Predictive modeling for Inflammatory Bowel Disease detection from endoscopic imaging

Nicolae Puica[1], Marco Chierici[2], Antonello Capistrano[3], Marcello Dorian Donzella[4], Antonio Colangelo[5], Venet Osmani[6], Giuseppe Jurman[†,7]

[1] Fondazione Bruno Kessler (FBK), Trento, I-38123, Italy [*]
nicolaerazvan.puica@gmail.com, https://orcid.org/0000-0002-5422-0573

[2] Fondazione Bruno Kessler (FBK), Trento, I-38123, Italy. chierici@fbk.eu,
https://orcid.org/0000-0001-9791-9301

[3] GPI S.p.A., Trento, I-38123, Italy. antonello.capistrano@gpi.it

[4] GPI S.p.A., Trento, I-38123, Italy. marcello.donzella@gpi.it,
https://orcid.org/0000-0002-6422-1641

[5] GPI S.p.A., Trento, I-38123, Italy. antonio.colangelo@gpi.it

[6] Fondazione Bruno Kessler (FBK), Trento, I-38123, Italy. vosmani@fbk.eu,
https://orcid.org/0000-0001-7306-2972

[7] Fondazione Bruno Kessler (FBK), Trento, I-38123, Italy. jurman@fbk.eu,
https://orcid.org/0000-0002-2705-5728

[†]corresponding author

Keywords: Artificial Intelligence, Inflammatory Bowel Disease, Endoscopy, Predictive Models, Diagnosis.

Abstract. The SI-CURA project (*Soluzioni Innovative per la gestione del paziente e il follow up terapeutico della Colite UlceRosA*) is an Italian initiative aimed at the development of Artificial Intelligence (AI) solutions to discriminate pathologies of different nature – Inflammatory Bowel Disease (IBD), Ulcerative Colitis (UC) and Crohn's Disease (CD) – based on endoscopic imaging of patients (P) and healthy controls (N). In particular, in this study a prototype deep learning (DL) framework is developed identifying disease patterns through three binary classification tasks, namely *P versus N*, *UC versus CD* and *UC versus N*. Starting from an imbalanced dataset of endoscopic imaging, the problem is tackled by training different ResNet architectures within a reproducible pipeline, eventually combined by means of an ensemble learning strategy. The models achieved high performance of MCC$>$0.95 on the test set for *P versus N* and *UC versus N*, and MCC$>$0.6 on the test set for *UC versus CD*, providing further evidence of their potential as a decision support tool for endoscopy-based diagnosis.

1 Scientific Background

Inflammatory bowel diseases (IBD), including Crohn's disease (CD) and ulcerative colitis (UC), are chronic and recurrent diseases. CD patients have healthy parts of the intestine mixed in between inflamed areas, while UC induces a continuous inflammation of the colon. Further, CD may occur in all the layers of the bowel walls, while UC only affects the innermost lining of the colon. Although they both have an undetermined etiology, research advances have outlined some of the pathways leading to their

[*]Current address: PagoPA S.p.A., Rome, I-00187, Italy.

insurgence: 1) genetic predisposition associated with the environment induces disruption of the intestinal microbial flora, 2) the structure of the epithelial cells and of the immune system of the intestine determine the risk of developing the disease. IBD is diagnosed using a combination of endoscopy (for CD) or colonoscopy (for UC) and imaging studies, such as contrast radiography, magnetic resonance imaging, or computed tomography. Physicians may also check stool samples to make sure symptoms are not being caused by an infection or run blood tests to help confirm the diagnosis. Still, a definite diagnosis of IBD remains a challenging task [1], often affected by subjective judgement [2]. Several automated approaches have been published in the recent literature [2, 3] attempting to provide computational support to improve the diagnostic task, with machine learning (ML) models playing a major role. In this scenario, it comes as no surprise the emergence of Artificial Intelligence (AI) solutions based on Deep Learning (DL) models, exploiting the potential of artificial neural networks to tackle the task starting from different data [4], also including endoscopic testing [5]. The current proposal naturally embeds in this research line, combining recent DL architectures with more classical ML strategies such as ensemble learning to further enhance the achieved performance. The promising results obtained can support the clinicians in providing a more objective and reliable diagnosis, thus reducing the risk of misidentification of CD and UC, an important aspect considering different treatment options and follow-ups of the two conditions [5].

