Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network

Astrid de Maissin, Remi Vallée, Mathurin Flamant, Marie Fondain-Bossiere, Catherine Le Berre, Antoine Coutrot, Nicolas Normand, Harold Mouchère, Sandrine Coudol, Caroline Trang, et al.

Published online: 2021-06-21

Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network

OPEN ACCESS

Authors

Astrid de Maissin[1], Remi Vallée[2], Mathurin Flamant[3], Marie Fondain-Bossiere[4], Catherine Le Berre[4], Antoine Coutrot[2], Nicolas Normand[2], Harold Mouchère[2], Sandrine Coudol[5], Caroline Trang[4], Arnaud Bourreille[4]

Institutions

1 CHD La Roche Sur Yon, department of gastroenterology, La Roche Sur Yon, France
2 Nantes University, CNRS, LS2N UMR 6004, Nantes, France
3 Clinique Jules Verne, department of gastroenterology, Nantes, France
4 CHU Nantes, Institut des Maladies de l'Appareil Digestif, CIC Inserm 1413, Nantes University, Nantes, France
5 CHU de Nantes, INSERM CIC 1413, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, Nantes, France

submitted 25.11.2020
accepted after revision 4.3.2021

Bibliography
Endosc Int Open 2021; 09: E1136–E1144
DOI 10.1055/a-1468-3964
ISSN 2364-3722

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Corresponding author
Arnaud Bourreille, Institut des maladies de l'Appareil Digestif, CHU Nantes, 1, place A Ricordeau, F-44093 Nantes, France
Arnaud.bourreille@chu-nantes.fr

Supplementary material is available under https://doi.org/10.1055/a-1468-3964

ABSTRACT

Background and study aims Computer-aided diagnostic tools using deep neural networks are efficient for detection of lesions in endoscopy but require a huge number of images. The impact of the quality of annotation has not been tested yet. Here we describe a multi-expert annotated dataset of images extracted from capsules from Crohn's disease patients and the impact of the quality of annotations on the accuracy of a recurrent attention neural network.

Methods Images of capsule were annotated by a reader first and then reviewed by three experts in inflammatory bowel disease. Concordance analysis between experts was evaluated by Fleiss' kappa and all the discordant images were, again, read by all the endoscopists to obtain a consensus annotation. A recurrent attention neural network developed for the study was tested before and after the consensus annotation. Available neural networks (ResNet and VGGNet) were also tested under the same conditions.

Results The final dataset included 3498 images with 2124 non-pathological (60.7 %), 1360 pathological (38.9 %), and 14 (0.4 %) inconclusive. Agreement of the experts was good for distinguishing pathological and non-pathological images with a kappa of 0.79 ($P<0.0001$). The accuracy of our classifier and the available neural networks increased after the consensus annotation with a precision of 93.7 %, sensitivity of 93 %, and specificity of 95 %.

Conclusions The accuracy of the neural network increased with improved annotations, suggesting that the number of images needed for the development of these systems could be diminished using a well-designed dataset.

Introduction

Since its advent in the late 1990 s, wireless capsule endoscopy has become an irreplaceable tool for the exploration of the small bowel (SB), and subsequently, for the entire gastrointestinal tract [1]. Small bowel capsule endoscopy (SBCE) allows complete noninvasive exploration of the SB and its add-on value for the detection of mucosal lesions compared to conven-

tional endoscopy and cross-sectional radiologic imaging has been clearly demonstrated in many pathologies.

Over time, SBCE has proven to be a consistent tool for detection and evaluation of the severity of Crohn's disease (CD) lesions located in the SB [2–5]. Although the performance and acceptance of SBCE have been demonstrated [6], its use has not yet become standardized among the gastroenterology community for management of the patients with CD for a number of reasons.

One of the main limitations is the time and expertise needed for a reading by gastroenterologists. Indeed, the capsule captures around 2 to 6 frames per second, depending on the SB peristalsis, and produces 50 000 to 60 000 images for each video. On average, SBCE analysis requires 30 to 60 minutes to be read, with an inherent risk of missed lesions.

