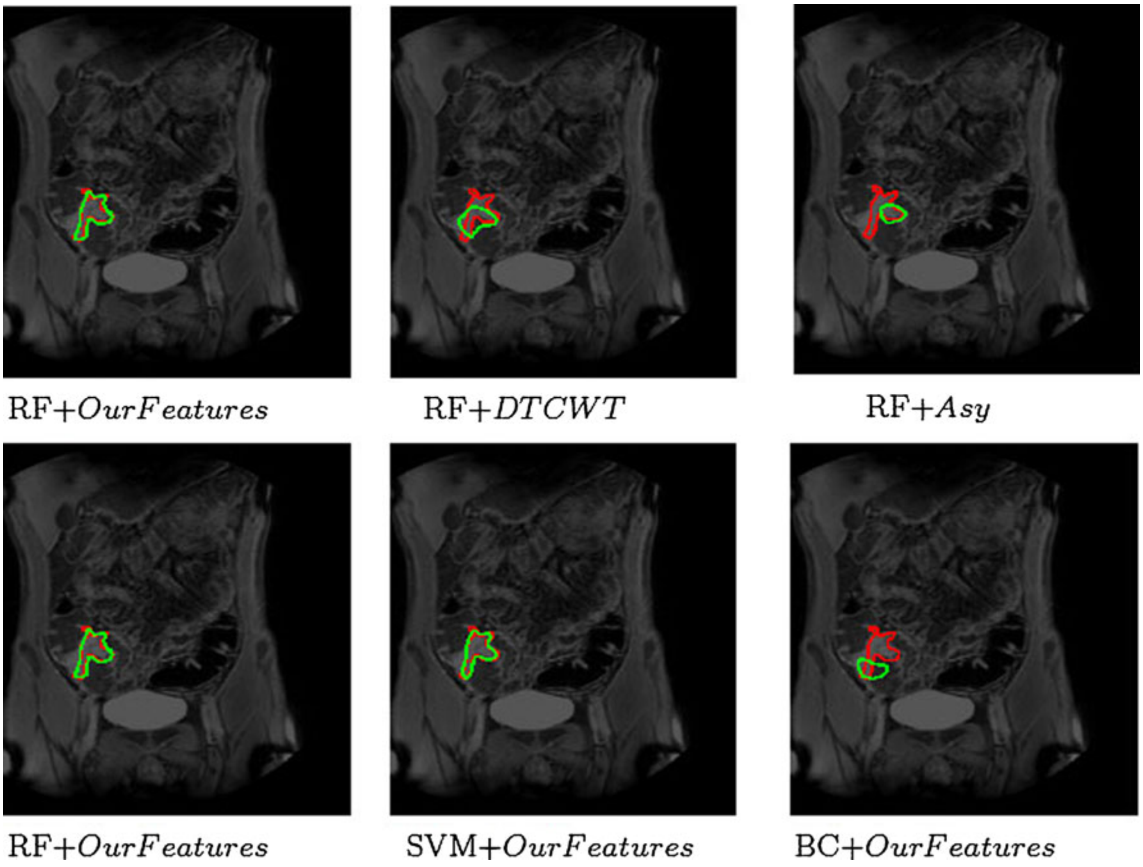


Fig. 4 Box plots for *Stage* 2 classification: **a** Accuracy; **b** Sensitivity. **c** ROC curves for three classifiers using *Ourfeatures*

Thus, it is expected that *Tex* would be a more accurate measure than *DTCWT*, and it is reflected in the quantitative measures for *Tex* (Table 2) and *DTCWT* (Table 1).

Asy Versus Shape The difference between the two features is that, while *Asy* operates on the whole patch, *Shape* operates on different sectors of the patch. They also differ in their approach to determining asymmetry. *Asy* calculates asymmetry as the difference in "positive" and "negative" histograms of orientation distributions. On the other hand, *Shape* characterizes asymmetry as the entropy of orientation angles in each sector. A comparison of the results for *Shape* (Table 2) and *Asy* (Table 1) show that their performance is very similar, with $p=0.13$.


