

RESEARCH ARTICLE

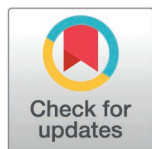
SmartHisto: Bayesian active learning for histology images

Sriram Vijendran^{1*}, Bailey Arruda², Tavis K. Anderson², Oliver Eulenstein^{1*}

1 Department of Computer Science, Iowa State University, Ames, Iowa, United States of America,

2 National Animal Disease Center, Agricultural Research Service, United States Department of Agriculture, Ames, Iowa, United States of America

* sriramv@iastate.edu (SV); oeulens@iastate.edu (OE)



OPEN ACCESS

Citation: Vijendran S, Arruda B, Anderson TK, Eulenstein O (2026) SmartHisto: Bayesian active learning for histology images. PLoS Comput Biol 22(9): e1013611. <https://doi.org/10.1371/journal.pcbi.1013611>

Editor: Virginia E. Pitzer, Yale School of Public Health, UNITED STATES OF AMERICA

Received: October 10, 2025

Accepted: August 24, 2026

Published: September 17, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1013611>

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under

Abstract

Accurate and efficient characterization of biological images is crucial for advancing systems biology and medical research. Recent advancements in deep learning and image processing have enabled neural network models to rapidly accelerate image analysis by utilizing large expert-annotated datasets. However, in histopathology, the size of whole-slide images makes expert annotation expensive, limiting the acquisition of sufficiently large annotated datasets and posing a major challenge for developing automated, AI-driven image analysis pipelines. To address this limitation, we propose a novel active learning-based framework to train image segmentation models interactively. Our approach employs a Bayesian neural network to identify informative regions in unlabeled images rather than entire images, making expert labeling more cost-effective. We validate our framework on multiple benchmark datasets with variable staining at fixed magnifications, demonstrating substantial reductions in annotation requirements. Notably, our method achieves a mean IoU of 0.75, significantly outperforming competing approaches, which averaged 0.60.

Author summary

Histopathology is fundamental to investigating tissue and immune responses, host-pathogen interactions, and disease mechanisms. However, histopathology is highly resource-intensive and requires specialized training, dramatically increasing the costs of annotating whole-slide images and, consequently, the expenses of large-scale studies involving numerous labs and specialists. We developed a computational tool to overcome these challenges, implementing a robust uncertainty-based sampling algorithm in conjunction with a next-generation Bayesian Convolutional Neural Network. This algorithm can be used for hypothesis testing and discovery by reducing the reliance on large, precisely annotated training datasets required in automated image analysis pipelines. The base model, when trained to identify lung tissue types using a small set

the [Creative Commons CC0](#) public domain dedication.

Data availability statement: The source code used to produce the results and analyses presented in this manuscript are available from GitHub repository: https://github.com/flu-crew/histology_segmentation. The pulmonary dataset is available at Zenodo with DOI: [10.5281/zenodo.18421739](https://doi.org/10.5281/zenodo.18421739).

Funding: This work was supported in part by the Agricultural Research Service, United States Department of Agriculture (USDA-ARS project numbers 5030-32000-231-000D to BA and TKA; and 5030-32000-231-095-S to OE); the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services (Contract No. 75N93021C00015 to TKA); and the SCINet project of the USDA-ARS (USDA-ARS project number 0500-00093-001-00-D to TKA). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

Competing interests: The authors have declared that no competing interests exist.

of annotated images, outperforms state-of-the-art models and can efficiently annotate thousands of images much more quickly than a human. Models trained by the proposed algorithm will serve as a standardized approach for pathologists and disease researchers to train automated image segmentation pipelines for large-scale histopathology.

Introduction

Semantic image segmentation, or pixel-level classification, is the task of partitioning an image into distinct regions or segments, each representing different objects or areas of interest. In recent years, deep neural networks trained on large, expert-annotated datasets have significantly advanced this field, achieving near-human performance as dataset sizes have increased [1,2]. Semantic image segmentation has been successfully applied to domains such as automated driving [3], intelligent medical technology [4–6], image search engines [7,8], industrial inspection [9–11], and augmented reality [1].