Therefore, the development of computer-aided diagnosis (CAD) tools for automatic detection of lesions is highly desirable. With the development of artificial intelligence, CAD tools are increasingly being used to assist physicians in interpreting medical images in many domains, including gastroenterology [7, 8]. In the field of endoscopy, CAD tools have been tested mainly for polyp detection and characterization, and using SBCE for detection and the characterization of polyps as well as vascular lesions [9, 10]. CAD tools also have been tested less frequently for detection of SB mucosal lesions in patients with CD. Various systems have been developed to limit the time required for gastroenterologists to analyze SBCE images. These systems can be classified into two main categories: algorithms based on extraction of well-identified features based on support vector machine and algorithms based on deep neural networks. The latter results in better performance in classification but requires more data for proper training. These algorithms currently are evaluated using different datasets containing a variable proportion of "pathological" and "non-pathological" images. Frequently, details concerning the modalities to attribute the "real truth" to each image remain unclear. Sometimes, no detail is available on the process used to classify the images [11–13]. Most of the time, the images are selected by gastroenterology fellows supervised by a capsule expert [14, 15] or by only one expert in capsule endoscopy [16]. When the annotation by two gastroenterologists is discordant, the final decision is accorded by a third one [17]. Only in the CAD-CAP dataset, images were reviewed by a group of experts after their selection by gastroenterology fellows but the inter-observer concordance was not detailed [15]. Moreover, datasets contain mainly images of ulcers and erosions and do not reflect the great variety of mucosal lesions visible in the SB of patients with CD, such as stenosis, edema, erythema. and mucosal break [11–14, 17]. The interobserver variability for annotation of capsule images in CD patients is unknown and the impact of possible errors of annotation on the performance of CAD tools has never been evaluated.

The objectives of this study were first, to elaborate a dataset of carefully annotated images of SBCE extracted from patients with CD with a precise description of the process; second, to make it available for free to the scientific community to train their own CAD tools; and third, to evaluate the increased accuracy of CAD algorithms based on deep learning at each step of creation of the dataset.

Materials and methods

Dataset

The CrohnIPI dataset is a multicentric dataset approved by the French Data Protection Authority and by the "groupe nantais d'éthique dans le domaine de la santé," the Ethics Committee of Nantes, France. Three French endoscopic units participated in the dataset. Four readers were involved in annotation of images: one gastroenterology fellow and three experts. The first reader, experienced in conventional endoscopy, had been trained to read capsule videos for the purpose of the study and was responsible for selection of images. The three experts had more than 10 years of experience in management of patients with inflammatory bowel disease (IBD) and had read more than 200 SBCE in patients with CD. Three annotation rounds were carried out to obtain a consensus annotation as close as possible to the "real truth." A research group specializing in computerized analysis of medical images was responsible for development of the convolutional recurrent attention neural network.

Data-collection and definition of the lesions

Third-generation SBCE videos (Pillcam SB3 system, Medtronic, Minnesota, United States) acquired between 2014 and 2018 from patients with CD or who had undergone exploration for suspicion of CD and registered in the three participating endoscopy units were retrospectively collected and de-identified. Only videos of patients with CD who had evidence of presence of SB lesions were selected. The successive still frames, in JPEG format, were extracted from the videos (without loss of quality), allowing their annotation. Clinical and demographic data at the time of the capsule were registered.

Pathological lesions were defined as follows [18]:

- Erythema: Area of reddish villi.
- Edema: Enlarged/swollen/engorged villi.
- Aphthoid ulceration: Diminutive loss of epithelial layering with a whitish center and a red halo, surrounded by normal mucosa.
- Ulceration: Mildly depressed or frankly deep loss of tissue with a whitish bottom compared to the surrounding swollen/edematous mucosa. Ulcerations were classified according to their largest diameter between 3 and 10 mm or > 10 mm.
- Stenosis: Narrowing of the intestinal lumen withholding or delaying the passing of the capsule.

Selection of images and first round of annotation

Frames were selected and annotated by the initial reader (AM). All the frames selected by the reader, pathological or not, were considered as the frames of interest. Each frame of interest was extracted and included in the CrohnIPI dataset. When a frame contained more than one lesion, all the lesions were annotated. Images with a questionable lesion (type or presence) were labeled as "inconclusive." Pathological images were selected regardless of the position of the lesion into the frame.

The same proportion of non-pathological control frames, containing no lesion, was extracted from the same set of videos. Control frames were randomly and not successively selected independent of the presence of bubbles, residue, or turbid liquid to avoid any bias of learning.

Second round of annotation

All the frames selected by the initial reader, non-pathological and pathological, were reviewed by three experts in IBD and capsule endoscopy (AB, MF, CT). All the images were assigned to each reader in random order. Experts were blind to the other experts' annotations and had access only to still frames.

The same definitions of the lesions were used to annotate the frames. When the frames contained more than one lesion, all the lesions were registered by the readers.