2 Materials and Methods

2.1 *Data description and preprocessing*

The SI-CURA dataset includes 14,226 three-channel RGB endoscopic images of different sizes split between "Positive" patients (P: 11,404) and "Negative" healthy controls (N: 2,822). Positive samples are further labeled as Ulcerative Colitis (UC), Crohn's disease (CD), and Inflammatory bowel disease (IBD). Table 1 illustrates the sample stratification for each class and Figure 1 shows an example of positive and negative images.

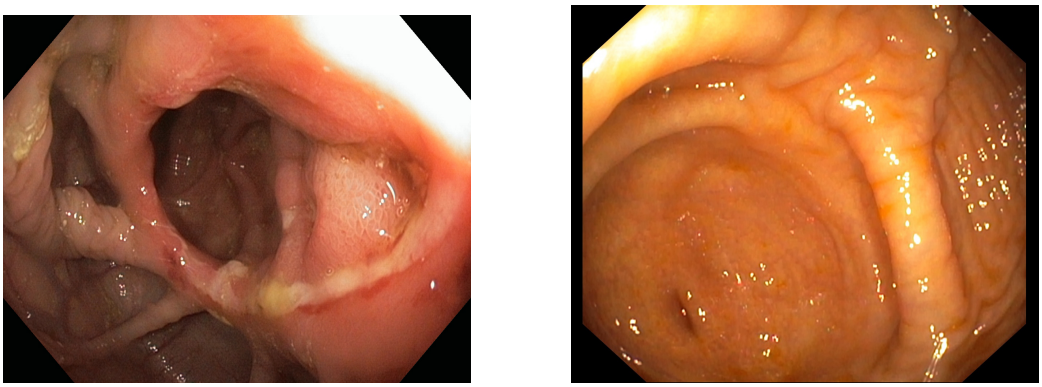
Table 1: Number of elements in each class, in the format "raw data (preprocessed data)". P, positive; N, negative; UC, Ulcerative Colitis; CD, Crohn's disease; IBD, inflammatory bowel disease.

P			N
UC	CD	IBD	
4,388 (3,594)	5,949 (4,098)	1,067 (823)	2,822 (2,815)

An image editor was used to remove undesired artifacts, such as signs, writings, medical instruments, black, white, and corrupted images. Further, all BMP images were converted to JPG with lossy compression to have uniform data formats across the dataset. Table 1 details the number of elements in each class after preprocessing.

For model development and evaluation, the full dataset was split into three main subsets: training (70% of total sample size), validation (20%), and test (10%). Since the P data is $3\times$ the N data, under- and over-sampling were used to balance the class distributions. Data augmentation is a technique commonly used to increase the amount of relevant data in the original dataset, thus providing the neural network with more examples. The following techniques were used: HSV format conversion, Random Horizontal Flip, and Random Vertical Flip, all of which had an equal probability of being applied.

Several experiments were performed with different batch and image sizes. The best results were obtained with a batch size of 32, 64, and 128, depending on the model's



(a) (b)

Figure 1: An example of positive (a) and negative (b) images within the SI-CURA dataset.

depth. To speed up the training time we opted for images of 700x700 pixels, equal to the average image dimension.

2.2 *Deep learning architecture*

We first focused on a binary classification task to discriminate healthy subjects from patients (N-P). Afterwards, we investigated the discrimination between Ulcerative Colitis versus Crohn's Disease (UC-CD) in order to learn deep features of specific IBD subgroups. Finally we focused on discriminating between Ulcerative Colitis versus Negative (UC-N).

We tackled the tasks through a transfer learning approach using the following variants of Residual Networks (ResNet) [6]: ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. The choice of using pre-trained models leverages the information already learned by the network, after an extensive training on the large ImageNet dataset. The head of the ResNet networks was swapped with the one performing binary classification, that is, an untrained sequential module that ends with a linear transformation with two outputs. Moreover, the rectified linear unit function (ReLU) [7] and dropout technique ($p = 0.5$) were used in this module: the former to avoid the vanishing gradient problem and the latter to increase regularization, preventing the coadaptation of the neurons. The weights of the layers were frozen except for the last two , which were updated with a small variation of the learning rate. The optimizer used for the training was Adam [8]. The actual number of epochs was determined by early stopping, monitoring the validation loss. The maximum number of epochs was set to 200, a value high enough to allow the net to learn as much as possible until the early stopping technique takes place. The learning rate was set based on the learning rate finder [9]. Only the layers immediately preceding the classifier were trained. Since the model is pre-trained on the large ImageNet dataset, these layers need only a small amount of fine-tuning to achieve a good prediction performance. In particular, the learning rate values for the `layer4` (just before the classifier) and `layer3` are reduced by a third and by a ninth, respectively.