Despite its success, semantic image segmentation has seen limited application in histopathology, a critical discipline for understanding host-pathogen interactions and immune responses. Histopathology is a specialized technique that involves microscopic examination of tissue samples to investigate and analyze disease mechanisms. However, histologic evaluation requires years of specialized training and costly equipment. Consequently, only a fraction of collected specimens is analyzed, with vast amounts of archival material remaining underutilized. Moreover, the high costs associated with image annotation create bias and accuracy challenges [12,13], limiting the effectiveness of machine learning-based analysis pipelines.

Active Learning (AL) offers a promising approach for facilitating access and utilization of archived images to train AI models, where informative subsets are sampled from a pool of unlabeled data for annotation by human experts, with the goal of selecting the fewest samples to learn a robust neural network model [14,15]. AL heuristics quantify the information within a sample by prediction uncertainty [16–22]. AL has produced highly accurate models in classification tasks with data scarcity due to high data annotation costs [23,24].

However, AL is challenging to apply in image segmentation [25–28] due to two key factors: 1) sample selection and 2) uncertainty quantification. Individual pixels in an image must be classified and may carry varying degrees of uncertainty [26,28]. Images containing only a few pixels with high uncertainty can skew sample selection and confound sample ranking, resulting in biased and suboptimal models. Furthermore, classical neural network models, also called point-estimate models, cannot differentiate between uncertainty due to ambiguity in the data-generating process (aleatoric uncertainty) or lack of information on the data-generating process (epistemic uncertainty). Epistemic uncertainty arises due to a lack of understanding and can be reduced by acquiring more data. In comparison, aleatoric uncertainty is due to inherent noise and is hence irreducible.

To address these challenges, we propose SmartHisto, a novel sample selection heuristic designed for the semantic segmentation of histology whole-slide images. SmartHisto trains a Bayesian Neural Network (BNN) [29,30] instead of a point-estimate model to improve uncertainty quantification and identify the individual structures that are present in each image.

Related Work. AL is a methodology for achieving high model performance with minimal labeled data by iteratively selecting the most informative samples for annotation. Computationally, AL is a set selection problem, where in each iteration, a subset of all available data points should be selected to maximize some measure of information gain. Such informative samples are sequentially selected, annotated by an oracle, and then used to train a machine learning model, such as a Neural Network. AL has been tackled using uncertainty-based [25,28,31,32], diversity-based [33], and expected model change-based approaches [27,34]. These three methods have previously been integrated within several deep learning models.

While AL has been used extensively in classification [23,24] where the model attempts to assign a single label for a sample, application of AL in image segmentation using classical point-estimate model [25–28] remains challenging. Unlike point-estimate models, BNNs [29,30] provide a more robust measure of uncertainty that differentiates between aleatoric and epistemic uncertainty. BNNs sample model parameters from distributions rather than use individual fixed values; BNNs can better identify and quantify sources of uncertainty than classical Neural Networks [35]. The capability to distinguish novel from familiar information makes the application of BNNs in AL frameworks more attractive for complex tasks such as semantic image segmentation.

Our Contribution. This article proposes SmartHisto, a novel sample selection heuristic for semantic image segmentation using BNNs. SmartHisto differentiates between *aleatoric* and *epistemic* uncertainty and employs superpixeling [36] to cluster pixels into superpixels based on the measured uncertainty. A novel divergence-based uncertainty metric is then computed to rank samples according to their superpixels, optimizing sample selection and streamlining label acquisition.

Our experimental studies on a standardized benchmark dataset and a specially curated lung pathology dataset indicate that our proposed sampler produces superior models to traditional samplers while using fewer training samples. Finally, our comparison study indicates that models trained using SmartHisto outperform current state-of-the-art (SOTA) image segmentation models, such as QBC [37] and DEAL [38], without compromising memory efficiency.

The structure of this paper is as follows. The Materials and methods section formalizes the architecture of the BNN used for image segmentation in conjunction with algorithms to quantify uncertainty and select samples for annotation. The Results section presents the metrics used to evaluate the proposed method, an experimental setup with publicly available benchmarking datasets, and a comparison study that contrasts our approach with current state-of-the-art AL algorithms. Finally, the Discussion section concludes the article with future directions and potential applications.