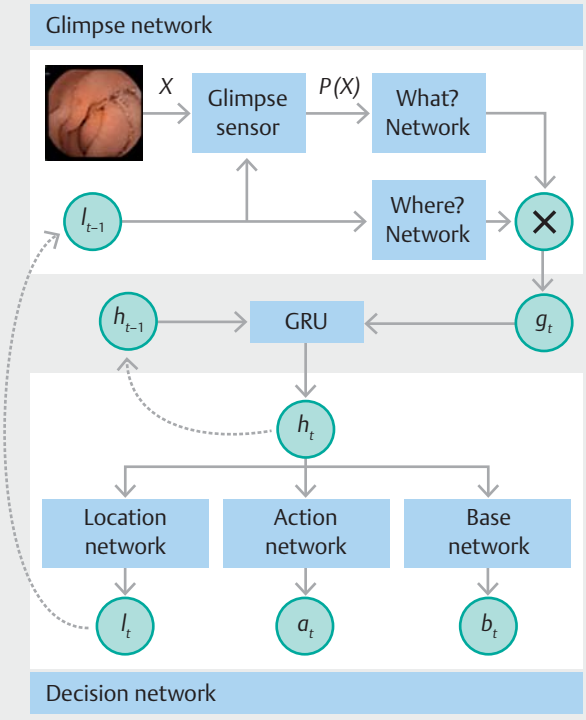
At the end of the process, annotations obtained from the initial reader and the experts were re-encoded to keep only one annotation per image. If several lesions were annotated on one frame, only the most severe one was retained. The severity of lesions was graded from the most to the least severe as follows: stenosis, ulceration > 10 mm, ulceration between 3 and 10 mm, aphthoid ulceration, edema, erythema [19]. Following this, concordance analysis between the experts was calculated and discordant images between the four readers were identified and registered for further analysis.

Third round of annotation

All the frames with a discordant annotation were reviewed. At this phase, only one lesion per frame was labeled, the most severe. The four readers met in three sessions to obtain a consensus annotation of the frames, considered as the "real truth" for each image. To prevent that one reader influenced the others, each of the readers had to give their opinion first, on 20 successive images, in the presence of an independent investigator not involved in reading the images. If necessary, a short adjacent video sequence could be retrieved, including 10 frames upstream and 10 frames downstream from the index frame. At the end of this process, discordant frames were classified as inconclusive and were excluded from the dataset for further analysis.

Recurrent attention neural network

The global architecture of the CrohnIPI network is described in ▶ **Fig. 1**. The proposed architecture [20, 21] mimics human visual attention. Humans sample the incoming visual information by moving their eyes on the most relevant targets. Similarly, this network samples relevant areas in the image thanks to a sequential attention mechanism. This network is a compound of three distinct parts: the Glimpse Network that extracts a small area of interest in the image, the Decision Network gives a first decision and proposes a new position of interest, and the last element is the gated recurrent unit (GRU) that propagates this information looping over four repetitions (patch – decision – region proposal). An endoscopic image X was provided as input to the network. The Glimpse Sensor then extracted a patch p(X) from the original image according to $l_t = (x, y, z)$ where x and y are the coordinates, initialized at the center of the image during the first loop, and z was a zoom coefficient. This patch was then resized to keep a fixed size at the network input.



▶ **Fig. 1** Global architecture of the attention recurrent network. At each time *t*, we provide the Glimpse sensor with an endoscopic image X and the location *lt*−1 of the patch to extract from the original image. Two independent neural networks, the What? Network and the Where? Network, will then extract information related to the content and location of the patch. A gated recurrent unit (GRU) will then merge the characteristics previously extracted by the network to produce the current system state *ht*. From this state, three sub-networks will independently produce *lt*, the position of the next patch to extract, *at*, a vector containing a score associated with each class and *bt*, the baseline from which is calculated the reward for reinforcement learning.

The subnetwork What? Network, based on VGG16, pretrained on ImageNet, allowed to extract the characteristics of the patch p(X) [22]. Only the 12 first layers of VGG16 and the first fully-connected layer were preserved. In parallel, localization information l_{t-1} goes through the Where? Network composed of 2 fully-connected layers, thus allowing the extraction of the characteristics relative to the position of the patch p(X). The two characteristic vectors produced by this network were combined in a new characteristic vector gt at the output of the non-linearity then contained the "Where?" and "What?" information extracted by the Glimpse Network at time *t*. A GRU allowed to merge the characteristics extracted at time *t* by the network with those extracted at the previous time contained in the previous internal state ht-1 of the GRU [23]. This internal state of the GRU was reused at the next time step.

From the new internal state produced by the GRU, the Action network produced a vector associating a score to each class of lesion. The Baseline Network allowed it to calculate

the reward associated with each prediction to allow the reinforcement learning. This reward permitted evaluation of the results of all decisions and back-propagation of this error through the network to train the localization in an unsupervised way. Thus, the network increased the probability of locations maximizing the reward function. If the network classified the image correctly, the reward was worth the number of views placed on the image minus the sum of baselines calculated by the network. A cross-entropy loss function was also used in addition of the reward function to help network convergence.

Statistical analysis

Analyses were performed using statistical R Project software (version 3.5.3). For all statistical analyses, $\alpha = 0.05$ was considered as an acceptable threshold for type I error.

Continuous variables are described using means ± standard deviations or median (25^{th} to 75^{th} percentile) and interquartile range. Categorical variables are described as raw counts and percentages. Whether after the first, second, or third round of annotations of lesions, there was no missing data and all readers annotated all the frames.