RF+$OurFeatures$ RF+$DTCWT$ RF+Asy

RF+$OurFeatures$ SVM+$OurFeatures$ BC+$OurFeatures$

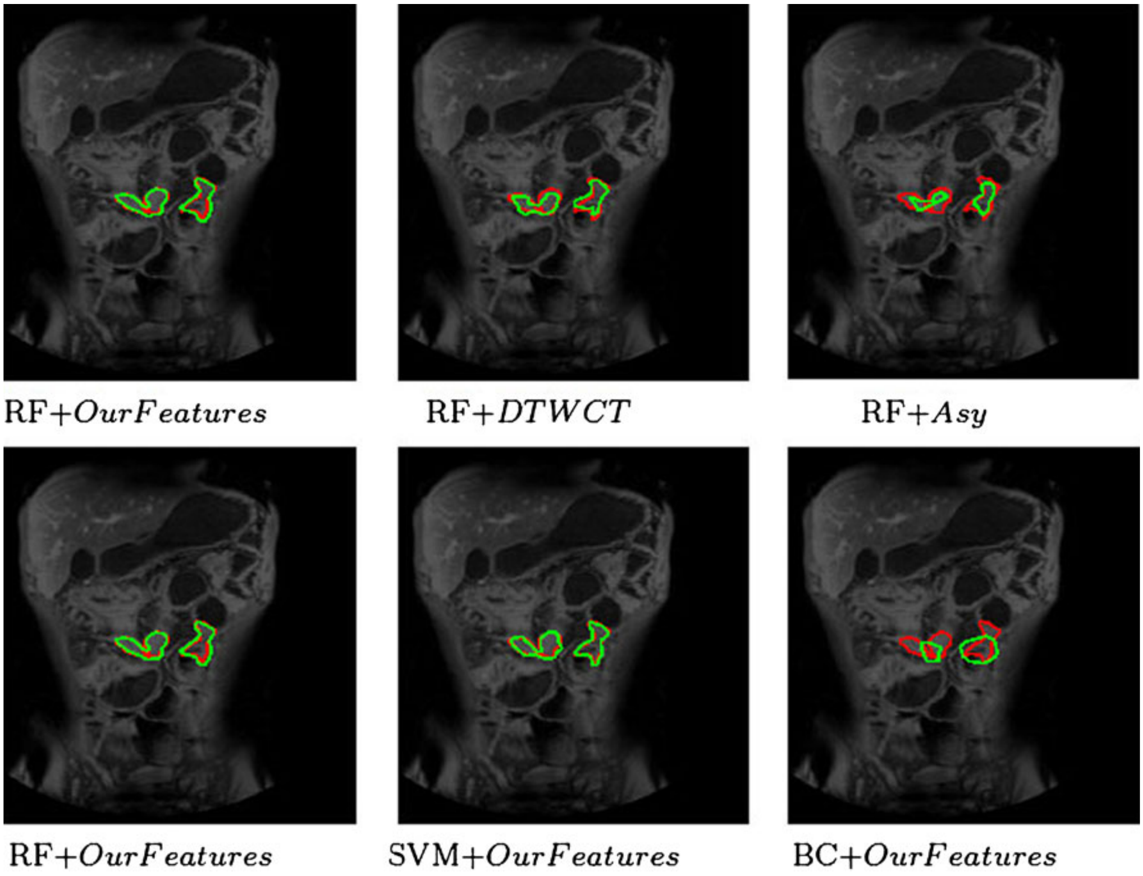
Fig. 5 Visual results for CD detection in Patient 23. *First row* shows results of RF using different features (our features, *DTCWT*, and *Asy*). *Second row* shows performance of different classifiers (RF, SVM, and BC) using our features. The two images in the first column are identical and are shown for continuity

Classification Results for Stage 2

Those samples that are identified as intestine in the first stage are considered in Stage 2 for further classification into diseased or normal. Suppose there are *N* number of intestine samples at the beginning of Stage 1, out of which *Nd* are diseased and *Nn* are normal, i.e., *Nd*+*Nn*=*N*. In *Stage* 1, let *N*2 intestine samples be correctly classified, out which *Nd*2 are diseased and *Nn*2 are normal. The sensitivity of *Stage* 1 is calculated using *N*2 and *N*. The *N*2 number of samples is now considered for classification in *Stage* 2. At the end of *Stage* 2, suppose the number of correctly classified diseased samples to be *Nd*3 and the number of correctly classified normal samples, *Nn*3. The different quantitative measures for *Stage* 2 are based on the original number of samples at the beginning of *Stage* 1, i.e., *Nd* and *Nn*. Below, we shall define the different measures:

