A Model Development Pipeline for Crohn's Disease Severity Assessment from Magnetic Resonance Images

Peter J. Schüffler[1,*], Dwarikanath Mahapatra[1], Jeroen A.W. Tielbeek[2], Franciscus M. Vos[2, 3], Jesica Makanyanga[4], Doug A. Pendsé[4], C. Yung Nio[2], Jaap Stoker[2], Stuart A. Taylor[4, 5], and Joachim M. Buhmann[1]

[1] Dept of Computer Science, ETH Zurich, Universitätstrasse 6, Zurich, CH
[2] Dept of Radiology, Academic Medical Center, Meibergdreef 9, Amsterdam, NL
[3] Quantitative Imaging Group, Delft University of Technology, Lorentzweg 1, Delft, NL
[4] Centre for Medical Imaging, University College London, 250 Euston Road, London, UK
[5] Dept of Radiology, University College Hospital London, 235 Euston Road, London, UK
peter.schueffler@inf.ethz.ch

Abstract. Crohn's Disease affects the intestinal tract of a patient and can have varying severity which influences treatment strategy. The clinical severity score CDEIS (Crohn's Disease Endoscopic Index of severity) ranges from 0 to 44 and is measured by endoscopy. In this paper we investigate the potential of non-invasive magnetic resonance imaging to assess this severity, together with the underlying question which features are most relevant for this estimation task. We propose a new general and modular pipeline that uses machine learning techniques to quantify disease severity from MR images and show its value on Crohn's Disease severity assessment on 30 patients scored by 4 medical experts. With the pipeline, we can obtain a magnetic resonance imaging score which outperforms two existing reference scores MaRIA and AIS.

Keywords: Crohn's Disease, abdominal MRI, CDEIS, MaRIA, AIS.

1 Introduction

Crohn's Disease is a chronic Inflammatory Bowel Disease that often affects the terminal ileum and colon causing inflammation, stenoses, fistula and ulcers. Symptoms of the disease include abdominal pain, diarrhea and weight loss due to a malfunction of the bowel. While the exact cause of Crohn's Disease (CD) is not known, it is thought to be a multifactorial mixture of environmental influences having an adverse effect on the immune system of genetically predisposed people. The treatment of patients in different stages of the disease include autoimmune suppressives and antibiotics as well as surgery of affected parts of the bowel in severe cases. CD patients undergo a regularly examination in which the severity of CD is determined. The severity and activity of the disease has direct influence on the current treatment strategy.

* Corresponding author.

H. Yoshida et al. (Eds.): Abdominal Imaging 2013, LNCS 8198, pp. 1–10, 2013.

A state-of-the-art score for this disease severity is the Crohn's Disease Endoscopic Index of Severity (CDEIS), which is determined by ileo-colonoscopy. Severity scores are assigned to the five bowel segments *rectum, sigmoid and descend colon, transverse colon, ascended colon, terminal ileum.* Three points are added to the mean segmental score if ulcerated stenoses or non-ulcerated stenoses are present to define the patient's CDEIS. Colonoscopy in general is time consuming, uncomfortable (partly painful) for the patient and has inherent limitations as e.g. hindered accessibility after stenoses. Therefore, there is ongoing research into alternative imaging methods such as magnetic resonance imaging (MRI). In this paper, we investigate to what extent MRI can serve as basis for the measurement of CD severity. We make following contributions: 1) developing a new systematic pipeline for model generation for CD severity assessment based on MRI; 2) proposing a plausible method for feature selection within this pipeline; 3) developing a new MRI based severity score for Crohn's Disease with the help of this pipeline. The new score is evaluated with its Pearson correlation to the CDEIS and compared to existing scores. Our findings can help in determining the relevant features to be addressed when it comes to automated MRI analysis in this context. Fully automated localization, calibration and segmentation are recently central bottlenecks for computer driven feature extraction.