Ensemble learning is a machine learning paradigm where multiple models ("weak learners") are trained to solve the same problem and finally combined to achieve better results. Two ensemble methods were used in this study: averaging prediction and stacking. While the former keeps the models separate and averages the predictions, the latter combines the predictions through a final classifier, resulting in a meta-model. This meta-model was further trained for several epochs by using the early stopping approach. As further discussed in the Results section, all N-P, UC-CD, and UC-N sample predictions benefit the most from the ensemble method. The final classifier has two input features for each weak learner (six features overall) and two outputs features. Specifically, ResNet34, ResNet50, and ResNet101 were used in the N-P and UC-N cases, and

ResNet34, ResNet50, and ResNet152 in the UC-CD case. Other combinations were tried, but they performed worse.

The metrics used for assessing the model performance in both training and evaluation stages are Matthews Correlation Coefficient (MCC), true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), and negative predictive value (NPV). The MCC was used as the main metric because it is particularly suitable for binary and multiclass classification [10] and it is generally regarded as a balanced performance measure that can be used even if the classes are of very different sizes. We also computed the training and validation losses and the 95% studentized bootstrap confidence intervals for the training MCC.

The following frameworks and libraries were used: PyTorch v1.5.0 with torchvision v0.6.0a0, CUDA 10.2, Jupyter-notebook v6.1.4, and Imbalanced-dataset-sampler library [1]. The packages were installed in a Anaconda virtual environment for improved reproducibility. The analyses were run on a ppc64le server with 128 CPUs (max 4023 MHz, min 2394 MHz) and 4 Tesla P100-SXM2-16GB GPUs.

3 Results

Table 2 shows the average training MCC with 95% confidence intervals for all classification tasks, as well as the test set MCC, TPR, TNR, PPV, and NPV. The results are ranked task-wise by the test set MCC. The table reports the results of the meta-models obtained by the stacking ensemble method: the average-prediction ensemble models performed consistently worse for all tasks and are not shown. According to the table, the top three nets (or weak learners) for N-P are ResNet50, ResNet34, and ResNet101. While the first one is trained with an unbalanced training data loader, the latter two nets are trained using a balanced data loader.

Table 2: Metrics values for the best trained nets, ranked by decreasing test set Matthews Correlation Coefficient (MCC); TNR, true negative rate; TPR: true positive rate; NPV, negative predictive value; PPV, positive predictive value; tr, ts: training and test set; resnetX-Y-Z: meta-model obtained by stacking ensemble of the three weak learners resnetX, resnetY, and resnetZ.

Task	Net	MCC_{tr}	MCC_{ts}	TNR_{ts}	TPR_{ts}	NPV_{ts}	PPV_{ts}
N-P	resnet34-50-101	0.979 (0.979,0.980)	0.959	1.000	0.979	0.940	1.000
	resnet50	0.964 (0.961,0.968)	0.959	0.996	0.980	0.943	0.999
	resnet34	0.959 (0.955,0.963)	0.957	0.996	0.979	0.940	0.999
	resnet101	0.963 (0.960,0.966)	0.957	0.996	0.979	0.940	0.999
UC-CD	resnet34-50-152	0.570 (0.570,0.570)	0.612	0.819	0.795	0.778	0.833
	resnet34	0.513 (0.493,0.535)	0.600	0.819	0.782	0.768	0.831
	resnet152	0.366 (0.347,0.386)	0.582	0.894	0.680	0.710	0.880
	resnet50	0.508 (0.491,0.526)	0.573	0.699	0.863	0.818	0.766
UC-N	resnet34-50-101	0.971 (0.965,0.976)	0.953	0.964	0.993	0.994	0.955
	resnet50	0.958 (0.952,0.965)	0.947	0.953	1.000	1.000	0.943
	resnet34	0.951 (0.938,0.967)	0.944	0.950	1.000	1.000	0.940
	resnet101	0.951 (0.943,0.961)	0.943	0.964	0.982	0.986	0.955

As for the UC-CD task, the best results were obtained by ResNet34, ResNet50, and ResNet152 (Table 2). The first two nets were trained with a balanced data loader and the last one with an unbalanced data loader. In this task, the classes are not as imbalanced as in N-P (Table 1): however, the dataset balancing technique helped maintain balanced the

[1] https://github.com/ufoym/imbalanced-dataset-sampler

confusion matrices in the training and evaluation phase. As expected, the meta-model improves over the weak learners on the test set (Table 2).

