Automatic Detection of Bowel Disease with Residual Networks

Robert Holland[1], Uday Patel[2], Phillip Lung[2], Elisa Chotzoglou[1], and Bernhard Kainz[1(✉)]

[1] Department of Computing, Imperial College London, BioMedIA, London, UK
{robert.holland15,e.chotzoglou16,b.kainz}@imperial.ac.uk
[2] St. Mark' Radiology, London North West University Healthcare NHS Trust, London, UK
{udaypatel2,phillplung}@nhs.net

Abstract. Crohn's disease, one of two inflammatory bowel diseases (IBD), affects 200,000 people in the UK alone, or roughly one in every 500. We explore the feasibility of deep learning algorithms for identification of terminal ileal Crohn's disease in Magnetic Resonance Enterography images on a small dataset. We show that they provide comparable performance to the current clinical standard, the MaRIA score, while requiring only a fraction of the preparation and inference time. Moreover, bowels are subject to high variation between individuals due to the complex and free-moving anatomy. Thus we also explore the effect of difficulty of the classification at hand on performance. Finally, we employ soft attention mechanisms to amplify salient local features and add interpretability.

1 Introduction

1.1 Motivation

Most people suffering from Crohn's disease are younger than 35 and the cost of their treatment exceeds £500 million in the UK alone. Symptoms include inflammation of tissue anywhere along the gastrointestinal tract. However, it is most commonly found in the terminal ileum (where the small and large intestine meet). While there is no cure, early detection can vastly improve quality of life.

A successful algorithm would assist radiologists in more accurate diagnosis and follow-up of Crohn's disease. This would be of particular benefit to radiologists with limited experience of Crohn's disease imaging or who encounter patients with Crohn's disease uncommonly. Such an algorithm could also be used to triage patients so that severe cases can be reviewed more immediately, or to perform a secondary review to the radiologist and flag potentially missed cases.

I. Rekik et al. (Eds.): PRIME 2019, LNCS 11843, pp. 151–159, 2019.
https://doi.org/10.1007/978-3-030-32281-6_16

1.2 Study Outline and Contributions

Performance of classification tasks on the bowels is degraded by the intrinsic complexity and noise of the anatomy. While Crohn's disease can inflame the entire gastrointestinal (GI) tract, radiologists typically study the terminal ileum when making a diagnosis [1]. The first question we consider is the extent to which is it possible to classify IBD Crohn's disease from an MRI volume using vanilla deep learning methods. To establish this baseline, we first localise to the ROI using the patient-specific coordinates of the terminal ileum provided by a radiologist. We demonstrate that this semi-automatic technique performs comparably to the current standard for evaluating Crohn's with MRI, the MaRIA score [8], while requiring only a fraction of the preprocessing. We also explore how the difficulty and inflammation severity of a sample affects classification performance.

The assumption will then be dropped, such that we are forced to work only with population-specific knowledge, resulting in weaker localisation. Precision and recall degrade as the now fully-automatic algorithm encounters a worse signal-to-noise ratio (SNR). Finally, we show that in the absence of overfitting soft attention mechanisms [9] improve performance through amplification of salient local features.

2 Related Work

Currently there are no deep learning methods deployed in the clinic to assist diagnosis of Crohn's disease. Diagnosis is determined entirely by radiologists and clinical professionals who employ various *in vivo* and imaging techniques. Thus, our classification performance will be compared with the clinical standard, the MaRIA score. For their similarity in physical domain, we then review similar applications of deep learning to the abdomen.

2.1 Clinical Standards for Evaluating IBD

The first methods to standardise diagnosis of IBDs were endoscopic scoring systems, such as the Crohn's Disease Endoscopic Index of Severity (CDEIS). However, these incur practical issues; regular endoscopic examinations have several drawbacks related to '*invasiveness, procedure-related discomfort, risk of bowel perforation and relatively poor patient acceptance*' [8]. In fact, a meta-analysis of prospective studies has shown both MRE and CT to have a sensitivity and specificity of greater than 90% in diagnosing IBDs. To evaluate the MRI, radiologists visually examine the bowels slice by slice and look for high level features. Signs indicative of IBD include increase in T2 signal and thickness of the bowel walls. Rimola et al. [8] developed a scoring system, the MaRIA score, by first extracting these standardised imaging features through manual annotation by a radiologist and then fitting them in a regression model. MaRIA score was found to have a strong correlation with CDEIS. For the detection of disease activity it scored **0.81** for sensitivity and **0.89** for specificity.