1.1 Related Work

Two published MRI based CD related scores are used as reference: the MaRIA model [1] and the AIS [2]. The MaRIA score serves as a baseline for our experiments, since it is optimized for CDEIS correlation. However, it uses the critical feature of relative contrast enhancement (RCE), which is very time consuming and highly subjective to measure. The scores are defined as:

$$\boldsymbol{MaRIA} = 1.5 * wall\ thickness + 0.02 * RCE + 5 * edema + 10 * ulceration$$

$$\boldsymbol{AIS} = 1.79 + 1.34 * mural\ thickness + 0.94 * mural\ T2\ signal$$

2 Methods

To develop a MRI based severity score, we propose a new systematic machine learning pipeline, which in principal can be applied on similar problems in computational radiology (Fig. 1). (1) Driven by the target score CDEIS which is calculated on each bowel segment, the first step in the pipeline is segment-wise feature extraction from MRI volumes. (2) An exhaustive search is systematically performed throughout all features for feature and model selection. This step incorporates combinatorial feature selection, which are subjected to linear regression models, and patient-wise cross-validation of the models. After ranking the models according to the median correlation to CDEIS, we discover a class of models with similar performance. (3) We propose a method to identify this set by P-values and show how a feature distribution in this set can support the model selection in step 3. (4) The selected model is validated on a separated dataset which has not been used in steps 1-3.

Fig. 1. Model development pipeline proposed in this paper. Each step is modular and can be adjusted to various kinds of biomedical problems. The feature distribution in this pipeline is an unconventional but very effective step.

2.1 Dataset and Feature Extraction

In total, we use the data of 30 CD patients, of which 20 patients were used for brute force model generation and ranking (including cross-validation) and 10 patients for independent testing. These numbers are competitive for this disease (e.g. 16 patients for training / 26 for testing in [2]).

MRI scans of 30 luminal CD patients with written consent of data usage were collected at the Academic Medical Center (AMC), Amsterdam [3]. For bowel distention, patients drank 1,6l of Mannitol (2.5%, Baxter, Utrecht, The Netherlands). 1 hour later, MRI sequences were aquired with a 3.0T scanner (Intera, Philips Healthcare, Best, The Netherlands): axial and coronal T2-weighted single shot fast spin echo (SSFSE) sequences w/ and w/o fat saturation, as well as a coronal 3D T1-weighted spoiled gradient echo (SPGE) sequence with fat saturation. Thereafter, the injected contrast agent butylscopalaminebromide (20mg, Buscopan, Boehringer, Ingelheim, Germany) helped for a dynamic contrast enhanced (DCE-MRI) MRI sequence with gadobutrol (0.1 ml/kg, Gadovist 1.0 mmol/ml, Bayer Schering Pharma, Berlin, Germany). A coronal DCE-MRI sequence with 450 scans over 6 min (temporal resolution: 0.82s, spatial resolution: 2.78x2.78x2.5mm) was performed. Then, a second dose of butylscopalaminebromide (20mg) was injected, followed by postcontrast axial and coronal 3D T1-weighted SPGE sequences with fat saturation [3]. The DCE-MRI sequence was not used in this study. Fig. 2 shows two typical images of two patients. The scans were independently visually examined by four radiology experts. Each patient's bowel was partitioned into five segments. Bowel segments are visually identified by the radiologists in a standardized way: e.g. the ileocoecal valve, splenic flexure and hepatic flexure were used to separate the different colon segments. Each segment is then visually rated according to 17 CD related features by every expert:

Per bowel segment: *mural_thickness, muralT2, perimural_T2, pattern, enhancement_T1, comb_sign, abcess, fistula, length, wall_thickness, rce, edema, ulcers and pseudopolyps,*
Per patient: *lymph_nodes_pP, node_enhancement_pP and enlarged_lymphnodes_pP*

These features comprise MRI features described in the literature and used by most abdominal radiologists as identified in an international inventory, as well as those used in the two published scoring systems. For a medical description of these features,

please see [4]. Note that there is no computer screened feature available for CD, yet. We focus exactly on this aspect by investigating which visible features are relevant. However, research on automated MRI analysis with automated feature extraction is ongoing [5, 6].

