SegClass-CD: Segmentation-Guided Region-Based Classification of Crohn's Disease Imaging Findings from Multi-Contrast MRE

Sarah Teitz[1], Naama Gavrielov[2], Leah Gitelman[2], Gili Focht[3], Ruth Cytter-Kuint[4], Talar Hagopian[4], Elena Vainberg[3], Israel Cohen[1], Dan Turner[3], and Moti Freiman[2]

[1] Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering, Technion – Israel Institute of Technology, Haifa, Israel

[2] Faculty of Biomedical Engineering, Technion – Israel Institute of Technology, Haifa, Israel

[3] Juliet Keidan Institute of Pediatric Gastroenterology, Hepatology and Nutrition, The Eisenberg R&D Authority, Shaare Zedek Medical Center, The Hebrew University School of Medicine, Jerusalem, Israel

[4] Pediatric Radiology Unit, Radiology Department, The Eisenberg R&D Authority, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel

Abstract. Crohn's disease (CD) is a chronic inflammatory bowel disease whose activity and complications can be assessed using magnetic resonance enterography (MRE). Existing MRE-based approaches for automated CD assessment either rely on manually extracted features or are limited to coarse disease detection rather than classification of specific radiological findings. Furthermore, most approaches utilize only a single MRE contrast, despite the complementary information available from multi-contrast imaging. Therefore, we introduce a model that extracts eight CD radiological findings using T1- and T2-weighted MRE without handcrafted features. Our model consists of contrast-specific feature extractors, masked Generalized Mean (GeM) pooling, and a modality-weighting module preceding the classification head. We use a pretrained supervised segmentation model to extract sub-organ intestinal segmentations and use them to guide region-based feature pooling from the feature maps. We evaluate our model through two experiments: First, we compare the multi-contrast model, which uses both T1 and T2 contrasts, with single-contrast models to assess the contributions of each contrast. Second, we compare different modality weightings, including linear, learned, and attention-based. In the first experiment, the multi-contrast model achieves a weighted AUPRC of 0.683±0.011, outperforming the single-contrast T1 or T2 models, which achieved 0.657±0.010 and 0.677±0.010, respectively. In the second experiment, the attention-based weighting produced the best performance with a weighted AUPRC of 0.691±0.011. These results demonstrate the value of segmentation-guided multi-contrast MRE for automated classification of CD imaging findings.

Keywords: Crohn's disease · multi-contrast MRE · radiological findings classification.

1 Introduction

Crohn's disease (CD) is a chronic inflammatory bowel disease that primarily affects the terminal ileum and colon [4, 5]. Inflammation may extend through the bowel wall into surrounding mesenteric tissues, leading to complications such as strictures, bowel obstruction, and fistulas [15].

Magnetic resonance enterography (MRE) is widely used for disease monitoring due to its non-invasive evaluation of the entire bowel and associated transmural and mesenteric tissues [4, 5, 16]. To assess disease activity and severity, clinicians use MRE to evaluate mural findings such as bowel wall thickening and enhancement, extramural findings including comb sign, mesenteric edema, and fat stranding, and complication indicators such as fistulas, stenosis, and pre-stenotic dilation throughout the rectum, colon, and ileum [2, 10, 12].

To support objective and scalable interpretation of these imaging findings, several methods have been proposed for automated CD assessment using MRE. Liu et al. employed support vector machines with radiomic features extracted from T2-weighted images to predict CD diagnosis [9]. Chirra et al. used random forests trained on radiomics extracted from T2-weighted images to detect CD inflammation and fibrosis [3]. Yu et al. developed radiomics-based models for classifying muscular alterations using both T1-weighted images and a combination of T1-weighted and dynamic contrast-enhanced (DCE) images [17]. Zhang et al. presented a multi-contrast radiomic framework for detecting intestinal fibrosis [18]. Lamash et al. developed a semi-automated model to extract wall thickness and lumen radius by manually annotating a centerline on T1-weighted images [8]. Mahapatra et al. classified CD from T1-weighted images using random forests applied to automatically extracted volumes of interest based on superpixels [11]. Shand et al. evaluated three models for automatic CD detection, two of which used only T1-weighted MRE, while the third combined T1-weighted and diffusion-weighted imaging [14].

Most of these methods, particularly radiomics-based approaches, rely on manual annotation and hand-crafted features, making them expensive, time-consuming, and labor-intensive. Methods that avoid manual feature extraction typically perform simplified disease classification and do not identify specific CD imaging findings. Furthermore, except Zhang et al. [18], and a limited subset of the models proposed by Shand et al. [14] and Yu et al. [17], these methods utilize only a single MRE contrast, despite the routine clinical use of multiple contrasts for comprehensive disease assessment. Due to the limited availability of annotated medical imaging data, these studies are generally developed and evaluated on small datasets, often containing fewer than 200 patients, which restricts model robustness and scalability.

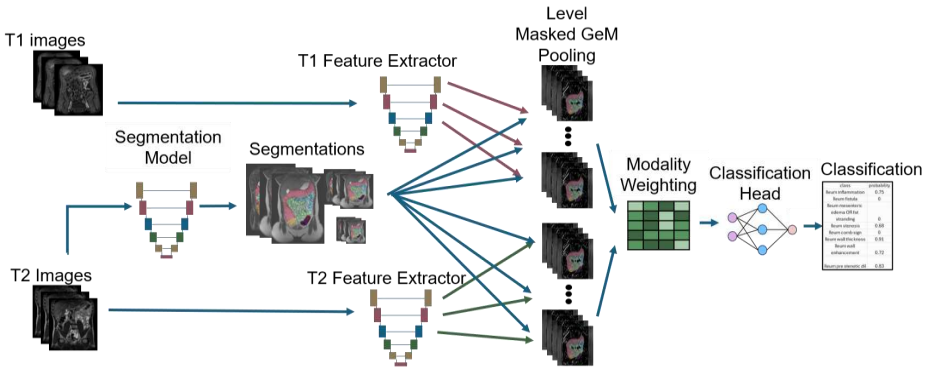


Fig. 1: Multi-contrast model. Segmentation is extracted using a pretrained model and T2-weighted images. T1- and T2-weighted images are processed by contrast-specific feature extractors. Feature maps are pooled using segmentation-guided, level-based, masked GeM pooling, fused using a modality-weighted mechanism, and passed to a multi-label classification head that predicts eight radiological Crohn's disease findings.

To address these limitations, there is a need for a fully automated, multi-contrast framework that detects specific CD imaging findings without manual annotation or handcrafted features.

Therefore, we propose an automated classification model that combines sub-organ intestinal segmentations with multi-contrast MRE to enable region-specific biomarker classification. Our key contributions are: 1) a fully automated pipeline for direct classification of radiological findings from imaging data without handcrafted features; 2) region-based biomarker classification enabled by sub-intestinal segmentations; 3) a multi-contrast feature integration that leverages complementary information from T1- and T2-weighted MRE;

2 Methodology

Our classification model first extracts contrast-specific features from T1- and T2-weighted MRE using pretrained feature extractors. To focus the model on the ileum, T2-derived segmentations are used to guide masked GeM pooling, producing fixed-length feature representations from the segmented regions. The pooled representations from both contrasts are then combined using a modality-weighting module before being passed to a multi-label classification head (Fig. 1).

2.1 Segmentation Extraction

T2-weighted ileum segmentations were extracted using a pretrained nnU-Net [7], initially trained on a publicly available MRE dataset [19] and subsequently fine-tuned on a small private dataset containing T2-weighted coronal masks of

the ileum, colon, sigmoid colon, rectum, jejunum, cecum, and ileocecal valve. As T1 segmentations were unavailable, these masks were used for both modalities.

2.2 Feature Extractors

T1- and T2-weighted images are fed into matching feature extractors, where each feature extractor is a pretrained nnU-Net. The T2 feature extractor is identical to the segmentation model used earlier, and the T1 feature extractor is pre-trained using the same protocol as the T2 feature extractor, with the same masks, except using T1-weighted images instead of T2-weighted images. During the training of the classification model, the feature extractors' decoders are fine-tuned to adapt to the downstream classification task, while the encoder remains frozen, to preserve contrast-specific representations. Output feature maps are extracted using forward hooks from the last 5 decoder layers, to leverage representation from multiple resolutions, without modifying the pretrained architecture.