Interobserver agreement for classification of images in the CrohnIPI dataset was evaluated using the inter-rater agreement Fleiss' kappa, with *P* value [24]. Kappa value ranged from –1 to 1, with 0 value indicating statistical independence and 1 value indicating perfect agreement between observers. A Fleiss' kappa between 0.41 and 0.60 can be construed as a moderate agreement, between 0.61 and 0.80 as a substantial agreement and above 0.81 as a perfect agreement.

Images from the CrohnIPI dataset were randomly distributed into three sets, with 70 % assigned to the learning phase, 10 % to the validation phase, and 20 % to the test phase. This last phase was for final evaluation of the algorithm performance. It is known that an unbalanced repartition of images over the classes can influence the algorithm's generalization capacity, and thus, its performance in real conditions. To avoid this phenomenon during the evaluation, a five-fold cross-validation was performed, involving five repeated measurements of the algorithm performance, each time using a different split in 70 %, 10 %, and 20 % of the images into the three groups and ensuring that each example had been used exactly once in the algorithm's test base for each measure.

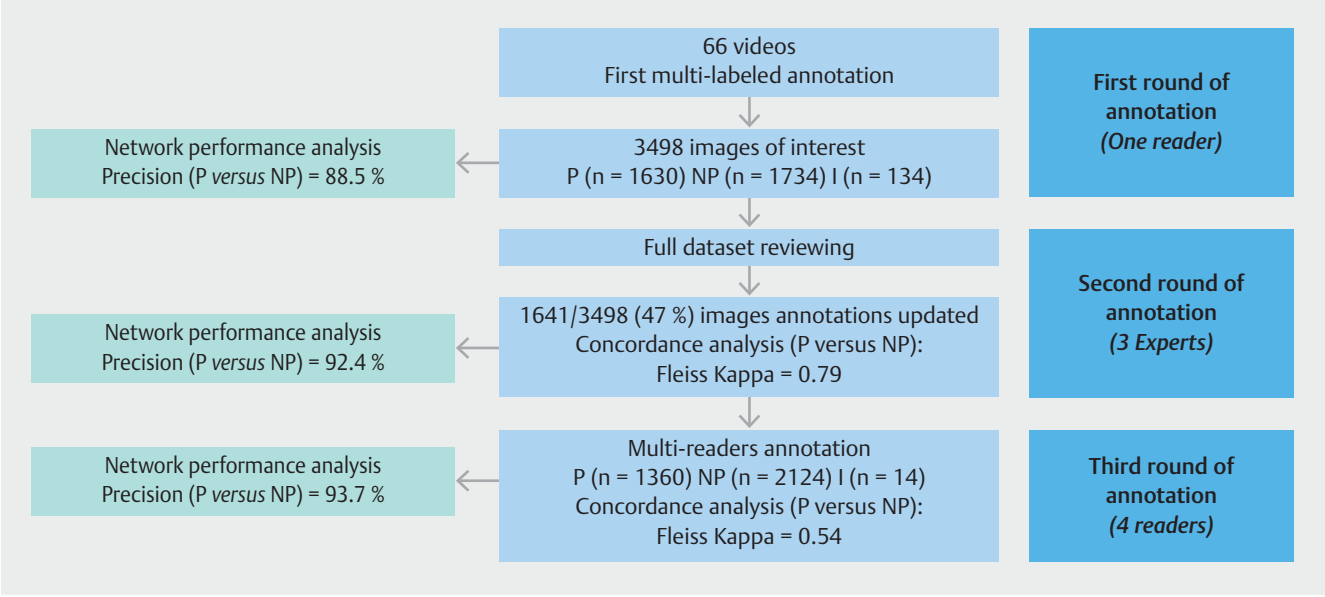
Performance of the networks in the CrohnIPI dataset were evaluated by calculation of mean sensitivity, specificity, and precision for each of the five tests in the cross-validation.

To evaluate the influence of multi-expert annotation, we trained and tested our neural networks on the first annotator dataset (AM) and then trained and tested on the multi-annotated dataset (AM, AB, CT, MF). A McNemar test was used to evaluate the differences in proportion of well-classified images by the neural network between the successive rounds of annotation.

Results

CrohnIPI dataset

Sixty-six videos containing at least one pathological frame obtained from 63 patients with CD were included in the dataset. Clinical and demographic characteristics of patients are detailed in **Supplementary Table 1**. Six patients were explored for a suspicion of CD, and at the end of the work-up, had a definitive diagnosis of CD based on the results of conventional



▶ **Fig. 2** Flowchart of the study. The final dataset was obtained after the selection of non-pathological (NP) and pathological (P) images of interest extracted from 66 SBCE performed in patients with CD by an initial reader. All the images were reviewed and annotated by three experts. The discordant images were read again by all four gastroenterologists to obtain a consensual annotation. Inconclusive images (I) were excluded of the dataset. The performance of the neural network has been tested at each step of the process as well as the concordance between readers.