Out of the potential 150 bowel segments, 7 segments could not be assessed by the radiologists and 6 segments were not accessible by endoscopy, resulting in 137 usable bowel segments with features and label. Every bowel segment has one CDEIS score as label from the colonoscopy as described before (score ranges from 0 to 38), allocated by one expert gastroenterologist. Since most segments are not affected by CD, 96 segments have a CDEIS score 0 and 41 segments a score greater than 0.

The inter-observer variance among the MRI features between the 4 radiologists is relatively low, implying that the feature findings in the scans are reproducible (*cf.* [4]). To cover the low variance between different domain experts and to exclude a potential bias according to one expert, the samples of the four observers are bundled together resulting in a dataset of 137*4=548 samples. Each sample has 17 features. 188 samples of 10 random patients were separated for later independent testing.

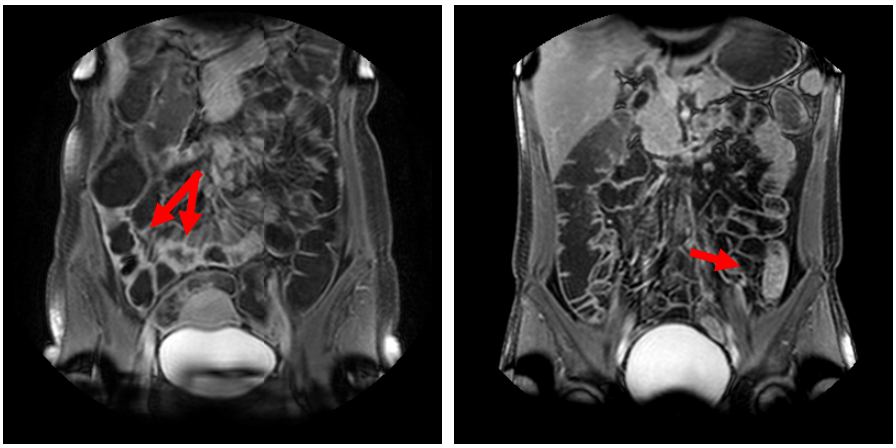

Fig. 2. Two typical MR images from two patients, showing parts of the small bowel and colon. Enhanced regions were annotated by a MD to be associated with CD.

2.2 Feature / Model Selection

An exhaustive search algorithm throughout all 17 features was conducted for feature selection. All 131071 possible combinations of the features were subjected to linear regression models [7], which are easily interpretable, computationally fast, less prone to overtraining and easily comparable to other approaches [8], [1] and [2]. Each model was validated in a 50-fold patient wise bootstrapped cross-validation experiment on the data of our 20 patient data (360 samples). In this scenario, 20 patients are randomly drawn *with replacement* out of the available 20 patients, resulting in roughly

13 different patients per draw. The model is trained on the samples of these patients and tested on the remaining 7 out-of-bag patients. The tested Pearson correlation is recorded and the procedure is repeated 50 times. In every iteration, a different random sample is drawn from the 20-patients dataset. Due to the highly variable nature of biological data, and due to the limited amount of training data, this bootstrapped validation approach provides a realistic simulation to mimic the heterogeneity of the whole population. Note that for every model, the same bootstrap-folds are drawn in each iteration to facilitate an accurate comparison of the different models. The stratification into complete patients is necessary to account for potential dependencies among samples within one patient. All data of a specific patient is contained either in the training set or in the test set, but not distributed over both sets during one iteration.

