CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn’s disease lesions recognition algorithms
Rémi Vallée, Astrid de Maissin, Antoine Coutrot, Harold Mouchère, Arnaud Bourreille, Nicolas Normand

To cite this version:
Rémi Vallée, Astrid de Maissin, Antoine Coutrot, Harold Mouchère, Arnaud Bourreille, et al.. CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn’s disease lesions recognition algorithms. SPIE Medical Imaging, Feb 2020, Houston, France. pp.61, 10.1117/12.2543584. hal-02518263

HAL Id: hal-02518263
https://hal.science/hal-02518263
Submitted on 20 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn’s disease lesions recognition algorithms

R. Vallée, A. De Maissin, A. Coutrot, H. Mouchère, A. Boureille, and N. Normand

Université de Nantes, CNRS, LS2N, F-44000 Nantes, France

CHU Nantes, Institut des Maladies de l’Appareil Digestif, CIC Inserm 1413, Université de Nantes, Nantes, France

ABSTRACT

Wireless capsule endoscopy (WCE) allows medical doctors to examine the interior of the small intestine with a noninvasive procedure. This methodology is particularly important for Crohn’s disease (CD), where an early diagnosis improves treatment outcomes. The counting and identification of CD lesions in WCE videos is a time-consuming process for medical experts. In the era of deep-learning many automatic WCE lesion classifiers, requiring annotated data, have been developed. However, benchmarking classifiers is difficult due to the lack of standard evaluation data. Most detection algorithms are evaluated on private datasets or on unspecified subsets of public databases. To help the development and comparison of automatic CD lesion classifiers, we release CrohnIPI, a dataset of 3498 images, independently reviewed by several experts. It contains 60.55% of non-pathological images and 38.85% of pathological images with 7 different types of CD lesions. A part of these images are multilabeled. The dataset is balanced between pathological images and non-pathological ones and split into two subsets for training and testing models. This database will progressively be enriched over the next few years in aim to make the automatic detection algorithms converge to the most accurate system possible and to consolidate their evaluation.

Keywords: Dataset, Video Capsule Endoscopic, Deep learning, Crohn’s disease

INTRODUCTION

Since the mid-1990s, the development of wireless capsule endoscopy (WCE) improves the diagnosis of the Crohn’s disease (CD) and its early treatment. This minimally invasive technology offers significant diagnostic performance due to its ability to fully explore the 3 to 4 metres of the small bowel. It is now well accepted by the entire scientific community. The diagnosis of CD is based on two scores: the Capsule Endoscopy Crohn’s Disease Activity Index (CECDAI) and Lewis’ score. These indicators depend on the number of lesions, their types and locations in the small bowel. As these scores are related to the number and precise identification of the lesions, it is necessary for gastroenterologists to view all 50,000 images generated by the WCE for each patient, which can be very time-consuming (30-60 minutes per read in most studies). Various systems have been developed to limit the time required for gastroenterologists to analyze the WCE data. These systems can be classified into 2 main categories. First, algorithms based on the extraction of well-identified features based on SVM. Second, algorithms based on deep neural networks. The latter obtain better classification performance but need more data in order to be trained properly. However, there is currently no public database to benchmark these algorithms. They are currently evaluated on different personal databases. The release of CrohnIPI comes to challenge this problem with 3498 carefully annotated images. Collecting large scale medical databases is difficult due to the limited number of available experts able to deliver reliable annotations. The CrohnIPI database has been designed to compensate for this lack of precious resource, making available the WCE images along with multi-label, multi-expert annotations.

Further author information: (Send correspondence to R. Vallée and H. Mouchère)
R. Vallée: E-mail: remi.vallee@univ-nantes.fr
H. Mouchère: E-mail: harold.mouchere@univ-nantes.fr
2. DATABASE DESCRIPTION

2.1 Method

The images that make up the database were collected using the PillCam3 and come from 63 different patients from the Nantes University Hospital. Their resolution is 320px by 320px. Table 1 describes the characteristics of the patients included in this database.