The best-performing network for the UC-N task is an ensemble model using ResNet50, ResNet34, and ResNet101 as weak learners trained with unbalanced, balanced, and unbalanced data loaders, respectively. The best metrics are shown in Table 2.

4 Conclusions

A prototype DL framework is developed, based on ResNet architectures merged by ensemble learning, able to identify disease patterns from endoscopic images in the context of different IBDs, namely Ulcerative Colitis and Crohn's Disease. The DL models achieve a test set performance MCC=0.959 for the classification of healthy controls versus IBD patients, MCC=0.612 for Ulcerative Colitis versus Crohn's Disease, and MCC=0.953 for Ulcerative Colitis versus healthy individuals. Overall, the results reached in all the three classification problems indicate a very good to excellent predictive performance, highlighting the potential of the framework to evolve into a valuable tool for the clinicians in their diagnostic tasks. Nonetheless, despite the encouraging results, the current study should be considered as a proof of concept rather than a consolidated pipeline, and as such there is large room for improvement, that we have already planned as future developments. For instance, model architectures other than ResNet could be evaluated, and different loss functions can be used to overcome the data limitation involving unbalanced classes. Moreover, the nets could be trained for more epochs and/or more combinations of hyperparameters could be evaluated. Furthermore, alternative ensemble models forming different combinations can be tested, even trained with images in different color spaces. The model outcome can also be improved by enhancing the training data, even through augmenting techniques creating synthetic data: for example, generative adversarial networks. Finally, strengthening the resampling strategy will further improve the overall reproducibility of the study, while the analysis of the data trajectories across the DL layers can provide valuable hints regarding the model interpretability.

Acknowledgments

This work was supported by GPI S.p.A. through the Contract 210202009864 between GPI S.p.A. and FBK within the framework Project "*SI-CURA – Soluzioni Innovative per la gestione del paziente e il follow up terapeutico della Colite UlceRosA*" supported by the POR Puglia FESR-FSE 2014-2020 Innonetwork.

References

[1] L. Negreanu, T. Voiosu, M. State, A. Voiosu, A. Bengus, and B. R. Mateescu, "Endoscopy in inflammatory bowel disease: from guidelines to real life," *Therapeutic Advances in Gastroenterology*, vol. 12, 2019.

[2] G. Chen and J. Shen, "Artificial Intelligence Enhances Studies on Inflammatory Bowel Disease," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 570, 2021.

[3] G. E. Tontini, A. Rimondi, M. Vernero, H. Neumann, M. Vecchi, C. Bezzio, and F. Cavallaro, "Artificial intelligence in gastrointestinal endoscopy for inflammatory bowel disease: a systematic review and new horizons," *Therapeutic Advances in Gastroenterology*, vol. 14, 2021.

[4] S. Cohen-Mekelburg, S. Berry, R. W. Stidham, J. Zhu, and A. K. Waljee, "Clinical applications of artificial intelligence and machine learning-based methods in inflammatory bowel disease," *Journal of Gastroenterology and Hepatology*, vol. 36, no. 2, pp. 279–285, 2021.

[5] S. Sundaram, T. Choden, M. C. Mattar, S. Desai, and M. Desai, "Artificial intelligence in inflammatory bowel disease endoscopy: current landscape and the road ahead," *Therapeutic Advances in Gastrointestinal Endoscopy*, vol. 14, 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, 2016.

[7] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolutional Network." arXiv:1505.00853, 2015.

[8] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv:1412.6980, Proceedings of the 3rd International Conference for Learning Representations (ICLR), 2015.

[9] L. N. Smith, "Cyclical Learning Rates for Training Neural Networks," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, IEEE, 2017.

[10] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, 2020.