1. Accuracy: the number of diseased and normal samples that were correctly classified, i.e., (*Nd*3+*Nn*3)/*N*.
2. True-positives (TP): number of correctly classified diseased samples(*Nd*3).
3. True-negatives (TN): number of correctly classified normal samples (*Nn*3).
4. False-positives (FP): number of normal samples classified as diseased (*Nn*-*Nn*3).
5. False-negatives (FN): number of diseased samples identified as normal. (*Nd*-*Nd*3).
6. True-positive rate (TPR) : $TPR = \frac{TP}{TP+FN} = \frac{N_{d3}}{N_d}$. It is the same as sensitivity and recall.
7. True-negative rate (TNR) : $TNR = \frac{TN}{TN+Fp} = \frac{N_{n3}}{N_n}$. It is the same as specificity.
8. 8. Precision $= \frac{\mathrm{TP}}{\mathrm{TP+FP}} = \frac{\mathrm{N_{da}}}{\mathrm{N_{d3}+N_n-N_{n3}}}$. It is the fraction of retrieved instances that are relevant. In this case, retrieving diseased samples is relevant.

Table 3 shows the accuracy and sensitivity after the final classification stage, i.e., an indicator of performance over the original *N* samples. Although BC's accuracy and TPR in *Stage* 1 was comparable to RF and SVM, it showed significantly worse results for *Stage* 2. From the results of BC, we infer that the feature values of diseased and normal samples are not linearly separable. This supports our use of features from multiple scales. The high difference of measures between BC


RF+$OurFeatures$

RF+$DTWCT$

RF+Asy

RF+$OurFeatures$

SVM+$OurFeatures$

BC+$OurFeatures$

Fig. 6 Visual results for CD detection in Patient 16. *First row* shows performance of RF using different features (our features, *DTCWT,* and *Asy*). *Second row* shows performance of different classifiers (RF, SVM, and BC) using our features. The two images in the first column are identical and are shown for continuity

and the other classifiers in *Stage* 2 as compared with the corresponding values in *Stage* 1 can be explained as follows. The extracted features of diseased and normal samples have lower differences than the feature values of background and intestine (normal and diseased) samples. Thus, in *Stage* 1, BC did a good job because intestine and background samples formed "clusters" which were reasonably distant. However, diseased and normal samples seem to form clusters that are not very far apart. This would explain the low accuracy, sensitivity, and specificity values for BC in *Stage* 2. These observations make a stronger case for RF and SVMs using RBF kernels.

Figure 4 shows the box plots of accuracy and sensitivity, and ROC curves when we consider *Stage* 2 as different classification process without any link to *Stage* 1. The accuracy and sensitivity are calculated only over the samples that pass *Stage* 1, i.e., *TRP* (Sensitivity)= $\frac{N_{da}}{N_d}$; Accuracy=$(N_{d3}+N_{n3})/(N_{d2}+N_{n2})$. Obviously, the values will be higher than those reported in Table 3 (which are based on all the original number of samples). The box plots and ROC curves indicate that a high percentage of each sample type is correctly classified by both SVM and RF. This is highly desirable because, ultimately, we would like to detect the diseased regions from abdominal MRI. During tenfold cross-validation, sensitivity and specificity values are obtained for each run which are then used to draw the ROC curves.

Results on Real Patient MRI

Figure 5 shows visual results for CD detection in Patient 23. After classification, we generally get one large cluster and smaller isolated clusters. Isolated clusters with less than 10 pixels were removed, and the large cluster was transformed into a single continuous region using contour fitting. The first row shows the detection results obtained using the RF classifier with the three methods namely our features, *DTCWT* and *Asy*. The manually annotated diseased regions are shown in red while the result of automatic detection is shown in green. The figures show that our features give the best performance. *DTCWT* and *Asy* both detect less number of diseased pixels, which can prove to be critical in decision making systems. They also identify many background or normal pixels as diseased (false-positives), while using our features the number of false-positives is very low.