3 Results

3.1 Model for Severity Assessment

In the exhaustive search approach, all 131071 (2^{17}-1) possible linear regression models were ranked by their median cross-validated correlation to the targeted CDEIS. The top-ranked "**Model 1**" has a cross-correlation of r=.65, which tends to be higher than the MaRIA (r=.56, P=0.07, Paired t-test) and is significantly higher than the AIS (r=.55, P=0.01), even if they are retrained (both r=.60, P<0.01), meaning an improvement of 8-18 %.

$$\boldsymbol{Model1} = 2.89 * enhancement\ T1 + 7.08 * combsign + 4.95 * edema + 16.62 * ulcers$$

Model 1 uses only 4 of 17 features, illustrating that more features do not necessarily improve the correlation. Fig. 3 shows the cross-validated correlation of model 1 to CDEIS, compared to two alternative scores (Model 23 and 63), a random model, the MaRIA and the AIS. We will explain Model 23 and 63 in the next section.

3.2 Feature/Model Distribution

Although the top model shows a significant high performance, it may not be reported as the final problem solution, since the second best and third best models (and so on) show similar results in the cross-validation (Fig. 3). Actually, the correlation coefficients of the best 116 models do not differ significantly from Model 1 (P<0.05, Bonferroni corrected for multiple testing [9]).

Fig. 4A shows the ranked median correlations of all 131071 possible models with a color code for the number of features used by the corresponding models. From a machine learning point of view, it is difficult to justify the top ranked model to be the final solution, while there exists a class of models with statistically similar performance. We solve this problem with a feature distribution within this class, which

enables selecting favorite models according to additional criteria. As depicted in Fig. 4B, the feature distribution among the best 116 models reveals the prominent importance of the features *comb_sign*, *ulcers* and *enhancement_T1*, as they appear in nearly all of these models. The model using only these three features is ranked on position 23 and has a median correlation to CDEIS of r=.64.

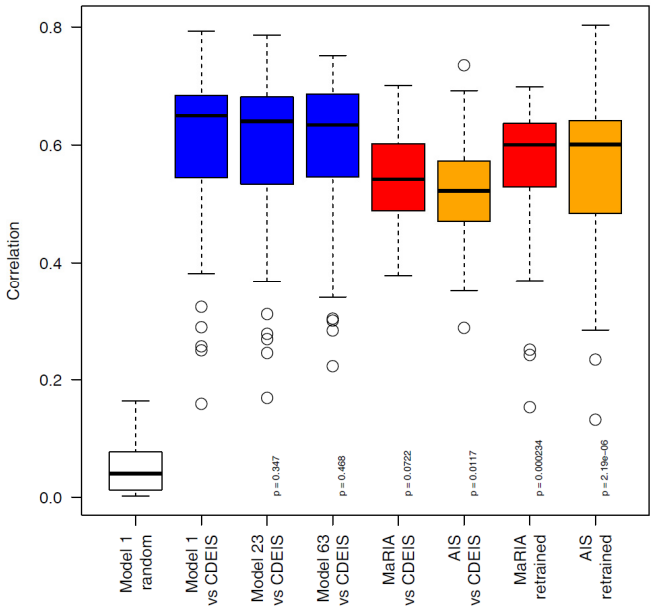

Fig. 3. Pearson correlation performance of our top 3 MRI models. Y-axis: absolute correlation coefficient. For comparison, a random model (left, white) is trained on the features of Model 1, but with randomly permuted CDEIS label before cross-validation. Top ranked models 1, 23 and 63 (blue) have a median correlation of r=.65 to r=.63. The MaRIA (red) and AIS (orange) are cross-validated on our dataset with and without retraining (i.e. adjusting the weights to our dataset and using the weights as proposed by the models (see Introduction)). P-values (paired t-test) express the significant differences to Model 1.

$$\boldsymbol{Model23} = 4.51 * enhancement\ T1 + 8.21 * combsign + 18.37 * ulcers$$

If a low number of features is wished, we can deliver Model 23 as a proper solution of the problem. Indeed, the contribution of additional features to the three will not result in statistically higher cross-validated correlation. Nevertheless, we continue with Model 1, as a low number of features is not mandatory for the specific CD problem.

