DataSet Description

The Quiron database has 245 WSI scored by a pathologist according to H. pylori density as NEGATIVE (Healthy), LOW DENSITY and HIGH DENSITY. Of the 245 patients included in the study, 117 (47.8% of the total) are classified as NEGATIVE, while 128 are classified as POSITIVE (LOW and HIGH DENSITY) for the presence of H. pylori. Each WSI contains 2 sections of several gastric mucosa samples and one section extracted from an external positive control to check the validity of the immunohistochemical staining.

1. Images

1.1 WSI example

These are the original WSI in .tiff format. Each .tiff file contains 3 digitalized sections with several tissue samples digitalized at high resolution (40 augmentations).

1.2 Patches

Patches of size 256x256 were cropped along the border of the tissue samples of one of the sections. Cropped images are inside a folder identified by the PatID and the tissue section as PatID_Section#. Patches can be explored at http://ivendis.cvc.uab.cat/Digipatics/, with user isbi_user and password isbi_pass.

From a subset of 123 cases (77 positive ones), an expert pathologist annotated 1211 patches (161 being positive ones) extracted from the first slide. The number of positive annotated patches has been augmented by cropping windows along tissue border from the point the annotated patches were extracted from. They are labelled with the tag '_Aug'. The set of annotated cases was complemented with images from negative cases in order to balance the set at diagnosis level. These subset of cases with annotated patches can be used for CrossValidation of methods, while the remaining cases are an independent HoldOut set for verification of reproducibility.

2. Folder Structure

Figure 1 shows the main structure of folders. The CrossValidation folder contains patches from patients with annotations of presence of helicobacter in patches, as well as, global diagnosis of the sample. This set should be used for training and validation of models for, both, patch classification and patient diagnosis. The HoldOut folder contains patches cropped from patients that only have global diagnosis. This cases can be used to assess reproducibility. The diagnosis of the whole set of patients is given in the file PatientDiagnosis.csv.

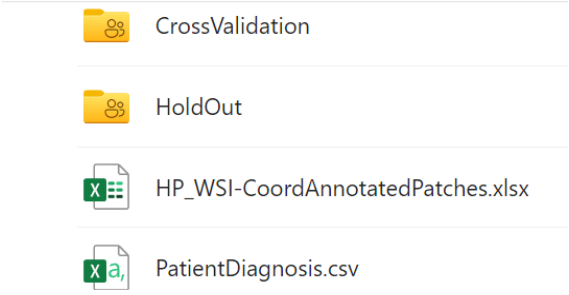


Figure 1. Data Base Folder Main Directory. CrossValidation contains patches from patients with annotations of presence of helicobacter in patches for training and validation of models. HoldOut contains patches cropped from patients that only have global diagnosis.

The CrossValidation has to set of patches, the Annotated ones that have identified the presence of helicobacter and the total set of patches Cropped from the tissue sample. The annotated ones are a subset of the Cropped. Annotations of presence of the bacteria in this set is given in the file HP_WSI-CoordAnnotatedPatches.xlsx. For positive patches, the set of augmentations extracted from neighboring pixels are also included. The folder also contains the extra negative patches extracted from healthy patients. This cases are not annotated in HP_WSI-CoordAnnotatedPatches.xlsx.

Both folders contain one subfolder for each patient's patches. Each patient folder is identified by the patient code or PatID (B22-#) and the tissue section number (1,0): PatID_#Section. The HolsOut set also follows the same distribution in subfolders containing the patches cropped for each patient.

Figure 2 shows the distribution and folder structure of the CrossValidation set.