Challenges in computing the MaRIA score include differentiation of diseased segments from those that are collapsed, variability of disease presentation and image degradation caused by motion [2]. Additionally, the aforementioned metrics used in the MaRIA score must be calculated by a radiologist in the terminal ileum, the transverse, ascending, descending and sigmoid colon and the rectum, which is a timely and costly procedure.

2.2 Machine Learning for the Automated Detection of IBD

Machine learning can automatically extract local features in the presence of noise, and combine them to make more complex decisions. Thus, it promises to automate the collection of low level features and, as we determine in this work, the diagnosis. Some attempt has been made to automate the collection of features specifically for calculation of the MaRIA score; in 2013 Schüffler et al. [7] used random forests to segment diseased bowels. However, this technique first requires a radiologist to indicate the section of diseased bowel to evaluate. Moreover at the time of the study it required one hour per patient. As far as we can see, there are no studies that use deep learning to directly diagnose IBD from imaging data. Moreover, there are comparatively few medical imaging challenges that focus on the abdomen (notably KiTS19 and CHAOS19) compared to other domains, and as far as we can see, none that regard IBD.

Typically, the medical imaging community has been more focused on tasks such as tumour, lesion and anatomical segmentation. This is evidenced in 'A Survey of Deep Learning in Medical Image Analysis' [5], detailing that '*Most papers on the abdomen aimed to localize and segment organs, mainly the liver, kidneys, bladder, and pancreas*'. A more recent review paper, 'An overview of deep learning in medical imaging focusing on MRI' [6], describes continued progress in segmentation, registration and image synthesis, but regarding diagnosis and prediction it advises to consult the list from the previous review [5] indicating that the main focus still lies in segmentation. Indeed, newer studies on the task of prediction and diagnosis concentrate on the brain, kidney, prostate and spine, but do so via segmentation rather than direction prediction.

Thus, it may be the case that the optimal method for diagnosing Crohn's IBD operates by first segmenting the terminal ileum. Abdominal segmentation has been attempted, though not including the terminal ileum [3]; dice scores were high for larger anatomy (e.g. liver at 95.3 ± 0.7) but significantly reduced for smaller anatomy similar in function to the terminal ileum (e.g. duodenum at 65.5 ± 8.9). Furthermore, they go on to describe the limitations of CNNs for inference in the bowels, commenting that '*It is very challenging for the CNN to learn stable representative features for the digestive organs because the appearances, shapes, and sizes of these organs are highly unstable from day to day depending on different food intake and digestion process*' [3].

To summarise, there are no studies making direct diagnosis of IBD using deep learning on images. Furthermore, there are also no learning algorithms since the random forests [7] diagnosing IBD from MR volumes. Segmentation is typically preferred to direct diagnosis due to the increased dimensionality of

the annotations. As such, we compare our baseline performance to the reported binary classification performance of the MaRIA score in classification of Crohn's disease.

3 Data

MRI data has been acquired on a Philips Achieva 1.5 T MR System with acquisition parameters as outlined in Table 1. Use of de-identified data has been consented by the local ethics committees.

Table 1. MRI acquisition parameters. Number of signal averages (NSA); Turbo spin echo (TSE)

Planes	Sequence	FOV [mm]	TR/ TE	Slice [mm]	Matrix	NSA	Time [s]
Axial	e-THRIVE (T1 FFE / TFE)	375	5.9/3.4	3	212×160	1	20.7×2
Coronal	Single shot TSE (T2 TSE)	375	554/120	3	300×213	1	21.1
Axial	Single shot TSE (T2 TSE)	375	587/120	3.5	304×255	1	22.3×2

The Crohn's MRI dataset is divided into healthy, mild, moderate and severe (with fistulation) terminal ileal inflammation. These represent severity levels 0, 1, 2 and 3 respectively which were originally calculated using the MaRIA score. As there are no terminal ileal ground-truth segmentations available, the only other annotation is the centroid coordinates of the terminal ileum.