Table 5 Accuracy percentage and computation time for multi-scale feature extraction and comparison with other methods

	Asy	DTCWT	Our Features	1 Scale	2 Scales
Accuracy (%)	81.7 (1.2)	81.9 (1.2)	88.9 (1.5)	84.3 (1.9)	80.1 (1.5)
Computation time (s)	349 (14)	404 (21)	463 (24)	415 (22)	381 (17)

Values indicate mean (standard deviation)

The second row of the same figure shows the detection results using our features with different classifiers, namely RF, SVM, and BC. The two figures in the first column are the same but have been shown for consistency. Although our features give the best performance using RF, the results vary depending upon the choice of classifier. SVM's performance is close to that of RFs in terms of high true-positives and low false-positives. However, BC shows significantly poor performance than the other two classifiers with low true-positives and high false-positives because of its limited ability to handle non-linearly separable data. We would like to point out that, if we use a linear kernel in SVMs, it performs poorly than RFs, and is closer to BC's performance. Figure 6 shows the detection results on Patient 16, where there are two diseased regions to be detected. Again, our features perform better than *DTCWT* and *Asy*; RF and SVM show similar performance, and BC performs poorly.

The manual annotations can be treated as ground truth segmentations. The pixels identified as diseased or normal can be grouped together to form automatic segmentations. Thus, we can calculate measures like Dice metric (DM) and Hausdorff distance (HD) between the two regions. The average DM and HD values are summarized in Table 4.

Importance of Multi-scale Feature Extraction

Although multi-scale feature extraction leads to a slight increase in computation time, it contributes toward to higher classification accuracy. Table 5 summarizes the performance of our features for different neighborhoods using RF classifiers over all samples. Results are shown only after Stage 2, although the analysis was carried out for both Stage 1 and Stage 2. *1 Scale* refers to features extracted from a 25×25 neighborhood; *2*

Table 4 Average DM and HD values after *Stage 2* using different features and classifiers

	Asy			DTCWT			*OurFeatures*		
	SVM	BC	RF	SVM	BC	RF	SVM	BC	RF
DM (%)	84.1 (1.1)	79.3 (2.4)	84.7 (1.5)	85.3 (2.1)	80.1 (2.3)	85.6 (1.4)	90.3 (2.1)	81.8 (2.1)	90.9 (1.2)
HD (pixels)	3.4 (2.1)	6.3 (2.8)	3.5 (1.8)	3.1 (1.2)	6.7 (2.2)	3.2 (2.1)	2.1 (1.1)	4.8 (1.6)	2.0 (1.1)

Values indicate mean (standard deviation)

Scales refers to features extracted from a 25×25 and 30×30 neighborhoods. The computation time is shown for the whole pipeline consisting of Stage 1 and Stage 2 for a 100×30 manually drawn ROI. The results clearly indicate that *Our Features* (multi-scale feature extraction) improves accuracy measures significantly over 1 *Scale* ($p < 0.01$ from *t*tests) and *2 Scales* ($p < 0.018$). However, the corresponding extra computation time is not significantly high. Therefore, we can include multi-scale features without a large increase in computation time.

Conclusion

In this paper, we have proposed a method to identify regions in the human gastrointestinal tract that are afflicted with Crohn's disease. Higher order intensity and texture statistics, and shape asymmetry information are extracted at multiple scales and used to discriminate between diseased, normal, and background regions. Higher-order statistics capture image properties that are not discernible to the human eye. Our shape asymmetry measure is simple to compute and is informative in detecting diseased regions. We compare our features with a wavelet–transform-based feature (*DTCWT*) and asymmetry descriptor (*Asy*) using three standard classifiers. Experimental results show that our designed feature vector performs better than *Asy* and *DTCWT*. Our results indicate that Crohn's disease can be detected from MR images and, thus, reduce reliance on invasive procedures like colonoscopy and biopsy. With further improvements of our method in the future, we can hope to build a reliable detection and CD classification system.