Another criterion for model selection might be low variance during cross-validation. Model 63 (r=.63) shows lowest variance among the top ranked models:

$$\boldsymbol{Model63} = 2.06 * muralT2 + 7.19 * combsign + 3.12 * length + 0.03 * rce$$

As in our specific problem, it is more important not to use the expensive RCE feature than to have low variance, we do not consider Model 63 as final solution.

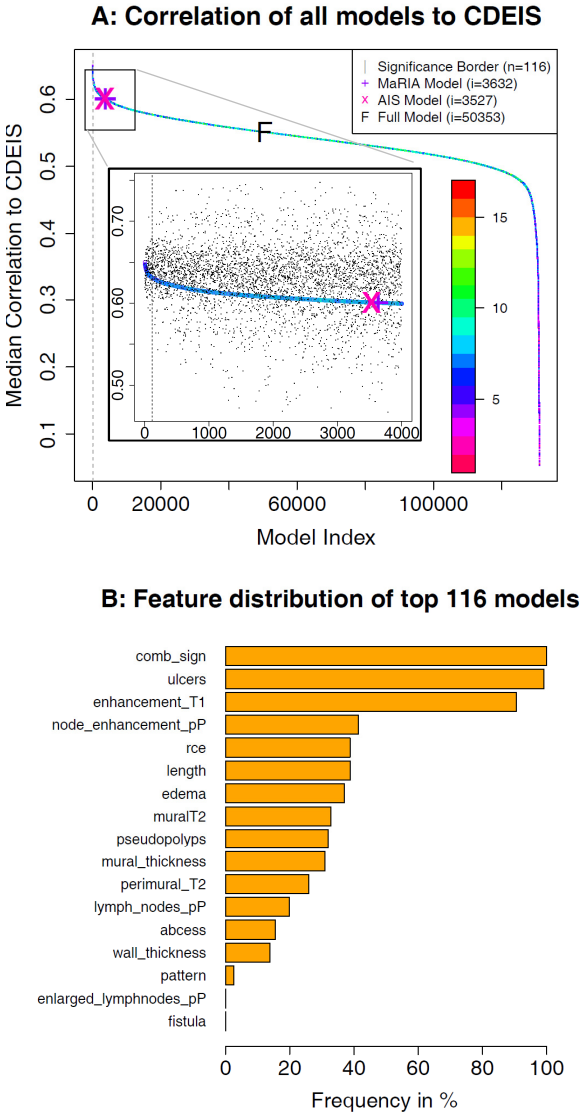

Fig. 4. A: Median cross-validated Pearson correlations of all 131071 possible models. X-axis: model index of ranked models. Color encodes the number of features used per model. The univariate models have weaker correlation. Models 1-116 (gray vertical line in zoom rectangle) do not differ statistically significantly to model 1 (paired t-test $p \geq 0.05$, Bonferroni corrected for multiple testing). +, x and F show the position of the models using MaRIA features, AIS features or all 17 features, respectively. Additional to the colored median CV correlation, the zoom rectangle shows to every model the corresponding correlation on the external test set. **B:** Feature distribution of MRI features in the best 116 models, which do not differ significantly in performance (gray line in Fig. 4A). The top three of the predictors were used in nearly 100% of all 116 models, illustrating their importance for regression.

3.3 Independent Model Testing

Since the result of the cross-validation experiment was used for model selection (by ranking), an independent validation step is needed to consider the issue of overfitting. This test performance should ideally perform in the same range as the cross-validation correlation. As stated in section Methods, we use the samples of 10 randomly chosen patients of the original data as test set. All other 20 patients are then used for training. The test results are reported in Table 1. Model 1 outperforms all other models on the test set by 6-25%. The significance of the differences of the models in the test set are tested in a leave-one-out cross-correlation.

Table 1. Pearson correlation to CDEIS computed by bootstrapped cross-validation and on 10 independent patients. Although MaRIA and AIS are fully parameterized models, we considered them either with **ret**rained weights (ret) or **app**lied without retrained weights (app) before testing. The independent test set is stratified by the four observers. N is the number of bowel segments in the sets. P-values of paired t-tests show the partly significant differences to Model 1.