Individuals are ranked by classification *difficulty*; an ordering determined by the radiologists who annotated the data. While we cannot formally describe how an MRI volume of a patient might be *difficult* to annotate, we can theorise that it means the symptoms of Crohn's disease are hard to spot or are borderline. These difficulties may correspond to those discussed in computing the MaRIA score discussed in Sect. 2.1. Indeed, we see that as the severity of inflammation decreases the difficulty increases in Table 2 (average difficulty is 35.0).

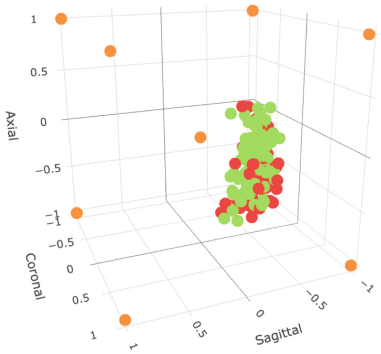
Table 2. Distribution of inflammation and suggested classification difficulty

Inflammation class	Frequency	Average difficulty
Healthy	100	N/A
Mild	34	39.1
Moderate	29	35.3
Severe	7	19.1

Formally, let $\{\mathbf{i}_j\}_{j=1}^N, \{\mathbf{d}_j\}_{j=1}^N$ such that $\forall j\ \mathbf{i}_j, \mathbf{d}_j \in \mathbb{R}^3$ be the set of physical locations of the terminal ilea and the dimensions of the j^{th} patient respectively. Then let the proportional ileal location be $\mathbf{p}_j = \frac{\mathbf{i}_j}{\mathbf{d}_j}$ and suppose that

$$\forall j\ \mathbf{p}_j \sim \mathcal{N}(\mu, \Sigma)$$

The distribution of $\{\mathbf{p}_j\}_{j=1}^N$ is shown in Fig. 1. Given that $\hat{\mu} < 0$ and $\hat{\Sigma}$ is small we observe that the terminal ileum is usually confined to one octant of the volume. From this distribution we can define a bounding box that we expect to contain all ilea. We make use of this assumption in preprocessing (see Sect. 3.1). We also observe from $\hat{\Sigma}$ that most variation is in the axial direction. This is expected as the method by which we determine the patient's size is most uncertain in this direction (patient dimensions were determined by region growing).



$$\hat{\mu} = \begin{bmatrix} -0.192 \\ -0.171 \\ -0.111 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 0.012 & -0.005 & -0.014 \\ -0.005 & 0.019 & 0.017 \\ -0.014 & 0.017 & 0.042 \end{bmatrix}$$

Fig. 1. Terminal ileal population distribution (normalised to [–1, 1])

3.1 Application Variants

We can localise to the ROI using the coordinates of the terminal ileum by extracting a small surrounding window, resulting in the *Localised* dataset. However, in the fully-automatic variant, we are forced to extract a larger region using the estimated distribution shown in Fig. 1 resulting in the *Generic* dataset. The effect of localisation strength on performance is detailed in Sect. 5. Localisation is crucial for mitigating overfitting, but also for permitting larger batch sizes, since the *Generic* and *Localised* techniques result in 95.8% and 99.4% volume reductions respectively.

4 Method

We are interested in the binary classification power of vanilla deep learning frameworks. As such, due to its efficient use of parameters, we chose ResNet [4]

- this affords us larger batch sizes, which are restricted by the dimensionality of the scans. Our custom Resnet uses exclusively 3^3 filters and ReLU activation. Refer to our network specification in Table 3. Each set of residual blocks, $\mathbf{d}_j$, begins with a downsampling layer via strided convolution. The residual blocks are followed by a classification module, comprising a global average pooling layer which allows us to feed inputs of variable size to the network. It also reduces the number of learnable parameters in the model as it is followed by a dropout fully connected layer resulting in two output neurons, as in binary classification.

We also add soft attention layers as described in Attention-gated Sononet by Schlemper et al. [9]. These act as a gate for signal by learning the compatability between pixel-wise features at a large scale and more global, discriminative features taken before the final soft-max layer. This is then normalised to form the attention map (see Fig. 2 for examples) and the dot product is taken with the pixel-wise features to produce attended features. These too pass through a classification module and their prediction is weighted against that of the original network's. To extend our custom Resnet we add an attention layer before the final downsampling layer. This multi-scale technique assists the network in identifying local, salient features such as the terminal ileum and is shown to improve performance in the absence of overfitting.