Our system also has its limitations: (1) manual definition of ROI is not always possible. Therefore, we are working on a reliable method to detect regions of disease activity (i.e., the ROIs). Once an approximate ROI is identified, each pixel within it can be further analyzed for disease activity; (2) a two-stage classification takes lot of time which can be reduced by efficient coding and faster feature extraction; (3) post-processing to remove isolated pixel clusters increases the manual involvement. We are working on a solution using segmentation frameworks (like graph cuts) that inherently impose spatial smoothness constraints and produce a coherent segmentation of diseased regions.

Acknowledgment This research was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013): the VIGOR++ Project (grant agreement number 270379).

References

1. Mary JY, Modigliani R: Development and validation of an endoscopic index of the severity for Crohns Disease: a prospective multicentre study. Gut. 30(7):983–989, 1989
2. Rimola J, Rodriguez S, Garcia-Bosch O, et al: Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. Gut 58:1113–1120, 2009
3. Vos FM, Tielbeek J, Naziroglu R, Li Z, Sch*ue*ffler P, Mahapatra D, Alexander Wiebel, Lavini C, Buhmann J, Hege H, Stoker J, and van Vliet L: "Computational modeling for assessment of IBD: to be or not to be?," in *Proc. IEEE EMBC*, pp. 3974–3977, 2012
4. Mahapatra D, Sch*ue*ffler P, Tielbeek J, Buhmann JM, and Vos F.M: "A supervised learning based approach to detect Crohn's Disease in abdominal MR volumes," in *Proc. MICCAI workshop Computational and Clinical Applications in Abdominal Imaging(MICCAI-ABD)*, pp. 97–106, 2012
5. Bodily KD, Fletcher JG, Solem CA, et al: Crohn disease: mural attenuation and thickness at contrast-enhanced CT enterography correlation with endoscopic and histologic findings of inflammation. Radiology 238(2):505–516, 2006
6. Horsthuis K, Bipat S, Bennink RJ, Stoker J: Inflammatory bowel disease diagnosed with US, MR, scintigraphy, and CT meta-analysis of prospective studies. Radiology 247(1):64–79, 2008
7. Bhushan M, Schnabel JA, Risser L, Heinrich MP, Brady JM, and Jenkinson M:"Motion correction and parameter estimation in DCEMRI sequences: application to colorectal cancer.," in *MICCAI*, pp. 476–483, 2011
8. Schunk K: Small bowel magnetic resonance imaging for inflammatory bowel disease. Top Magn Reson Imaging. 13 (6):409–25, 2002
9. Atasoy S, Mateus D, Meining A, Yang G-Zh, and Navab N: "Targeted optical biopsies for surveillance endoscopies," in *MICCAI*, pp. 83–90, 2011
10. Cheng J, Tao D, Wong DWK, Lee BH, et al.:"Focal biologically inspired feature for glaucoma type detection," in *MICCAI, part 3*, pp. 91–98, 2011
11. Mahapatra D, Saini MK, and Y. Sun: "Illumination invariant tracking in office environments using neurobiology-saliency based particle filter," in *IEEE ICME*, pp.953–956, 2008
12. Mahapatra D, Winkler S, and Yen SC: "Motion saliency outweighs other low-level features while watching videos," in *SPIE HVEI.*, pp. 1–10, 2008
13. Mahapatra D and Sun Y: "Registration of dynamic renal MR images using neurobiological model of saliency," in *Proc. ISBI*, pp. 1119–1122, 2008
14. Mahapatra D, Sun Y: Mrf based intensity invariant elastic registration of cardiac perfusion images using saliency information. IEEE Trans. Biomed. Engg. 58(4):991–1000, 2011
15. Mahapatra D and Sun Y: "Nonrigid registration of dynamic renal MR images using a saliency based MRF model," in *Proc. MICCAI*, pp. 771–779, 2008
16. Mahapatra D and Sun Y: "Joint registration and segmentation of dynamic cardiac perfusion images using MRFs.," in *Proc. MICCAI*, pp. 493–501, 2010
17. Mahapatra D, Sun Y: Integrating segmentation information for improved MRF based elastic image registration. IEEE Trans. Imag. Proc. 21(1):170–183, 2012
18. Mahapatra D and Sun Y: "Orientation histograms as shape priors for left ventricle segmentation using graph cuts," in *In Proc: MICCAI*, pp. 420–427, 2011
19. Mahapatra D and Sun Y: "Using saliency features for graphcut segmentation of perfusion kidney images," in *In 13th International Conference on Biomedical Engineering*, pp. 639–642, 2008
20. Pauly O, Glocker B, Criminisi A, and D. Mateus: "Fast multiple organ detection and localization in whole-body MR Dixon sequences," in *MICCAI*, pp. 239–247, 2011
21. Kelm BM, Mittal S, Zheng Y, et al.: "Detection, grading and classification of coronary stenoses in computed tomography angiography," in *MICCAI*, pp. 25–32, 2011