▶ **Table 1** Number (%) of images containing each type of lesion and non-pathological images for each expert and agreement between experts.

	Erythema	Edema	Aphthoid ulceration	Ulceration 3–10 mm	Ulceration >10 mm	Stenosis	Non-pathological	Inconclusive
Expert 1	112 (3.2)	202 (5.8)	442 (12.6)	315 (9.0)	266 (7.6)	82 (2.3)	2037 (58.2)	42 (1.2)
Expert 2	68 (1.9)	58 (1.7)	372 (10.6)	357 (10.2)	95 (2.7)	296 (8.5)	2243 (64.1)	9 (0.3)
Expert 3	97 (2.8)	73 (2.1)	143 (4.1)	618 (17.7)	273 (7.8)	237 (6.8)	2011 (57.5)	46 (1.3)
Category-wise Fleiss Kappas	0.31	0.27	0.48	0.35	0.50	0.58	0.79	0.22

▶ **Table 2** Proportion of images identically classified by the experts with regards to the number of images with disagreement, according to the type of lesions. Three levels of labels are considered with their respective inter-observer agreement (Fleiss Kappa coefficient)

		Images classified as pathologic or non-pathologic	Images classified as non-pathologic or containing stenosis or ulcerations or edema/erythema	Images classified as non-pathologic or containing any type of lesions (S, U>10, U3–10, AU, O, E)
		Ratio of agreed and total images N/N (%)		
Non-pathologic		1827/2345 (78)	1827/2345 (78)	1827/2345 (78)
Pathologic	S	1134/1614 (70)	80/323 (25)	80/323 (25)
	U>10		658/1300 (51)	74/369 (20)
	U3–10			117/850 (14)
	AU			103/555 (19)
	O		39/406 (10)	16/250 (6)
	E			10/197 (5)
Inconclusive		0/94 (0)	0/94 (0)	0/94 (0)
Total		2961/3498 (85)	2604/3498 (74)	2227/3498 (64)
Fleiss Kappa coefficient		0.79	0.68	0.57

S, stenosis; U>10, ulceration>10 mm; U 3–10, ulceration between 3 and 10 mm; AU, aphthod ulceration; O, edema; E, erythema.

endoscopy and histology; SB lesions detected by the capsule were considered as CD lesions. From the videos, 3498 images were extracted and annotated by the first reader. According to this first reading, 1630 frames (46.6%) contained at least one lesion, 1734 (49.6%) were considered as non-pathological, and 134 were inconclusive (3.8%) (▶ **Fig. 2**).

After the second round of annotations, when distinguishing between pathological and non-pathological images, 537 images (15%) were differently labeled by at least one expert among three. Of the images, 2345 (68%) were coded at least once as non-pathological, 1614 (46%) as pathological, and 94 (2%) as inconclusive. Details of the lesion characterization by each expert are summarized in ▶ **Table 1**. ▶ **Table 2** presents the agreement between experts according to three different modalities of classification of images: non-pathological versus pathological; non-pathological versus stenosis or all types of ulcerations or edema and erythema; and non-pathological versus each type of lesion. The agreement among experts was good for distinguishing between pathological and non-pathological images with a kappa coefficient of 0.79 ($P<0.0001$). With intermediate coding of lesions, the global inter-observer agreement was substantial (k = 0.68, $P<0.0001$). With the finest coding of the lesions, the global inter-observer agreement was moderate (k = 0.57, $P<0.0001$).

Considering the annotations by the initial reader and the three experts, 1641 of 3498 images (47%) were annotated differently by at least one, corresponding to a moderate agreement (Fleiss' kappa = 0.54; $P<0.0001$). All 1641 images were reviewed during the third round of annotation to achieve a consensus annotation as close as possible to the "real truth." At the end of the process, 2124 of 3498 images were considered non-pathological (60.7%), 1360 pathological (38.9%), and 14 (0.4%) were inconclusive. Details of the consensual annotation are summarized in **Supplementary Table 2**.

Performance of the recurrent attention neural network

The task assigned to the algorithm was to classify images as non-pathological or pathological. Pathological images were defined as each image containing at least one lesion (see definition above) (▶ **Fig. 3**). The neural network was trained successively after the first, second, and third rounds of annotation.

▶ **Fig. 3** Identified lesions retained to define pathological images. The images show: **a** Two examples of erythema, **b** edema, **c** aphtoïd erosions, **d** ulcerations, and **e** stenosis.