Table 3. Our ResNet configuration for input volume of size $31 \times 87 \times 87$

Layer	Channels	Blocks	Resultant feature map
$\mathbf{d}_1$	64	3	$16 \times 44 \times 44$
$\mathbf{d}_2$	128	3	$8 \times 22 \times 22$
$\mathbf{d}_3$	256	3	$4 \times 11 \times 11$
Global average pooling			256
Dense layer			2

4.1 Training and Evaluation

Loss is computed as cross entropy between the logits and the ground truth labels. We use Adam with $\beta_1 = 0.9, \beta_2 = 0.99$ for the first and second order moment coefficients respectively, and a learning rate of $5 \cdot 10^{-6}$. Due to the reduction in volume (see Sect. 3.1) we can use batch sizes of 64, a significant portion of the training set, which somewhat mitigates the intrinsic sample variability by producing more accurate gradient estimates.

Most deep learning frameworks were designed to train on vast datasets. Since we have many millions of parameters but relatively few samples, augmentation is necessary to artificially inflate the dataset. We capture variation present in anatomy and acquisition by including a mix of rotation (about the axial plane), horizontal flipping and random cropping.

All results are determined by four-fold cross validation on stratified training and testing sets, allowing us to evaluate the network on the entire dataset. The limited size of the dataset introduces an upper bound on overall binary classification accuracy of 92.45% (p-value 0.05), and just 84.8% on a single fold.

5 Results

Metrics were recorded when the loss was lowest for each fold. We will refer to Table 4, containing the results for the combined predictions over the whole dataset, as well as the best performing fold, and detailing the effect of the attention mechanism. It also compares the two levels of localisation that distinguish the *Localised* and *Generic* datasets.

Table 4. Best and average cross-fold binary classification performance for all application variants (formatted by precision/recall, and where A and H denote the abnormal and healthy classes respectively)

Attention		Generic region		Localised region	
		A	H	A	H
✗	Average	0.61/0.20	0.62/0.91	0.76/0.69	0.79/0.85
	Best	0.73/0.47	0.71/0.88	0.93/0.82	0.89/0.96
✓	Average	0.59/0.14	0.61/0.93	**0.79/0.80**	**0.86/0.85**
	Best	0.60/0.35	0.66/0.84	**0.94/0.94**	**0.96/0.96**

In most cases performance is reduced on the underrepresented class of abnormal patients. Moreover, performance is signficantly increased on localised data, and achieved best performance with attention mechanisms - this variant achieves weighted f-1 score **0.83**, demonstrating a strong correlation with the MaRIA score.

However, there is a large disparity between the best fold and the cross-fold average. In fact, the performance of any given fold was found to be highly dependent on the difficulty of the test set. Here we consider the difficulty of the abnormal samples only, assuming that healthy individuals present similar difficulty. We find that difficulty of the best fold was merely 31.3 while the worst was 42.3. Moreover, for the *Localised* variant with attention mechanisms, the average difficulty of incorrectly predicted abnormals was high, at 51.78, and of the seven severely inflamed individuals none were incorrectly classified (see Table 5). Classification power consistently increases with inflammation severity.

The limited size of the dataset introduced severe overfitting in training, forcing us to restrict the depth of the network and degrading overall performance. Furthermore, larger networks performed worse on the *Generic*, or population-specific, variant due to the reduction in SNR. This introduced difficulties in comparing variants on a standardised architecture.

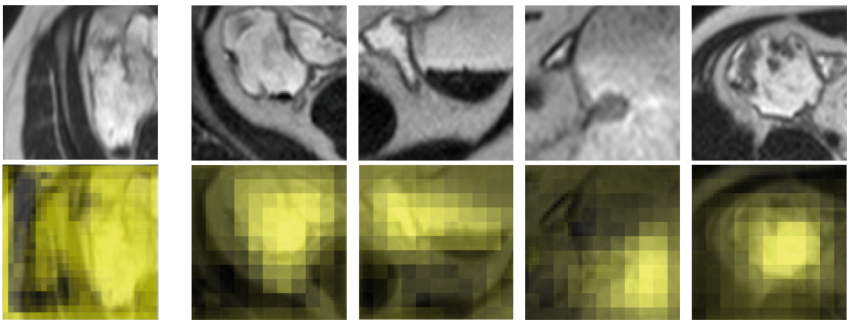
5.1 Attention

Attention mechanisms were found to exacerbate overfitting in scenarios with a low SNR but otherwise boosted performance. This can be seen by observing that attention boosts performance on the *Localised* dataset but degrades it