Materials and methods

The proposed approach consists of two significant state-of-the-art advances: model initialization, referred to as *learner*, and uncertainty estimation. The BNN structure is described in Section Bayesian Neural Network Architecture, where the parameters of a regular point-estimate model are replaced with distributions from which parameters are sampled during inference. To train a BNN for semantic image segmentation, a loss function that capitalizes on the dual-parameter initialization is formalized in Section Evaluating Loss in a Bayesian Neural Network. Next, a heuristic to approximate the pixel-wise uncertainty is introduced in Section Estimating Pixelwise Uncertainty. Finally, to address the challenge of annotating individual pixels, a superpixeling method is described in Section Selecting Regions for Annotation.

Ethics statement

The pig tissues used in this study were collected from animals housed in the research facilities at the National Centers for Animal Health campus (Ames, IA, USA); the study protocol and animals were cared for in compliance with the guidelines

established by the Institutional Animal Care and Use Committee of the National Animal Disease Center, Agricultural Research Service, U.S. Department of Agriculture (USDA-ARS-NADC, protocol number ARS-22-1001).

Bayesian neural network architecture

The target learner architecture used in this study is based on the UNet [39] architecture. This approach replaces the regular point estimate initialization used by AL paradigms with parameters, each sampled from Gaussian distributions. The approach uses a smaller version of the traditional UNet architecture with weights drawn from independent Gaussian distributions.

BNNs replace the deterministic network's weight parameters with distributions over these parameters. Instead of optimizing the network weights directly, our approach optimizes the network using the marginalized weight distribution. The BNN is denoted by $\mathcal{M}$, and the model prediction for some input image x is given by $\mathcal{M}(x)$. The model likelihood is defined as $p(y|\mathcal{M}(x))$. Given a dataset with m samples with input features, X , and target labels, Y , Bayesian inference computes the posterior distribution over the weights given by $p(w|X, Y)$.

A distribution-driven initialization results in a learner that behaves as an ensemble of models trained with the same dataset, i.e., an ensemble of models with identical priors but different posterior weight distributions. This form of parameter initialization results in a more reliable measure of uncertainty as the ensemble of models are highly similar; In this setting, computing the uncertainty of a prediction is equivalent to the variance in prediction per pixel.

The target learner is trained using Bayes by Backprop [40], where each parameter $w \sim \mathcal{N}(w_\mu, w_\sigma^2)$ is given by a pair (w_μ, w_ρ) , where w_σ is computed from w_ρ as given below.

$$w_\sigma = \log(1 + w_\rho), \quad (1)$$

Following the weight initialization, the weights and biases of the target learner are given by:

$$w = w_\mu + \log(1 + w_\rho) \times \epsilon \quad (2)$$

$$b = b_\mu + \log(1 + b_\rho) \times \epsilon \quad (3)$$

where ϵ is added Gaussian noise, i.e., $\epsilon \sim \mathcal{N}(0, 1)$.

The initialization for all convolution layers follows the methodology described in [41], where a single convolution layer is equivalent to two sequential convolution operations.

The final output layer uses a sigmoid activation to compute individual class probabilities for each pixel. This approach allows the learner to better estimate the uncertainty of a superpixel, thereby identifying samples in the unlabeled pool of images from which to train. The proposed formulation of the target learner allows efficient characterization of the different types of uncertainty, discussed further in Section Estimating Pixelwise Uncertainty.

Evaluating loss in a Bayesian neural network

Various methods exist to compute the loss in semantic segmentation. Common approaches include Mean Intersection over Union (mIoU) [42,43], Jaccard Similarity [44,45], and Tversky Loss (TI) [46,47]. We use Dice Binary Cross entropy (DiceBCE) [44] to evaluate the quality of the predicted mask, whose formulation accommodates the hierarchical classes frequently seen in histopathology. DiceBCE is computed by adding the binary cross entropy loss with the *Dice Loss*, which is calculated as:

$$Dice\ loss = 1 - \frac{1}{m} \sum_{i \in X} \left(\frac{2 \times TP_i + s}{2 \times TP_i + FP_i + FN_i + s} \right), \quad (4)$$

where s is a smoothing parameter and m is the number of samples in the dataset/mini-batch. FP_i , FN_i , TP_i , and TN_i denote the number of false positives, false negatives, true positives, and true negatives in the learner's prediction of sample i in X , respectively.

The *Binary Cross entropy loss (BCE)* is given by:

$$BCE = \frac{1}{m} \sum_{i=1}^m (p_i \log p_i + (1 - p_i) \log (1 - p_i)), \quad (5)$$

where m is the number of samples.