2.3 Masked GeM Pooling

Each level of feature maps is paired with either the sample's segmentation or a downsampled version of the segmentation, so both share the same spatial resolution. Those are aggregated using a level-based masked Generalized Mean (GeM) pooling to create a fixed-length representation regardless of segmentation size. Unlike standard GeM pooling [13], which performs adaptive pooling over the entire feature map, masked GeM pooling restricts aggregation to the segmented region, combining GeM's learnable balance between average and max pooling with feature aggregation over the anatomically relevant region of interest for classification. For channel c, the pooled feature map is computed as:

$$PFM_c = \left(\frac{\sum_{\mathbf{u}} M(\mathbf{u}) x_c(\mathbf{u})^p}{\sum_{\mathbf{u}} M(\mathbf{u})} \right)^{\frac{1}{p}} \tag{1}$$

where $x_c(\mathbf{u})$ denotes the activation of channel c at spatial voxel $\mathbf{u}$, $M(\mathbf{u})$ is the corresponding segmentation mask value, and p is a learnable pooling parameter shared across all channels within the feature level.

2.4 Modality Weighting

Modality weighting is applied to the output of the GeM pooling due to discrepancies in the quality of features produced from T2 images versus those from T1 images, as the segmentations are limited to T2 contrast. We therefore experiment with four fusion strategies that determine each contrast contribution:

1. Linear Weighting: A weighted combination of the pooled feature maps (PFMs) is computed as: $(1 - \alpha) \cdot T1_{PFM} + \alpha \cdot T2_{PFM}$, where α is a learnable parameter. A separate weighting parameter is learned for each class and then passed to a two-layer classifier.

2. Learned weighting: The PFMs from each modality are first projected into a latent space of 256 using fully connected layers. The projected features are then processed by a two-layer MLP to generate modality-specific scores, which are normalized using a Softmax function. The resulting weights are applied to the projected PFMs to produce a class-specific representation for the single-layer classification head.
3. Attention-based weighting: The T1 PFMs and the T2 PFMs are fed into separate two-layer MLPs to project to a smaller-sized latent space of 128. The resulting embeddings are stacked and then fed into a two-headed self-attention module. Outputs from the attention are flattened and input to a shared single-layer classification head.
4. No modality weighting: As a baseline, no modality fusion weighting is applied. Instead, the pooled features from both modalities are directly provided to a classification head consisting of two fully connected layers.

2.5 Optimization and Training

We train the model for 15 epochs with a batch size of 8, using the AdamW optimizer and an ExponentialLR scheduler with $\gamma = 0.99$ and an initial learning rate of 0.0001. Loss is Binary Cross Entropy with label smoothing ($\epsilon = 0.1$) and per-class positive weights ($N_{\text{neg}}/N_{\text{pos}}$) to address label imbalance. Training data is augmented with random gamma, random bias field, and random noise, as in nnU-Net training protocols. Spatial augmentations were avoided to preserve disease characteristics. The model is trained using 5-fold cross-validation, with an 80% training and 20% validation split. The best model is determined by the highest weighted average precision on the validation set.

2.6 Single Contrast Model

The single contrast model is a leaner version of the multi-contrast model. There is only a single feature extractor, corresponding to the input image contrast. Additionally, modality weighting is removed, and the classification head is extended to two fully connected layers, with a dropout of p=0.2 between them.

3 Experiments and Results

3.1 Data and Preprocessing

We utilize data from a subset of the epi-IIRN study [6], comprising 4,562 coronal SSH-T2 scans and 4,562 coronal T1-weighted scans acquired during the second post-contrast phase of the mDIXON-W BH MRE sequence. Of those, 20 samples have T2-weighted segmentation masks for intestinal sub-organs, without corresponding T1-weighted masks. All scans in our data subset were acquired from a single Health Maintenance Organization among the 4 available ones, to ensure consistency with the domain on which the segmentation model was developed.