22. Iglesias JE, Jiang J, Liu C-Y, and Tu Z: "Classification of Alzheimers disease using a self-smoothing operator," in *MICCAI*, pp. 58–65, 2011
23. Zhang D, Wang Y, Zhou L, Yuan H, Shen D: Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55(3):856–867, 2011
24. Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM: "Detection of prodromal Alzheimer's via pattern classification of MRI. Neurobiology of Aging 29(4):514–523, 2008
25. Avni U, Greenspan H, and Goldberger J: "X-ray categorization and spatial localization of chest pathologies," in *MICCAI*, pp. 199–206, 2011
26. Irving B, Goussard P, Gie R, Todd-Pokropek A, and P. Taylor: "Identification of paediatric tuberculosis from airway shape features," in *MICCAI*, pp. 133–140, 2011
27. Xu R, Hirano Y, Tachibana R, and Kido S: "Classification of diffuse lung disease patterns on high-resolution computed tomography by a bag of words approach," in *MICCAI*, pp. 183–190, 2011
28. Liu Z, Smith L, Sun J, Smith M, and R. Warr: "Biological indexes based reflectional asymmetry for classifying cutaneous lesions," in *MICCAI*, pp. 124–132, 2011
29. Berks M, Chen Z, Astley S, and Taylor C: "Detecting and classifying linear structures in mammograms using random forests," in *IPMI*, pp. 510–524, 2011
30. Kovalev VA, Petrou M, Bondar YS: Texture anisotropy in 3D images. IEEETrans. Imag. Proc 8(3):346–360, 1999
31. Julesz B, Gilbert EN, Shepp LA, Frisch HL: Inability of humans to discriminate between visual textures that agree in second-order statistics-revisited. Perception 2(4):391–405, 1973
32. Petrou M, Kovalev VA, Reichenbach JR: Three-dimensional nonlinear invisible boundary detection. IEEE Trans. Imag. Proc 15 (10):3020–3032, 2006
33. Manjunath BS, Ma WY: Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Mach. Intell 18 (8):837–842, 1996
34. Liu C, Wechsler H: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. Image Process. 11(4):467–476, 2002
35. De Valois RL, Albrecht DG, Thorell LG: Spatial-frequency selectivity of cells in macaque visual cortex. Vis. Res. 22(5):545–559, 1982
36. Kingsbury N: Complex wavelets for shift invariant analysis and filtering of signals. Applied and Computational harmonic analysis 10(3):234–253, 2001
37. Breiman L: Random forests. Machine Learning 45(1):5–32, 2001
38. Fuchs TJ, Buhmann JM: Computational pathology: challenges and promises for tissue analysis. Comp Med Imag Graphics 35(7–8):515–530, 2011
39. Montillo A, Shotton J, Winn J, Iglesias JE, Metaxas D, and Criminisi A: "Entangled decision forests and their application for semantic segmentation of ct images," in *MICCAI*, pp. 184–196, 2011
40. Bauer S, Nolte L-P, and Reyes M: "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *MICCAI*, pp. 354–361, 2011
41. Verma R, Zacharaki E, Ou Y, Cai H, Chawla S, Lee S, Melhem E, Wolf R, Davatzikos C: Multiparametric tissue characterization of brain neoplasms and their recurrence using pattern classification of MR images. Acad. Radiol. 15(8):966–977, 2008
42. Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm