The Journal of Open Source Software

# Histomics Label

**Brianna Major** [1], **Jeffery A. Goldstein** [2], **Michael Nagler** [1], **Lee A. Newberg** [1], **Abhishek Sharma** [2], **Anders Sildnes** [2], **Faiza Ahmed** [1], **Jeff Baumes** [1], **Lee A. D. Cooper** [2], **and David Manthey** [1]

**1** Kitware, Inc., New York, United States **2** Northwestern University Feinberg School of Medicine, Illinois, United States

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:** julia ferraioli

**Reviewers:**

- @cinnetcrash
- @Mmasoud1

**Submitted:** 20 June 2025
**Published:** unpublished



## Summary

`Histomics Label` is a software tool for the interactive development of machine-learning classifiers for whole slide pathology images. It is deployed as part of the Digital Slide Archive (Gutman et al., 2017; Kitware, Inc, 2025a), a web-based data management system for whole slide image datasets, and was built on top of HistomicsUI (Kitware, Inc, 2025c) and uses the HistomicsTK (Kitware, Inc, 2025b) image analysis tool kit.

Users label image regions or tissue structures to provide training data for classifiers and iteratively improve these classifiers by reviewing their output and providing additional labels. The interface uses heuristics to guide users to the most impactful examples to label, and supports bulk labeling of samples and review of labeled examples for collaboration. An example data generation pipeline is included that segments a whole slide image into superpixels, and generates feature embeddings for segmented regions using a foundation model.

## Statement of need

One of the limitations in developing classification models is the need for labeled data. In pathology and other medical fields, the expertise required for labeling and busy schedules of medical experts make labeling particularly challenging. For whole slide images, where each image can contain several billion pixels, navigating vast datasets in search of possibly rare tissue states can be very inefficient and frustrating. Software interfaces need to be optimized for the user experience and make the most of an expert's time and energy.

Other issues in labeling include the volume and accessibility of data. Software that must run local to the data requires that all data be copied and correctly versioned for the project. Using a web-client and server model with appropriate permission models, only requires that the data be on centrally managed server. This allows there to be a single, coordinate source of data for a project, and reduces the burden on individual users to only requiring a web browser and ordinary internet connection. This enables collaboration between multiple experts, or to allow experts to review the work of their trainees.

`Histomics Label` uses a technique called active learning to identify the unlabeled examples that can provide the most benefit to classifier performance and provides an intuitive workflow for presenting these examples to experts for efficient labeling. Data can be generated using a built-in pipeline that partitions whole-slide images into superpixels, or users can provide their own data from external cell or tissue segmentation algorithms. Users specify the categories that can be labeled and assign display properties like color, and can exclude categories from classifier training (for instance, for regions whose categories cannot be accurately determined). After labeling a few initial example regions, a classifier is trained and used to both predict the category of all regions and the unlabeled regions that provide the most classifier benefit. The

user can retrain the classifier at any time and review the classifier predictions and labels from other users. Labeling can also be performed by painting directly on the whole slide image with a brush tool.

For development, the initial segmentation uses superpixels generated with the SLIC (Achanta et al., 2012) algorithm. These are computed on whole slide images in a tiled manner so that they can work on arbitrarily large images, and the tile boundaries are properly handled to merge seamlessly. Once generated, segments are represented in one of two ways, either as two-dimensional patches, each centered in a fixed-sized square of masked pixels, or as one-dimensional feature embeddings, such as those generated from the huggingface UNI (Chen et al., 2024) foundation model. One of two basic models is trained based upon the segment representation. For two-dimensional patches, the model to be trained is a small-scale CNN implemented in tensorflow/keras or torch. For one-dimensional vectors, the model to be trained is a single-layer linear classifier. The certainty criteria for which segments should be labeled next can also be selected, and includes confidence, margin, negative entropy, and the BatchBALD (Kirsch et al., 2019) algorithm.

We had a placental pathologist provide feedback to validate the efficiency of the user interface and utility of the process.

## Basic Workflow

When starting a new labeling project, the user selects how superpixels are generated, which certainty metric is used for determining the optimal labeling order, and what features are used for model training. The labeling mode allows defining project labels and performing initial labeling. This mode can also be used to add new label categories or combine two categories if they should not have been distinct. Label categories can additionally be marked as excluded, which removes them from training and ensures that superpixels with those labels are no longer suggested for labeling.