The final *DiceBCE* loss is given as the sum of [Eq \(4\)](#) and [Eq \(5\)](#):

$$DiceBCE = Dice \text{ loss} + BCE$$

The target learner approximates the optimal posterior distribution of its parameters through variational inference by maximizing the log evidence lower bound. The learner, denoted by $\mathcal{M}$, estimates the output y for a given input x . The initial parameters (w_μ, w_ρ) represent the prior Bayesian distribution over the true function $\mathcal{M}$ that generated the data. The likelihood of the prediction is given by $p(y|\mathcal{M}, x)$. The parameters of $\mathcal{M}$ are estimated as $p(\mathcal{M}|X, Y)$, where X and Y are sets of inputs and outputs, respectively. Hence, the likelihood of the data is given by,

$$p(y|x; \mathcal{M}) = \int p(y|\mathcal{M})p(\mathcal{M}|x, X, Y)df \quad (6)$$

[Eq \(6\)](#) is intractable as an infinite number of parameters can be used to model $\mathcal{M}$. Hence, the parameters of $\mathcal{M}$ are approximated by introducing a parameter w , which simplifies the problem of finding the best possible parameters.

$$p(y|x; \mathcal{M}) = \int p(y|\mathcal{M})p(\mathcal{M}|x, w)p(w|X, Y)d\mathcal{M}dw \quad (7)$$

Here, the true distribution $p(w|X, Y)$ remains intractable, and is replaced with a variational distribution $q(w)$, estimated using variational inference [\[48,49\]](#). This distribution should be close to the posterior weight distribution of the true function $\mathcal{M}$, given by $p'(w)$. Kullback-Leibler (KL) Divergence [\[20,30,50,51\]](#) measures the distance between two distributions, which estimates the loss between the current and target distributions.

Prior approaches have suggested finding a variational approximation to the Bayesian posterior distribution on the weights [\[52\]](#). Variational learning finds the parameters θ of a distribution on the weights $q(w|\theta)$ that minimizes the KL Divergence of the true Bayesian posterior on the weights.

$$KL[q(w|X, Y)||p(w)] = \int q(w|X, Y) \log \frac{q(w|X, Y)}{p(w)} dw \quad (8)$$

The proposed method does this by sampling weights from the estimated distribution $q(w|X, Y)$ rather than the true posterior $p(w)$. In this approach, where the model parameters are sampled from Gaussian distributions, the KL Divergence can be computed as follows:

$$KL(q, p) = \int q(w|X, Y) \log \frac{q(w|X, Y)}{p(w)} dw$$

In the case of normal distribution, we can bind the range to $(-\infty, \infty)$.

$$\begin{aligned} \text{KL}(q, p) &= \int_{-\infty}^{\infty} q(w) \log q(w) dw \\ &\quad - \int_{-\infty}^{\infty} q(w) \log p(w) dw \end{aligned} \quad (9)$$

$$= \log \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} - \frac{1}{2} \quad (10)$$

The final loss functions combine the DiceBCE and a scaled KL divergence, resulting in the following:

$$\text{Loss} = \text{DiceBCE} + \kappa * \text{KL}[q(w|\theta)||p(w)], \quad (11)$$

where κ is a scaling factor for the KL divergence.

The pipeline we introduce trains the target learner for a given dataset X, Y following Algorithm 1.

Algorithm 1: Model Train

Input: labeled pool of images L_0 , Bayesian model $\mathcal{M}$, Epochs E , scaling factor κ

Output: Trained Learner

```

1: for epoch in  $E$  do
2:    $\mathcal{M}.\text{freeze\_mu}()$ 
3:   for  $\mathbf{b}$  in  $L_0$  do
4:      $\mathbf{x}, \mathbf{y} \leftarrow \mathbf{b}$ 
5:      $\bar{\mathbf{y}} \leftarrow \mathcal{M}(\mathbf{x})$ 
6:      $\text{loss}_m \leftarrow \text{DiceBCE}(\mathbf{y}, \bar{\mathbf{y}})$ 
7:      $\text{loss}_{kl} \leftarrow \text{KL}(\text{model.parameters})$ 
8:      $\text{loss} \leftarrow \text{loss}_m + \kappa * \text{loss}_{kl}$ 
9:    $\mathcal{M}.\text{update\_params}(\text{loss})$ 