<table>
<thead>
<tr>
<th>Patients, n</th>
<th>63</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women, n (%)</td>
<td>37 (40)</td>
</tr>
<tr>
<td>Age, median (IQR)</td>
<td>39 (28-49)</td>
</tr>
<tr>
<td>Duration of illness in years, median (IQR)</td>
<td>9.7 (2-16)</td>
</tr>
</tbody>
</table>

The various examinations required to monitor Crohn’s disease were carried out using the PillCam3 video capsule, and generated approximately one million images. These images were initially reviewed during the clinical examination by the referring gastroenterologist of each patient. Then, in a second step, an external expert reviewed the videos and annotated certain images, with the aim of selecting images that varied both in terms of pathologies they contained and in term of viewing conditions (images containing bubbles, blurred, with lesions at the edge of focal fields, etc). In order to be able to train deep neural networks more easily, the expert selected, to the best of their knowledge, the same quantity of pathological and non-pathological images. The images reviewed during this second phase were reviewed in their context, i.e. in the full video stream. Thus the expert could use previous and future frames to annotate the current frame.

During this second phase, 3498 images have been annotated. In order to consolidate the annotations obtained, they were then reviewed by 3 independent experts thanks to an application developed by our team (see fig. 1). In this last annotation phase, the images were seen out of context. Each image was reviewed independently of the previous and following images, and without any idea of the patient undergoing the examination.

In both annotation phases the experts had the choice between 7 different non-exclusive labels. These different labels and their associated descriptions are presented in Table 2. The possibility to answer “I don’t know” was also possible and referenced by the NC label.
Table 2. Label description and distribution in the database. 14% of the images have multiple labels (1.2 labels/image in average).

<table>
<thead>
<tr>
<th>Nomenclature</th>
<th>Label</th>
<th>Description</th>
<th>Example</th>
<th>Percentage of the base containing the label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphthoid ulceration</td>
<td>UA</td>
<td>Minimal loss of substance of the epithelial layer, with a white background, surrounded by an erythematous halo, lying on a normal mucous membrane.</td>
<td></td>
<td>9.45%</td>
</tr>
<tr>
<td>Stenosis</td>
<td>S</td>
<td>Narrowing of the intestinal lumen delaying or preventing the passage of the capsule (needs to be evaluated on a video).</td>
<td></td>
<td>5.86%</td>
</tr>
<tr>
<td>Edema</td>
<td>O</td>
<td>Appearance enlarged, swollen, engorged villi.</td>
<td></td>
<td>10.28%</td>
</tr>
<tr>
<td>Erythema</td>
<td>E</td>
<td>Isolated redness, often as a result of a local accumulation of venous blood.</td>
<td></td>
<td>13.25%</td>
</tr>
<tr>
<td>Ulceration of more than 10mm</td>
<td>U+10</td>
<td>Loss of depressed substance in relation to the surrounding mucous membrane which is swollen/oedematous and has a whitish background whose size is greater than 10mm.</td>
<td></td>
<td>7.94%</td>
</tr>
<tr>
<td>Ulceration of less than 10mm</td>
<td>U3-10</td>
<td>Loss of depressed substance in relation to the surrounding mucous membrane which is swollen/oedematous and has a whitish background whose size is inferior than 10mm.</td>
<td></td>
<td>14.60%</td>
</tr>
<tr>
<td>Non pathologic</td>
<td>NP</td>
<td>Normal intestinal image</td>
<td></td>
<td>60.55%</td>
</tr>
</tbody>
</table>
2.2 Dataset content

The proportion of each class is available in Table 2. Frames containing the labels UA, U+10, U3-10, E, O and S have been grouped into a so-called pathological class. Once the different classes of CD lesion have been grouped in a single pathological class, we aggregate the votes of the 3 annotators of the third phase, as shown in Table 3.

We made the choice to provide all users with the raw data, i.e. with the labels provided by each annotator on each image. We hope that this will allow the development of techniques for merging multi-label and multi-annotator annotations, in order to improve the automatic classification of pathological images.

<table>
<thead>
<tr>
<th>Label</th>
<th>Non pathologic</th>
<th>Pathologic</th>
<th>Don’t know</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selector</td>
<td>49.57% (1734)</td>
<td>46.60% (1630)</td>
<td>3.83% (134)</td>
<td>100% (3498)</td>
</tr>
<tr>
<td>3 annotators</td>
<td>60.55% (2118)</td>
<td>38.85% (1359)</td>
<td>0.60% (21)</td>
<td>100% (3498)</td>
</tr>
</tbody>
</table>

We calculated the agreements based on Fleiss’ kappa. We obtained an overall score of 78.35%, which corresponds to a good agreement rate. Table 4 shows the percentage of the base where all three experts agree, almost agree (2 vs 1), and completely disagree (all different).