Table 1: Prevalence of ileal Crohn's disease findings in the dataset.

Class	Prevalence [%]	Positive Cases (n/N)
Wall Thickness	42.26	1928/4562
Wall Enhancement	39.30	1793/4562
Inflammation	32.88	1500/4562
Stenosis	22.73	1037/4562
Pre-Stenotic Dilation	16.68	761/4562
Comb Sign	12.41	566/4562
Fistula	6.44	294/4562
Mesenteric Edema / Fat stranding	5.00	228/4562

Of these, 51% have at least one biomarker indicating Crohn's disease, while 49% have no indications from MRE scans. Classification labels for each MRE sample are extracted from radiologist reports by using Badash et al.'s HSMP-BERT model, previously trained and validated on a subset of this cohort [1]. In total, 8 ileum labels are relevant across T1- and T2-weighted images, each with a frequency greater than 3%. Table 1 summarizes label prevalence. All images and segmentations were resampled to a uniform resolution of [5.5,1,1] mm, with variable spatial dimensions. To mitigate partial segmentations and the lack of T1 segmentation masks, dilation is applied to the segmentation masks using three iterations. Intestinal regions are cropped after segmentation and dilation. Images and segmentations are then zero-padded to multiples of [16, 128, 128] to fit the pretrained nnU-Net. Z normalization is applied after sampling and cropping.

3.2 Experimental Evaluation

We conduct two experiments. In the first, we compare the multi-contrast model without modality weighting against the results of the T1 single-contrast model and the T2 single-contrast model. The second experiment compares the different modality weightings within the multi-contrast model. We evaluate model performance using AUPRC, Cohen's κ, and precision for each label. To determine at which threshold Cohen's κ and precision should be used, we utilize Youden's Index.

Table 2 shows overall performance, while Table 3 shows per-label AUPRC for single-contrast and prominent multi-contrast models. Figure 2 shows results from the first experiment. The multi-contrast model achieves a higher weighted AUPRC of 0.683 ± 0.011 compared to the T1 and T2 single-contrast models, which have weighted AUPRCs of 0.657 ± 0.010 and 0.677 ± 0.01, respectively. In addition, the multi-contrast model obtains a similar or higher AUPRC for almost every label, suggesting that the multi-contrast model leverages complementary information from both imaging contrasts rather than relying primarily on the stronger individual contrast. Although Comb Sign, Wall Enhancement, and Mesenteric Edema/Fat Stranding are evaluated clinically by a single contrast [10], those labels achieved high scores in the complementary contrast model, likely because of high correlation with other detectable labels.

Table 2: Overall performance comparison of single-contrast (T1 and T2) and multi-contrast models, with and without attention-based weighting.

Model	Weighted AUPRC ↑	Micro AUPRC ↑	Micro PPV ↑	Micro κ ↑
Multi-contrast, attention-based	$\mathbf{0.691 \pm 0.011}$	$\mathbf{0.567 \pm 0.063}$	0.496 ± 0.023	$\mathbf{0.454 \pm 0.020}$
Multi-contrast, no mod. weighting	0.683 ± 0.011	0.549 ± 0.029	0.496 ± 0.014	$\mathbf{0.454 \pm 0.019}$
T1 contrast	0.657 ± 0.010	0.537 ± 0.043	$\mathbf{0.500 \pm 0.020}$	0.441 ± 0.014
T2 contrast	0.677 ± 0.010	0.529 ± 0.032	0.473 ± 0.020	0.429 ± 0.020

Table 3: Per-label AUPRC performance comparison of single-contrast (T1 and T2) and multi-contrast models, with and without attention-based weighting.