We wanted to verify if the quality of the labeling improved the quality of the obtained model. Because inconclusive annotations cannot be used during training, we did the experiments twice. Indeed, generally in the machine learning domain, the more training data available, the better the results. Thus, as the number of inconclusive images decrease, an improvement would come from an increasing number of labeled images. The first experiments used only images conclusive in all three annotation steps (3331 images) and the second set of experiments used all available labels in the three rounds of annotation (respectively 3363, 3476 and 3484). ▶ **Table 3** presents the results of these two sets of experiments. In both situations, the precision, sensitivity, and specificity increased significantly as the annotation improved ($P = 0.014$). In the most favorable situation (more and better data), it obtained a final precision of 93.70 %, a sensitivity of 92.09 % and a specificity of 94.76 % for good detection of pathological/non-pathological images.

Three additional and available pre-trained networks: ResNet 34 and VGGNet 16 and 19, were tested on the CrohnIPI dataset using the same modalities to verify that the increased performance of our network following the improved quality of the labeling was reproducible [25, 26]. The results are shown in **Supplementary Table 3**. As for our recurrent attention neural network, the sensitivities, specificities, and accuracies increased progressively after each round of annotation, to reach a final accuracy of 94.58 %, 94.40 %, and 94.35 % for ResNet 34, VGG 16, and VGG 19, respectively.

The performance of our model, for each lesion, after the consensual annotation using all available labels is shown in the confusion matrix (▶ **Fig. 4**). The neural network classified correctly 2903 of 3484 (83 %) pathological, 2041 of 2124 (96 %) non-pathological images, and 502 of 705 (71 %) images of ulceration. Conversely, the performance was lower for the other lesions.

Discussion

The CrohnIPI is a CD specific and dedicated dataset of pathological and non-pathological images carefully reviewed by several experts to obtain the consensus annotation as close as possible to the "real truth" for training, validation, and testing of CAD tools. After three rounds of annotations, the dataset contained 3498 well-coded images with a large variety of CD mucosal lesions and non-pathological images chosen independently of the quality of bowel preparation reflecting as close as possible real-life conditions. The network performance was good with an accuracy reaching 93 %. Moreover, we demonstrated, using four different pre-trained networks, that the performance increased when the algorithm was tested on the multi-expert-annotated dataset rather than on the first-reader-annotated dataset, highlighting the major importance of high-quality annotation. For the first time, the interobserver agreement between experts and gastroenterology fellows for classification of SB mucosal lesion often seen in patients with CD was evaluated and confirmed the difficulty in correctly classifying these lesions.

The major strength of this study comes from the multi-reader annotation process described. The process that we used to create the CrohnIPI dataset has corroborated that readers agreed for classification of pathological and non-pathological images. However, the agreement concerning the different

▶ **Table 3** Neural network performance evaluated on CrohnIPI dataset after successive rounds of annotation.

	Number of images used	Accuracy (%)	Sensitivity (%)	Specificity (%)
	Using only common conclusive images for training			
One-reader annotation (1st round)	3331	88.53	86.24	90.67
Experts annotation (2nd round)[1]	3331	92.37	90.96	93.31
Consensual annotation (3rd round)	3331	93.70	92.89	94.76
	Using all conclusive images for training			
One-reader annotation (1st round)	3363	90.90	90.06	91.70
Experts annotation (2nd round)[1]	3476	91.83	88.45	94.00
Consensual annotation (3rd round)	3484	92.48	88.16	95.24

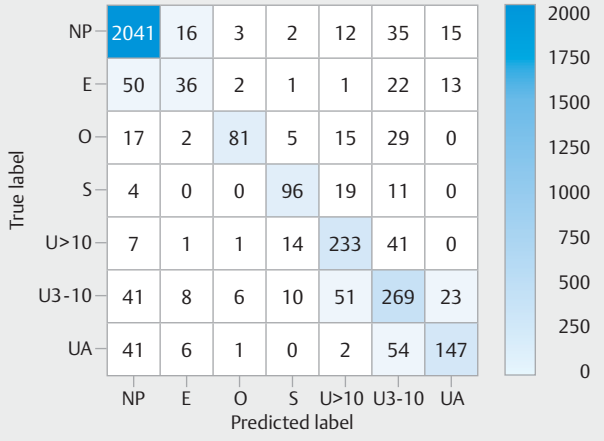
[1] Images were categorized as non-pathological or pathological when at least two readers among three were concordant.

types of SB lesions frequently seen in patients with CD was weaker. This discrepancy was not reduced by the fact that all readers learned capsule endoscopy in the same endoscopy unit and by use of a standardized definition of each lesion. Part of the disagreement could be due to the absence of common training before the annotation of images selected by the first reader. This could be tested later, during the next enrichment of the dataset with new images. The discrepancies concerning erythema and edema labeling also could be explained by the high variability and non-specificity of these lesions. This was highlighted by the difficulty to select typical images for the establishment of a consensus on the nomenclature and description of typical SB mucosal lesions of CD [14]. The difficulty in classifying these lesions had a negative impact on the performance of our neural network and highlights the need for a more rigorously developed dataset to properly train deep learning systems.