3. RESULTS

In this part we show the results obtained with state-of-the-art deep neural networks to establish a baseline. To evidence the importance of multi-expert annotation, we compare the results obtained by a model trained with a dataset annotated by a single annotator (the selector) to a model trained on a dataset annotated by several annotators. Most medical imaging databases rely on the diagnosis of only one annotator at a time. In order to evaluate the influence of a multi-expert annotation, we trained and tested state-of-the-art neural networks on the selector database and then trained and tested on the multi-annotation database. The results are presented in Table 5 and Table 6. Images containing the label “Don’t know” (3.83%) have been excluded from the training and testing sets. To balance the sample size between the two sets of data, we removed from the multi-annotated database 134 images with the label “I don’t know” annotated by the selector, as the number of images with the “I don’t know” label is greater in the selector database than in the multi-annotated one.

<table>
<thead>
<tr>
<th>2 classes</th>
<th>Accuracy</th>
<th>F1 score</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet34</td>
<td>93.67%</td>
<td>93.39%</td>
<td>94.42%</td>
<td>92.39%</td>
<td>97.87%</td>
</tr>
<tr>
<td>ResNext</td>
<td>93.01%</td>
<td>92.78%</td>
<td>93.10%</td>
<td>92.58%</td>
<td>97.82%</td>
</tr>
<tr>
<td>VGG19</td>
<td>92.60%</td>
<td>92.32%</td>
<td>93.21%</td>
<td>91.49%</td>
<td>97.15%</td>
</tr>
<tr>
<td>VGG16</td>
<td>92.48%</td>
<td>92.20%</td>
<td>92.83%</td>
<td>91.61%</td>
<td>97.55%</td>
</tr>
</tbody>
</table>

Table 6 shows that an increase in the quality of the annotations leads to a better performance for each network on each of the metrics. This improves the model in every way, as both the number of false positives and false negatives in the database are reduced. By merging the annotations by voting, the labels on each image become more robust. This makes it easier for the model to be trained, as the number of false positives and false negatives in the database is reduced.
Table 6. Results of different cross-validated training (5 folds) with merged labels.

<table>
<thead>
<tr>
<th></th>
<th>2 classes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>ResNet34</td>
<td>94.56%</td>
</tr>
<tr>
<td>ResNext</td>
<td>93.81%</td>
</tr>
<tr>
<td>VGG19</td>
<td>93.66%</td>
</tr>
<tr>
<td>VGG16</td>
<td>93.24%</td>
</tr>
</tbody>
</table>

The best performance is obtained with the ResNet34 network on the multi-annotation database, with an accuracy of 94.56% and an AUC (Area Under ROC Curve) of 98.23%.

4. CONCLUSION

We provide a dataset with 3498 WCE images independently reviewed by several gastroenterologist experts. 60.55% are non pathological, 38.85% present at least one of 7 different type of CD lesions. We chose to provide the raw annotated data, showing potential classification discrepancies between experts for some particularly ambiguous images. This approach repositions data quality at the center of the problem of evaluating deep learning algorithms in medical imaging. More specifically, it raises the question of handling multi-label, multi-expert annotations for the training of models for the detection and identification of CD lesions. Such uncertainty, far from adding noise to the results, constitutes valuable information that the model can use to improve its performance.

This database will be further improved by carrying out a final phase of consensus building among the different experts to assign a final label to each litigious images. We will also continue to progressively enrich the database with new WCE images, and more labels assigned by more gastroenterologist experts. We hope that this database will become the gold standard for the evaluation of CD lesion detection algorithms.

Data availability

Instructions to download the CrohnIPI dataset are available at [https://crohnipi.ls2n.fr/](https://crohnipi.ls2n.fr/).

Acknowledgments

This work was supported in part by an unrestricted grant from the IBD patient’s association François Aupetit and the interdisciplinary project CrohnIPI of Nantes university ([https://www.afa.asso.fr/](https://www.afa.asso.fr/)).

REFERENCES


