MHIST: A Minimalist Histopathology Image Analysis Dataset

* MHIST is a binary classification dataset of 3152 H&E stained Formalin Fixed Paraffin-Embedded fixed size images (224 by 224 pixels) of colorectal polyps from the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center
  + Taken from 328 whole slide images scanned by an Aperio AT2 scanner at 40x resolution
  + To increase the field of view, we compress the slides with 8x magnifica-
  + tion.
  + 345 MB of disk space
* Each with a gold-standard label determined by 7 gastrointestinal pathologists based on their majority vote (also includes each image’s annotator agreement level)
* 2 classes in the dataset: hyperplastic polyp (HP), usually benign, and sessile serrated adenoma (SSA), precancerous and requires a sooner follow-up examination
* The dataset includes: annotations, images, MD5SUMs.txt (contains a checksum that verifies if contents are downloaded correctly)
* File includes each image file name and the corresponding majority-vote label and degree of annotator agreement
  + expressed as the number of annotators who marked the image as SSA (6 represents 6/7 agreement with a ground truth of SSA and 2 would indicate 5/7 agreement with a ground truth of HP)
  + 16.7% of the cases were 4/7
* DeepSlide - open-source framework for histology image analysis available to develop deep learning models
* Through multiple example use cases it was shown that the best results were achieved with the pretrained ResNet-18 model (AUC 92.7%)
* Limitations:
  + Using fixed-size images is not as precise as using whole-slide images as whole-slide images contain much more information
  + The dataset does not include any demographic information about patients or any information on the size and location of the polyp

https://bmirds.github.io/MHIST/

https://medium.com/analytics-vidhya/mhist-a-new-public-histopathological-image-dataset-for-ml-community-22e557aa2e3