		Model 1	MaRIA ret	AIS ret	MaRIA app	AIS app
Cross-Validation (n=360)	μ ± σ median t-test P	.59 ± .15 .65	.56 ± .12 .60 2.3e-4	.56 ± .14 .60 2.2e-6	.54 ± .07 .54 .072	.52 ± .09 .52 .012
Test set (n=188)	μ ± σ median t-test P	**.69 ± .08** **.71**	.57 ± .06 .57 6.2e-318	.64 ± .15 .65 7.3e-19	.59 ± .07 .60 6.2e-256	.67 ± .13 .67 1.0e-166

4 Discussion and Conclusion

Four radiologists exhaustively recorded observations concerning 17 features per bowel segment in 30 patients. We could train classifiers to predict a score with high correlation to the real CDEIS, with r=0.65 as highest correlation factor. Compared to the already published MaRIA score, this is an improvement of 18% on our dataset. Note that the MaRIA score was reported to have a much better performance (r=0.82, [8]). This difference might result from the MaRIA score being originally developed on data with more severe CD cases and higher CDEIS, which tend to have clearer

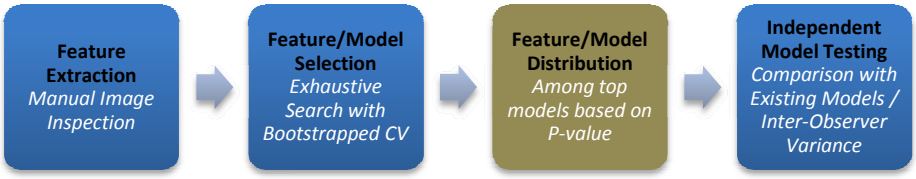

Fig. 5. The systematic model development pipeline as implemented in this paper. Applicable on various feature selection problems, we demonstrate the workflow on Crohn's disease data.

occurrences of the different features in the MRI scans. Also, the MaRIA patients experienced colonic preparation which results in better mucosal visualization at MRI. These effects might be due to a regional variation.

Our proposed feature selection workflow can principally be applied on various feature selection problems. For the CD problem, we implemented the development pipeline with following specifications (Fig. 5): In step 1, we use manually collected features by four domain experts. The ultimate vision of computational MRI processing might require automated feature extraction. However, to the best of our knowledge, no automated feature extraction for computational radiology tailored to CD severity assessment does exist. Our work reveals the importance of single features and, therefore, identifies targets that can be addressed for automated processing in further studies. In step 2, an exhaustive search among all 17 features has revealed that there is no clear prominent model solely suitable for CDEIS representation. Rather, we can have a set of models with different feature combinations but similar performance. This effect can at best been seen using a brute force algorithm instead of a heuristic approach, since in the first, really all models are considered. Further, this observation can often be made in analogue biomedical problems (e.g. [10]). The reasons for that might be manifold. First, there might be no evidence that the features do contain such a prominent relation to the examined target variable. Second, candidate features might indeed be related to the target, but not necessarily causally (directly), such that a mathematical model would only discover indirect (or "weaker") relations. These indirect relations are then discovered in an exhaustive search approach. Finally, from a machine learning point of view, implicit noise in real data, and subjectively measured features and labels always exacerbate the analyses of relations between the measurements. The pipeline respects this observation in step 3: based on the P-value, the set of top performing models is identified. Further, the feature distribution in this set supports the model selection process. Features that occur more often among the top models are assumed to have superior impact on the decision. A final independent validation step is proposed in this pipeline, importantly, since the result of the cross-validation has been used for model selection. Comparing the feature annotations of four different radiologists, our resulting model for CD severity quantification shows superior correlation performance to the state-of-the-art score than existing methods, with at the same time similar variance among different annotators.