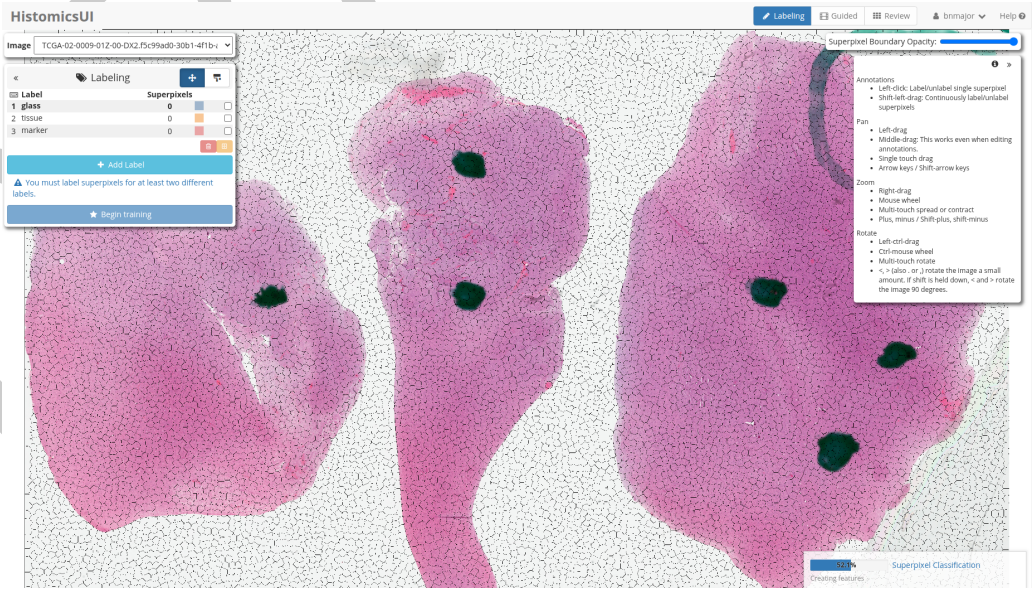


**Figure 1:** The Bulk Labeling interface showing one of the project images divided into superpixels with some categories defined. A user can "paint" areas with known labels as an initial seed for the guided labeling process

Once some segments have been labeled and an initial training process has been performed, additional segments are shown with their predictions. The user can use keyboard shortcuts or

The Journal of Open Source Software

the mouse to confirm or correct labels. These are presented in an order that maximizes the utility of improving the model based on the originally selected certainty metric.