Other strengths of this study come from the use of native images on which no transformation and no filter were applied, permitting us to hedge on colorimetric bias and, unlike with other datasets, to use all typical lesions of CD without limitation to ulcerations. Moreover, non-pathological images were extracted from the same videos as pathological images and the images were not selected with regard to the quality of preparation, presence or not of bubbles, residue, or lightness, so as to simulate real-life clinical practice as much as possible. It is noteworthy that, with a lower number of images, our neural network reached similar performance to other networks trained and tested on larger datasets. This is not specific to our network, as some available pre-trained networks, such as ResNet and VGGNet, have had similar performance greater than 90 %. This could be explained by the quality of the labelling of images and also by the ratio of pathologic to non-pathologic images. Our dataset contains as many normal as pathological images, while in larger datasets, most of the images are non-pathologic, even if the task assigned to the network is to detect images containing at least one lesion.

The performance of our neural network was insufficient, however, for classification of all types of lesions. The performance was acceptable concerning ulcerations and stenosis but not for edema and erythema. This is clearly due to the non-specificity of these lesions and the great variability of their presentation. Other reasons could be the limited number of images for these classes available in the dataset and the imbalance between the different types of lesions. To improve the performance of the neural network, a larger number of images may be necessary or a different training strategy (e. g. sampling or weighting hard examples). Large dataset solutions have been explored in other studies for detection of erosions and ulcerations with datasets of more than 15 000 images, reaching accuracy rates close to 90 % [11, 12]. By contrast, no data on edema and erythema have been published yet.

Other limitations of the study come from the unique source of the images i. e. the Pillcam SB3, preventing generalization of this CAD tool to other devices and, despite the multicenter design of the study, by the learning of capsule endoscopy in the



▶ **Fig. 4** Confusion matrix of the classifier based on the final dataset: predicted labels of each type of lesion, erythema (E), edema (O), stenosis (S), aphthoid ulceration (AU), ulceration 3–10 mm (U3–10) and ulceration > 10 mm (U > 10).

same endoscopy unit by the experts, which could have introduced a bias compared to real-life multicenter studies. Future development of algorithms needs to be tested on different sources of images to be used in clinical practice.

The promising performance of our recurrent attention convolutional neural network paves the way for development of new automatic diagnosis tools for clinical practice. Even if classical architectures allow us to achieve significant detection performance, our attentional network aimed to improve the explainability of the model decisions [20]. Indeed, it provides attention localizations, which could be used by experts to assess the network prediction. This will be explored in future research. All these tools need to be trained and tested on multiple and various images. The CrohnIPI dataset was built to be shared for free with the scientific community to facilitate and accelerate the development of such tools, which also will be accessible to gastroenterologists in the future. The CrohnIPI dataset will be enriched over time by including pathological and non-pathological images representing all types of lesion. A new process of annotation will be tested by selecting the frames of interest by the neural network itself with a posteriori validation by a group of experts in capsule endoscopy and IBD. This process should facilitate enrichment of the dataset by limiting the needed number of analyzed images by the experts. In the medium term, enrichment of the dataset should make it possible to classify each type of lesion rather than just as pathological or non-pathological.

Conclusions

In conclusion, we developed a dataset of images of SBCD lesions with a process allowing us to approximate the "real truth" as much as possible. We demonstrated that the performance of our deep neural network increased in parallel with the quality of annotations, highlighting the need for the best possible annotated dataset. The objective was to enlarge the base for the institutional research teams that need images for to develop new CAD systems.

The CrohnIPI dataset can be downloaded, on-demand, at http://crohnipi.ls2n.fr/. The dataset is for research use only and protected under the creative commons license CC BY-NC-ND (https://creativecommons.org/licenses/by-nc-nd/2.0/). All the data produced in part or in totality with the CrohnIPI dataset need to reference the original publication.

Acknowledgements

This work was supported, in part, by an unrestricted grant from the IBD patient's association Francois Aupetit (https://www.afa.asso.fr/) and the interdisciplinary project CrohnIPI of Nantes University.

Competing interests

Arnaud Bourreille received lecture or consultancy fees from Abbvie, MSD, Celltrion, Medtronic, Takeda, Janssen, Gilead, Galapagos, Ose immunotherapeutics, Roche, Ferring, Pfizer, Tillotts and research grants from Medtronic, Takeda, Maunakea technologies, Abbvie, MedAdvanced.
Catherine Le Berre received lecture or consultancy fees from Janssen, Gilead, AbbVie, Ferring, Janssen, MSD, Pfizer and Takeda.
Caroline Trang received lecture or consultancy fees from AbbVie, Amgen, Janssen, MaaT Pharma, MSD, takeda, Arena, CT scout.
Mathurin Flamant received lecture or consultancy fees fromAmgen, Abbvie, Biogen, Celltrion, Janssen, MSD, Pfizer, Takeda Tillots Pharma and research grants from Abbvie and biosynex.
Harold Mouchère and Nicolas Normand received research grants from MyScript and Apricity.