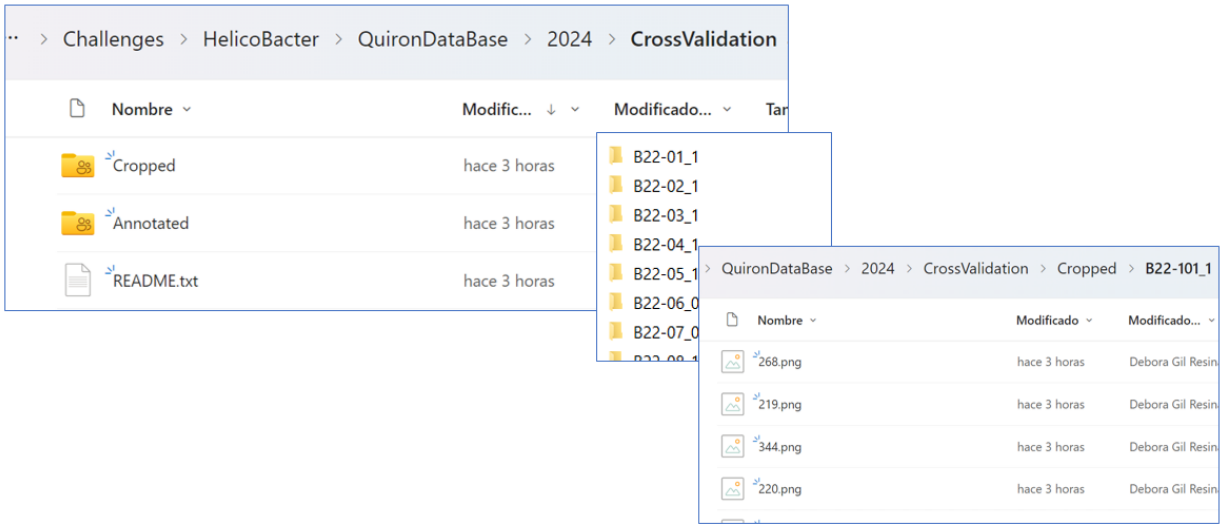


Figure 2. Folder distribution of the CrossValidation set

3. Metadata

3.1 Annotated Patches

The file HP_WSI-CoordAnnotatedPatches.xlsx (see Figure 3) contains information about patches analyzed by a pathologist regarding the presence of Helicobacter pylori on the patch. For each patch, we have the following information:

Pat_ID	Section_ID	Window_ID	i	j	h	w	Presence
B22-129	0	659	7477	11978	256	256	-1
B22-68	0	131	6597	12009	256	256	-1
B22-68	0	141	5100	10737	256	256	-1
B22-68	0	290	5015	14908	256	256	-1
B22-68	0	298	11626	13928	256	256	-1
B22-68	0	315	9541	11016	256	256	-1
B22-68	0	352	7536	9139	256	256	-1
B22-52	0	760	5664	12310	256	256	-1
B22-124	0	11	47727	30234	256	256	-1
B22-124	0	759	28997	17596	256	256	-1

Figure 3. Metadata of the annotated patches

- Pat_ID, Section_ID, Window_ID: Patient, Section and patch identifiers. Window_ID is the patch image filename. Pat_ID_Section_ID is the name of the folder containing the image (.png format)

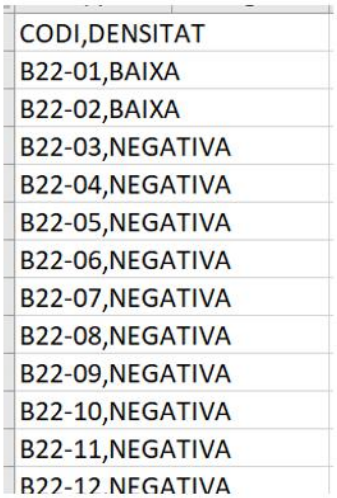
- i,j,h,w define the position of the patch in the .tiff image with respect the left upper corner of the image. Values are in matrix coordinates with access [i:i+h, j:j+w].

- Presence indicates whether H. pylori was observed in that patch (1) or not (-1). A 0 represents a patch that was not possible to classify.

The file HP_WSI-CoordAnnotatedPatches.xlsx only contains the annotations made by the pathologist. The folder Annotated contains extra patches extracted from healthy cases of the whole CrossValidation set. This extra cases are not in the excel file since they all correspond to negative patches.

3.2 Patient Diagnosis

The file PatientDiagnosis.csv (see Figure 4) contains information about the presence of Helicobacter pylori on the sampled tissue. The first column, CODI, has the Pat_ID of each WSI image, and the second column indicates the degree of H. pylori presence: Negative (NEGATIVA), Low (BAIXA), or High (ALTA). The positive cases can be group into a single group for binary classification (positive/negative) of the patient diagnosis.



CODI,DENSITAT
B22-01,BAIXA
B22-02,BAIXA
B22-03,NEGATIVA
B22-04,NEGATIVA
B22-05,NEGATIVA
B22-06,NEGATIVA
B22-07,NEGATIVA
B22-08,NEGATIVA
B22-09,NEGATIVA
B22-10,NEGATIVA
B22-11,NEGATIVA
B22-12,NEGATIVA

Figure 4. Contents of file PatientDiagnosis.csv