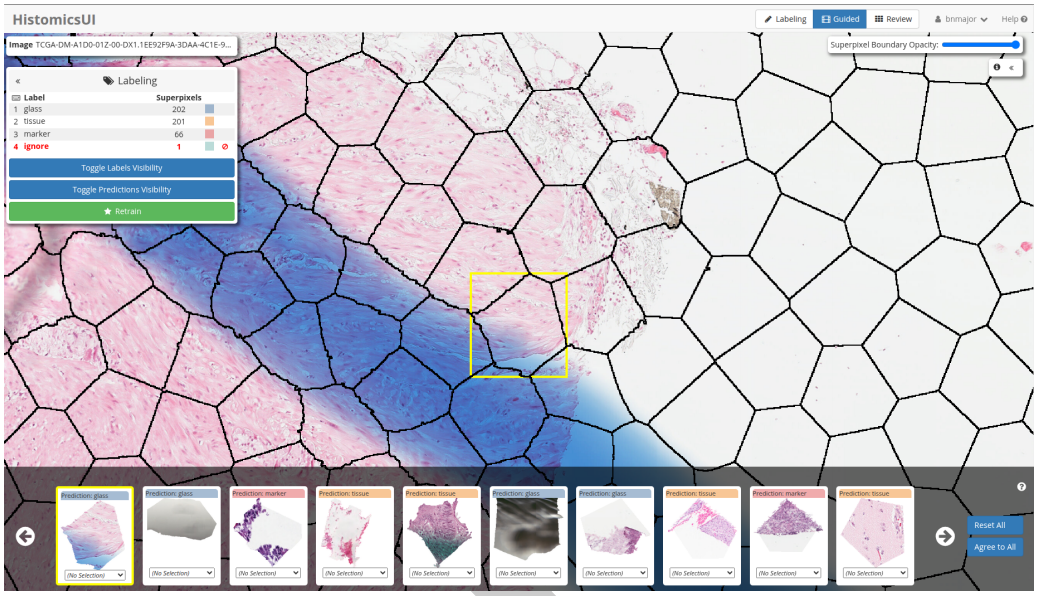


**Figure 2:** The Guided Labeling interface showing a row of superpixels to be labeled and part of a whole slide image

To check on overall behavior or correct mistakes, there is a review mode that allows seeing all labeled segments with various filtering and sorting options. This can be used to check agreement between pathologists or determine how well the model agrees with the manually labeled data.

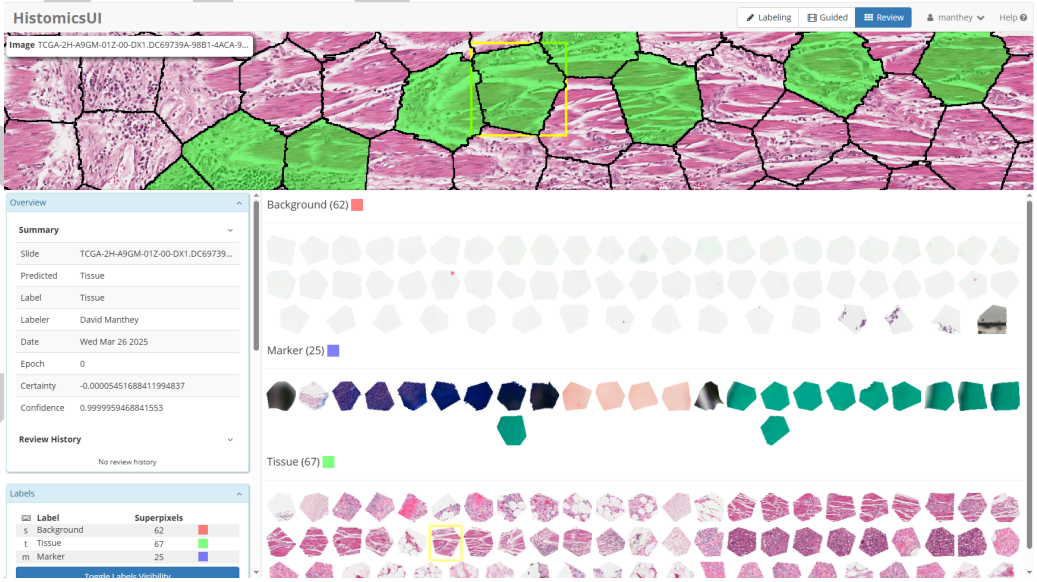


**Figure 3:** The Review interface showing labeled superpixels in each category

The whole slide image data in these figures are from data generated by the TCGA Research Network (**?**).

The Journal of Open Source Software

## Acknowledgements

This work has been funded in part by National Library of Medicine grant 5R01LM013523 entitled "Guiding humans to create better labeled datasets for machine learning in biomedical research".

## References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2282. https://doi.org/10.1109/TPAMI.2012.120

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, *30*(3), 850–862. https://doi.org/10.1038/s41591-024-02857-3

Gutman, D. A., Khalilia, M., Lee, S., Nalisnik, M., Mullen, Z., Beezley, J., Chittajallu, D. R., Manthey, D., & Cooper, L. A. D. (2017). The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. *Cancer Research*, *77*(21), e75–e78. https://doi.org/10.1158/0008-5472.can-17-0629

Kirsch, A., Amersfoort, J. van, & Gal, Y. (2019). BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. *CoRR*, *abs/1906.08158*. http://arxiv.org/abs/1906.08158

Kitware, Inc. (2025a). *Digital slide archive: A system for working with large microscopy images*. https://github.com/DigitalSlideArchive/digital_slide_archive

Kitware, Inc. (2025b). *HistomicsTK: A python package for the analysis of digital pathology images*. https://doi.org/10.5281/zenodo.14833780

Kitware, Inc. (2025c). *HistomicsUI: Organize, visualize, annotate, and analyze histology images*. https://doi.org/10.5281/zenodo.5474914