	AUPRC ↑			
Label	Multi-contrast attention-based	Multi-contrast no mod. weighting	T1 contrast	T2 contrast
Wall Thickness	$\mathbf{0.853 \pm 0.013}$	0.844 ± 0.014	0.826 ± 0.011	0.844 ± 0.018
Wall Enhancement	$\mathbf{0.813 \pm 0.016}$	0.805 ± 0.012	0.794 ± 0.018	0.800 ± 0.015
Inflammation	$\mathbf{0.678 \pm 0.024}$	0.676 ± 0.039	0.654 ± 0.035	0.664 ± 0.030
Stenosis	$\mathbf{0.628 \pm 0.036}$	0.620 ± 0.040	0.567 ± 0.019	0.617 ± 0.022
Pre-Stenotic Dilation	$\mathbf{0.588 \pm 0.051}$	0.564 ± 0.040	0.515 ± 0.037	0.576 ± 0.037
Comb Sign	$\mathbf{0.424 \pm 0.077}$	0.398 ± 0.053	0.388 ± 0.049	0.397 ± 0.056
Fistula	0.326 ± 0.052	$\mathbf{0.348 \pm 0.055}$	0.290 ± 0.030	0.273 ± 0.057
Mesenteric Edema /Fat Stranding	0.211 ± 0.068	$\mathbf{0.221 \pm 0.061}$	0.200 ± 0.078	0.202 ± 0.064

Figure 3 shows the results of the second experiment. Attention-based weighting achieves the best weighted AUPRC of 0.691±0.011 and the highest AUPRC for the labels: wall thickness, wall enhancement, pre-stenotic dilation, and comb sign. This improvement is likely due to self-attention, which enables interactions between T1 and T2 features before classification, reducing redundancy and enhancing feature representations.

Figure 4 shows feature attribution maps obtained from the multi-contrast model with attention-based weighting for representative patients with and without stenosis. In the stenosis-positive case, feature attributions are concentrated within a limited region of the ileum segmentation mask and appear to focus on areas of ileal narrowing or pre-stenotic dilation. In contrast, the stenosis-negative case exhibits attribution patterns that are either distributed throughout the segmentation mask or concentrated near its periphery.

4 Conclusion

We developed a fully automated multi-contrast pipeline for classifying eight radiological CD findings from real-world clinical MRE, eliminating the need for handcrafted features. To our knowledge, this is the first approach to do so.

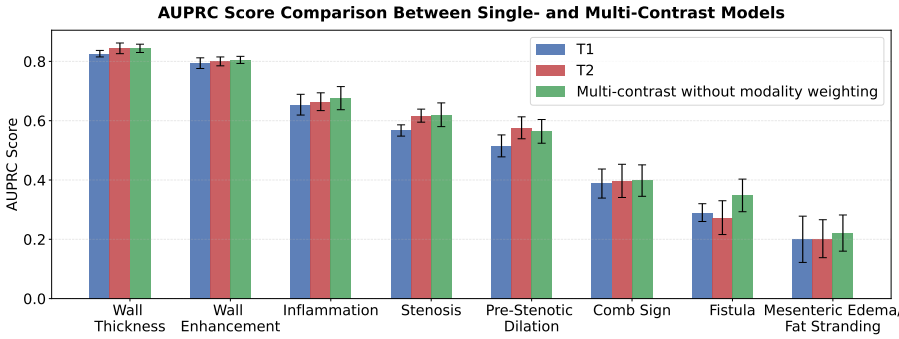


Fig. 2: AUPRC scores of the multi-contrast model without modality weighting vs. the single-contrast T1 model and single-contrast T2 model.