Table 5. Classification accuracy of best performing variant per inflammation class (for class support refer to Table 2)

Inflammation	Severe	Moderate	Mild	Healthy
Accuracy (%)	100.0	86.2	70.6	85.0

on *Generic*. We theorise that attention mechanisms can only become effective techniques to identify salient, local features within a network if the additional parameters they introduce are not accidentially misused for overfitting. There is evidence for this since the lowest cross entropy achieved by the best performing fold on the *Generic* dataset increased from 0.565 to 0.619 with the addition of attention mechanisms. Referring to Fig. 2, it also assisted us in *debugging* our network by highlighting that zero-padding allows the network to localise to regions that can be overfit on, such as bordering tissue (see Fig. 2a); mirror padding solves this issue. From Fig. 2b we deduce that the attention mechanism successfully identifies the relevant bowel section, reinforcing our confidence in the diagnoses.



(a) Zero-padding (b) Improved localisation with mirror-padding

Fig. 2. Attention maps on the *Localised* dataset (original slice and with attention overlayed on top and bottom rows respectively)

6 Discussion and Conclusion

In this work we demonstrated that a generic deep learning network, trained on a very small MRI dataset, correlates strongly to the MaRIA score, the current clinical standard, while requiring a fraction of the preprocessing by the radiologist. However, the framework is not without limitations in that performance is highly dependent on the level of localisation used in preprocessing and the difficulty rating of the classification at hand. Furthermore, the low dimensionality of

the output variable introduces statistical upper bounds on classification power and increases overfit. In this paper the evaluation criteria is solely based on expert radiologist assessment of the MRI data. Validation through colonoscopy is subject to future work, pending ethical approval.

Despite this, we observed very high classification power on the moderate to severely inflamed individuals, suggesting that this algorithm could provide secondary diagnoses to the radiologist in order to flag potentially missed cases. Overall, this pilot study highlights that deep learning is a very promising technique as a method for diagnosing disease in the bowels, and indicates that a larger dataset should continue to be collected for further evaluation. Finally, the limitations encountered through predicting low-dimensionality data might be alleviated by instead automating segmentation of the terminal ileum using deep learning, as a precursor to diagnosis. Thus, we recommend that terminal ileal ground-truth segmentations also be collected.

Acknowledgements. This research was kindly supported by Intel and hardware donations from Nvidia.

References

1. Chang, C.W., Wong, J.M., Tung, C.C., Shih, I.L., Wang, H.Y., Wei, S.C.: Intestinal stricture in Crohn's disease. Intest. Res. **13**(1), 19 (2015). https://doi.org/10.5217/ir.2015.13.1.19
2. Donagh, C., Walshe, T.M., Roche, C., Lohan, D., Cronin, C.G., Murphy, J.: Potential Pitfalls in MRI enterography-a pictorial review. Learning objectives (2012). https://doi.org/10.1594/ecr2012/C-2046, www.myESR.org
3. Fu, Y., et al.: A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. Med. Phys. **45**(11), 5129–5137 (2018). https://doi.org/10.1002/mp.13221
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**(December 2012), 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
6. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik **29**(2), 102–127 (2019)
7. Mahapatra, D., et al.: Automatic detection and segmentation of Crohn's disease tissues from abdominal MRI. IEEE Trans. Med. Imaging **32**(12), 2332–2347 (2013). https://doi.org/10.1109/TMI.2013.2282124
8. Rimola, J., et al.: Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. Gut **58**(8), 1113–1120 (2009). https://doi.org/10.1136/gut.2008.167957
9. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regionsin medical images. Med. Image Anal. **53**, 197–207 (2019). https://doi.org/10.1016/j.media.2019.01.012. http://www.sciencedirect.com/science/article/pii/S1361841518306133