References

[1] Iddan G, Meron G, Glukhovsky A et al. Wireless capsule endoscopy. Nature 2000; 405: 417–417

[2] Dionisio PM, Gurudu SR, Leighton JA et al. Capsule endoscopy has a significantly higher diagnostic yield in patients with suspected and established small-bowel crohn's disease: a meta-analysis. Am J Gastroenterol 2010; 105: 1240–1248

[3] Böcker U, Dinter D, Litterer C et al. Comparison of magnetic resonance imaging and video capsule enteroscopy in diagnosing small-bowel pathology: Localization-dependent diagnostic yield. Scand J Gastroenterol 2010; 45: 490–500

[4] Jensen MD, Nathan T, Rafaelsen SR et al. Diagnostic accuracy of capsule endoscopy for small bowel crohn's disease is superior to that of MR enterography or CT enterography. Clin Gastroenterol Hepatol 2011; 9: 124–129

[5] González-Suárez B, Rodriguez S, Ricart E et al. Comparison of capsule endoscopy and magnetic resonance enterography for the assessment of small bowel lesions in Crohn's disease. Inflamm Bowel Dis 2018; 24: 775–780

[6] Buisson A, Gonzalez F, Poullenot F et al. Comparative acceptability and perceived clinical utility of monitoring tools: a nationwide survey of patients with inflammatory bowel disease. Inflamm Bowel Dis 2017; 23: 1425–1433

[7] Le Berre C, Sandborn WJ, Aridhi S et al. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology 2020; 158: 76–94.e2

[8] Muhammad K, Khan S, Kumar N et al. Vision-based personalized Wireless Capsule Endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges. Future Gen Comp Sys 2020; 113: 266–280

[9] Yuan Y, Meng MQ-H. Deep learning for polyp recognition in wireless capsule endoscopy images. Med Phys 2017; 44: 1379–1389

[10] Leenhardt R, Vasseur P, Li C et al. A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy. Gastrointest Endosc 2019; 89: 189–194

[11] Aoki T, Yamada A, Aoyama K et al. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointest Endosc 2019; 89: 357–363

[12] Fan S, Xu L, Fan Y et al. Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. Phys Med Biol 2018; 63: 165001

[13] Alaskar H, Hussain A, Al-Aseem N et al. Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy Images. Sensors (Basel) 2019; 19: 1265

[14] Klang E, Barash Y, Margalit RY et al. Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. Gastrointest Endosc 2020; 91: 606–613

[15] Leenhardt R, Li C, Le Mouel JP et al. CAD-CAP: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. Endosc Int Open 2020; 8: E415–E420

[16] Ding Z, Shi H, Zhang H et al. Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. Gastroenterology 2019; 157: 1044–1054

[17] Wang S, Xing Y, Zhang L et al. Deep convolutional neural network for ulcer recognition in wireless capsule endoscopy: experimental feasibility and optimization. Comput Math Methods Med 2019; 2019: 7546215

[18] Leenhardt R, Buisson A, Bourreille A et al. Nomenclature and semantic descriptions of ulcerative and inflammatory lesions seen in Crohn's disease in small bowel capsule endoscopy: An international Delphi consensus statement. United European Gastroenterol J 2020; 8: 99–107

[19] Gralnek IM, Defranchis R, Seidman E et al. Development of a capsule endoscopy scoring index for small bowel mucosal inflammatory change: development of a capsule endoscopy scoring index. Aliment Pharmacol Ther 2007; 27: 146–154

[20] Vallée R, de Maissin A, Coutrot A. Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. IEEE 21st International Workshop on Multimedia Signal Processing (MMSP 2019), Sep 2019, Kuala Lumpur, Malaysia.

[21] Vallée R, de Maissin A, Coutrot A. CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms. Proc SPIE 2020; 11317: doi:10.1117/12.2543584

[22] Deng J, Dong W, Socher R et al. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conf Comput Vis Pattern Recognit Miami, FL: IEEE; 2009. 248–255. doi:10.1109/CVPR.2009.5206848

[23] Cho K, van Merrienboer B, Bahdanau D et al. On the properties of neural machine translation: encoder-decoder approaches. ArXiv14091259 Cs Stat 2014.

[24] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76: 378–382

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ICLR – 2015 conference paper. ArXiv: 1409.1556v6.

[26] Kaiming He, Xiangyu Z, Shaoqing R et al. Deep residual learning for image recognition. ILSVRC 2015 conference paper. ArXiv: 1512.03385v1.