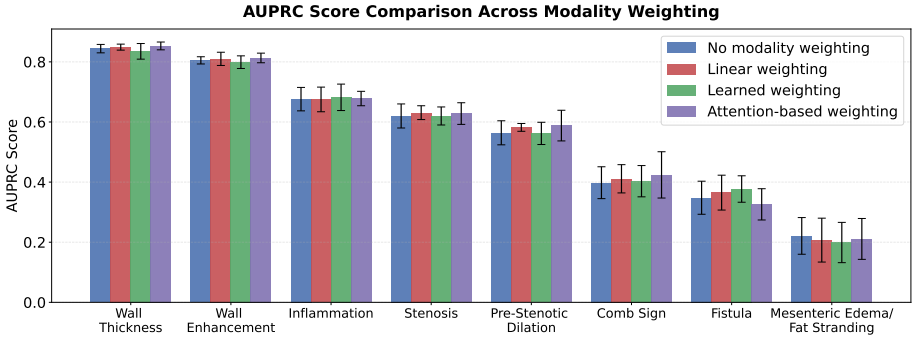
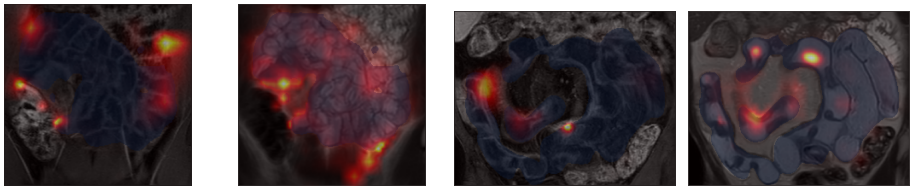


Fig. 3: AUPRC scores of the multi-contrast model with different modality weightings.



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

Fig. 4: Feature attribution maps for stenosis prediction. The ileum segmentation is overlaid in blue, while red–yellow regions indicate areas with high feature attribution. (a,b) T1-weighted and T2-weighted MR images from a patient without stenosis. (c,d) T1-weighted and T2-weighted MR images from a patient with stenosis.

Multi-contrast classification improved weighted and Micro AUPRC over either

contrast alone, with further improvement using attention-based weighting, supporting the benefit of integrating complementary T1 and T2 information.

Performance may be limited by severe class imbalance, correlated labels, and the lack of precise spatial correspondence between radiological findings and their locations within the scans. Class imbalance was addressed using BCE with per-class positive weights, while segmentation-guided pooling provides coarse anatomical localization. Performance remains highly dependent on segmentation quality, with mask dilation only partially mitigating segmentation inaccuracies and the lack of T1 segmentations. Future work will investigate finer anatomical localization and registration-based transfer of T2 segmentations to T1 images.

Acknowledgments The study was sponsored by the Leona M. and Harry B. Helmsley Charitable Trust. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosure of Interests The authors have no competing interests to declare.

Code https://github.com/SarahTeitz/SegClass-CD

References

1. Badash, Z., Ben-Atya, H., Gavrielov, N., Hazan, L., Focht, G., Cytter-Kuint, R., Hagopian, T., Turner, D., Freiman, M.: Hierarchical section matching prediction (hsmp) bert for fine-grained extraction of structured data from hebrew free-text radiology reports in crohn's disease (2025), https://arxiv.org/abs/2509.04519
2. Bane, O., Gee, M.S., Menys, A., Dillman, J.R., Taouli, B.: Emerging imaging biomarkers in crohn disease. Topics in Magnetic Resonance Imaging **30**(1), 31–41 (2021)
3. Chirra, P., Sleiman, J., Gandhi, N.S., Gordon, I.O., Hariri, M., Baker, M., Ottichilo, R., Bruining, D.H., Kurowski, J.A., Viswanath, S.E., et al.: Radiomics to detect inflammation and fibrosis on magnetic resonance enterography in stricturing crohn's disease. Journal of Crohn's and Colitis **18**(10), 1660–1671 (2024)
4. Dolinger, M., Torres, J., Vermeire, S.: Crohn's disease. The Lancet **403**(10432), 1177–1191 (2024)
5. Fiorino, G., Bonifacio, C., Malesci, A., Balzarini, L., Danese, S.: Mri in crohn's disease—current and future clinical applications. Nature Reviews Gastroenterology & Hepatology **9**(1), 23–31 (2012)
6. Friedman, M.Y., Leventer-Roberts, M., Rosenblum, J., Zigman, N., Goren, I., Mourad, V., Lederman, N., Cohen, N., Matz, E., Dushnitzky, D.Z., Borovsky, N., Hoshen, M.B., Focht, G., Avitzour, M., Shachar, Y., Chowers, Y., Eliakim, R., Ben-Horin, S., Odes, S., Schwartz, D., Dotan, I., Israeli, E., Levi, Z., Benchimol, E.I., Balicer, R.D., Turner, D.: Development and validation of novel algorithms to identify patients with inflammatory bowel diseases in israel: an epi-iirn group study. Clinical Epidemiology **10**, 671–681 (2018). https://doi.org/10.2147/CLEP.S151339, https://www.tandfonline.com/doi/abs/10.2147/CLEP.S151339, pMID: 29922093