Our proposed features and model for CD severity assessment should be further validated on larger datasets for clinical value. An interesting validation scenario would be the connection of MRI with histological data. While histology might serve as a better gold-standard than CDEIS, such data are far less available and often biased to severe cases where surgery was necessary. For future research, our feature ranking offers a basis for automated MRI feature extraction methods which focus on mimicking manual MRI features.

We have presented a systematic pipeline for feature and model selection in computational radiology and showed its benefits for the difficult problem of CD severity estimation based on MRI. With the pipeline, models could be generated with superior performance than existing score functions. We are convinced that this design of a medical imaging pipeline can easily be adapted to structurally similar problems in computational magnetic resonance imaging.

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

References

1. Rimola, J., Ordas, I., Rodriguez, S., Garcia-Bosch, O., Aceituno, M., Llach, J., Ayuso, C., Ricart, E., Panes, J.: Magnetic resonance imaging for evaluation of Crohn's disease: validation of parameters of severity and quantitative index of activity. Inflammatory Bowel Diseases 17, 1759–1768 (2011)
2. Steward, M.J., Punwani, S., Proctor, I., Adjei-Gyamfi, Y., Chatterjee, F., Bloom, S., Novelli, M., Halligan, S., Rodriguez-Justo, M., Taylor, S.A.: Non-perforating small bowel Crohn's disease assessed by MRI enterography: derivation and histopathological validation of an MR-based activity index. European Journal of Radiology 81, 2080–2088 (2012)
3. Ziech, M.L., Lavini, C., Caan, M.W., Nio, C.Y., Stokkers, P.C., Bipat, S., Ponsioen, C.Y., Nederveen, A.J., Stoker, J.: Dynamic contrast-enhanced MRI in patients with luminal Crohn's disease. European Journal of Radiology 81, 3019–3027 (2012)
4. Tielbeek, J.A.W., Makanyanga, J.C., Bipat, S., Pendsé, D.A., Yung Nio, C., Vos, F.M., Taylor, S.A., Stoker, J.: Grading Crohn's disease activity with MRI: Interobserver variability of MRI features, MRI scoring of severity and correlation with Crohn's Disease Endoscopic Index of Severity. AJR (2013)
5. Mahapatra, D., Schueffler, P., Tielbeek, J., Vos, F.M., Buhmann, J.M.: Crohn's Disease Tissue Segmentation from Abdominal MRI Using Semantic Information and Graph Cuts. In: Proc. IEEE ISBI 2013, San Francisco, pp. 358–361 (2013)
6. Mahapatra, D., Schueffler, P., Tielbeek, J.A.W., Buhmann, J.M., Vos, F.M.: A Supervised Learning Based Approach to Detect Crohn's Disease in Abdominal MR Volumes. In: Yoshida, H., Hawkes, D., Vannier, M.W. (eds.) Abdominal Imaging 2012. LNCS, vol. 7601, pp. 97–106. Springer, Heidelberg (2012)
7. Hastie, T., Tibshirani, R., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction. Springer, New York (2009)
8. Rimola, J., Rodriguez, S., Garcia-Bosch, O., Ordas, I., Ayala, E., Aceituno, M., Pellise, M., Ayuso, C., Ricart, E., Donoso, L., Panes, J.: Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. Gut 58, 1113–1120 (2009)
9. Bonferroni, C.E.: Il calcolo delle assicurazioni su gruppi di teste. In: Studi in Onore del Professore Salvatore Ortu Carboni, Rome, pp. 13–60 (1935)
10. Cima, I., Schiess, R., Wild, P., Kaelin, M., Schueffler, P., Lange, V., Picotti, P., Ossola, R., Templeton, A., Schubert, O., Fuchs, T., Leippold, T., Wyler, S., Zehetner, J., Jochum, W., Buhmann, J., Cerny, T., Moch, H., Gillessen, S., Aebersold, R., Krek, W.: Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. Proceedings of the National Academy of Sciences of the United States of America 108, 3342–3347 (2011)