7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
8. Lamash, Y., Kurugol, S., Warfield, S.K.: Semi-automated extraction of crohn's disease mr imaging markers using a 3d residual cnn with distance prior. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 218–226. Springer (2018)
9. Liu, R.X., Li, H., Towbin, A.J., Ata, N.A., Smith, E.A., Tkach, J.A., Denson, L.A., He, L., Dillman, J.R.: Machine learning diagnosis of small-bowel crohn disease using t2-weighted mri radiomic and clinical data. American Journal of Roentgenology **222**(1), e2329812 (2024)
10. Lo, R.W., Bhatnagar, G., Kutaiba, N., Srinivasan, A.R.: Evaluating luminal and post-operative crohn's disease activity on magnetic resonance enterography: A review of radiological disease activity scores. World Journal of Gastroenterology **31**, 107419 (2025). https://doi.org/10.3748/wjg.v31.i26.107419, https://www.wjgnet.com/1007-9327/full/v31/i26/107419.htm
11. Mahapatra, D., Schüffler, P.J., Tielbeek, J.A., Makanyanga, J.C., Stoker, J., Taylor, S.A., Vos, F.M., Buhmann, J.M.: Automatic detection and segmentation of crohn's disease tissues from abdominal mri. IEEE Transactions on Medical Imaging **32**(12), 2332–2347 (2013)
12. Minordi, L.M., Bevere, A., Papa, A., Larosa, L., Manfredi, R.: Ct and mri evaluations in crohn's complications: A guide for the radiologist. Academic Radiology **29**(8), 1206–1227 (2022). https://doi.org/10.1016/j.acra.2021.07.025, https://www.sciencedirect.com/science/article/pii/S1076633221003548
13. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP** (11 2017). https://doi.org/10.1109/TPAMI.2018.2846566
14. Shand, M., Manderfield, J., Singh, S., McLafferty, C., Sharma, Y., Sengupta, S., Fernandes, P., Koroulakis, D., Syed, S., Brown, D.E.: Identifying pediatric crohn's disease using deep learning to classify magnetic resonance enterography (mre) images. In: 2021 Systems and Information Engineering Design Symposium (SIEDS). pp. 1–6. IEEE (2021)
15. Tian, C., Ranasinghe, I.R., Hsu, R.: Crohn disease. StatPearls (2025)
16. Varyani, F., Samuel, S.: Can magnetic resonance enterography (mre) replace ileocolonoscopy for evaluating disease activity in crohn's disease? Best Practice & Research Clinical Gastroenterology **38**, 101621 (2019)
17. Yu, L., Cai, Y., Lin, S., Zhang, H., Yu, S.: Quantitative mri radiomics approach for evaluating muscular alteration in crohn disease: development of a machine learning-nomogram composite diagnostic tool. Abdominal Radiology **50**(11), 5069–5078 (2025). https://doi.org/10.1007/s00261-025-04896-x
18. Zhang, M., Zeng, Y., Fang, Z.n., Wang, Y.d., Zhang, R.n., Ye, Z., Cao, Q.h., Mao, R., Sun, C., Chen, Z.h., Huang, B., Li, X.h.: Mri radiomics enhances radiologists' ability for characterizing intestinal fibrosis in patients with crohn's disease. Insights into Imaging **15**(1), 165 (2024). https://doi.org/10.1186/s13244-024-01740-6
19. Zhong, Z., Huang, L., Feng, S.T., Lin, H., Wang, X., Lu, B., Cao, K., Li, X., Huang, B.: A comprehensive dataset of magnetic resonance enterography images with intestinal segment annotations. Scientific Data **12**(1), 425 (2025). https://doi.org/10.1038/s41597-025-04760-z