```

Estimating pixelwise uncertainty. There are two forms of uncertainty: Aleatoric uncertainty - uncertainty due to the inherent ambiguity of the data; and Epistemic uncertainty - uncertainty due to insufficient field knowledge. Aleatoric uncertainty cannot be reduced by adding more data to the training set, as this is a form of uncertainty caused by the inherent nature of the data. In contrast, Epistemic uncertainty can be addressed by including more samples with higher uncertainty. In Bayesian modeling, not only can the uncertainty of a sample be computed, but Aleatoric and Epistemic uncertainty can also be differentiated using the predictive variance of a sample [18,21,22,41].

In classification tasks, finding the total predictive variance to estimate the uncertainty of the sample can be useful. However, current approaches cannot be used directly in image segmentation; localization of uncertainty is not captured in current methods, which is necessary for image segmentation. In most cases, providing an image with high uncertainty to the oracle annotating images is the most useful. As such, sections with high uncertainty are highlighted for manual labeling, hence reducing the overall workload of manual annotation.

The learner's prediction performance on an unseen sample is computed as the average expectation given the trained learner posterior. Using the equation decomposition in [17], Epistemic uncertainty can be quantified in the image, which is then mapped to its corresponding superpixel. This produces a heatmap of uncertain sections that is then provided to the annotator for labeling. The final labeled image is then added to the labeled pool of images for the next active iteration.

The predictive variance can be decomposed on any new input $\mathbf{x}^*$ as:

$$\begin{aligned} \text{Var}(p(y^*|x^*)) &= E[yy^T] - E[y]E[y^T] \\ &= \frac{1}{T} \sum_{t=1}^T \text{diag}(\hat{p}_t) - \hat{p}_t \hat{p}_t^T \end{aligned} \quad (12)$$

$$+ \frac{1}{T} \sum_{t=1}^T (\hat{p}_t - \bar{p})(\hat{p}_t - \bar{p})^T, \quad (13)$$

where $\bar{p}$ is the average probability of the class, and T is the size of the active batch.

[Eq \(12\)](#) characterizes the Aleatoric uncertainty and [Eq \(13\)](#) characterizes the Epistemic uncertainty. The method makes use of differentiation between uncertainties caused due to the inherently noisy nature of the data and uncertainty caused due to the learner itself, allowing the learner to accurately find samples in the unlabeled pool of images that carry the most information. The method uses these superpixels to map the computed uncertainties, taking the average divergence of pixel predictions within each superpixel, and in doing so, provides a significant advantage: an objective metric that prioritizes groups of uncertain pixels over single uncertain pixels, maximizing information gain at each active iteration.

The Divergence score for a single superpixel is given by:

$$\text{Div}(S) = \sum_{i,j \in S} KL(P_i || P_j) \quad (14)$$

$$= \frac{1}{|S|} \sum_{i,j \in I} (P_i \log \frac{P_i}{P_j}), \quad (15)$$

where S is a superpixel.

This uncertainty metric uses both forms of uncertainty while evaluating images, fractionally weighting Aleatoric uncertainty to prioritize reducing Epistemic uncertainty. Further, the average superpixel divergence is computed for each unlabeled image while sampling the unlabeled pool of images. When an image is presented to the annotator, the individual superpixels are highlighted based on their respective divergence score, as shown in [Fig 1](#).

The total uncertainty for an unlabelled sample is computed as an average over L independent predictions, as shown in [Eq \(16\)](#)

$$\begin{aligned} \text{Uncertainty}(I) &= \frac{1}{L} \sum_{l=1}^L (\hat{p}_l - \bar{p})(\hat{p}_l - \bar{p})^T \\ &+ \alpha * \frac{1}{L} \sum_{l=1}^L (\text{diag}(\hat{p}_l) - \hat{p}_l \hat{p}_l^T) \\ &+ \frac{1}{|S|} \sum_{S \in I} \text{Div}(S), \end{aligned} \quad (16)$$

where α is a scaling factor for Aleatoric uncertainty, $\text{Div}(I)$ as described in [Eq \(14\)](#), S is a superpixel in the input sample, and I is an input sample.

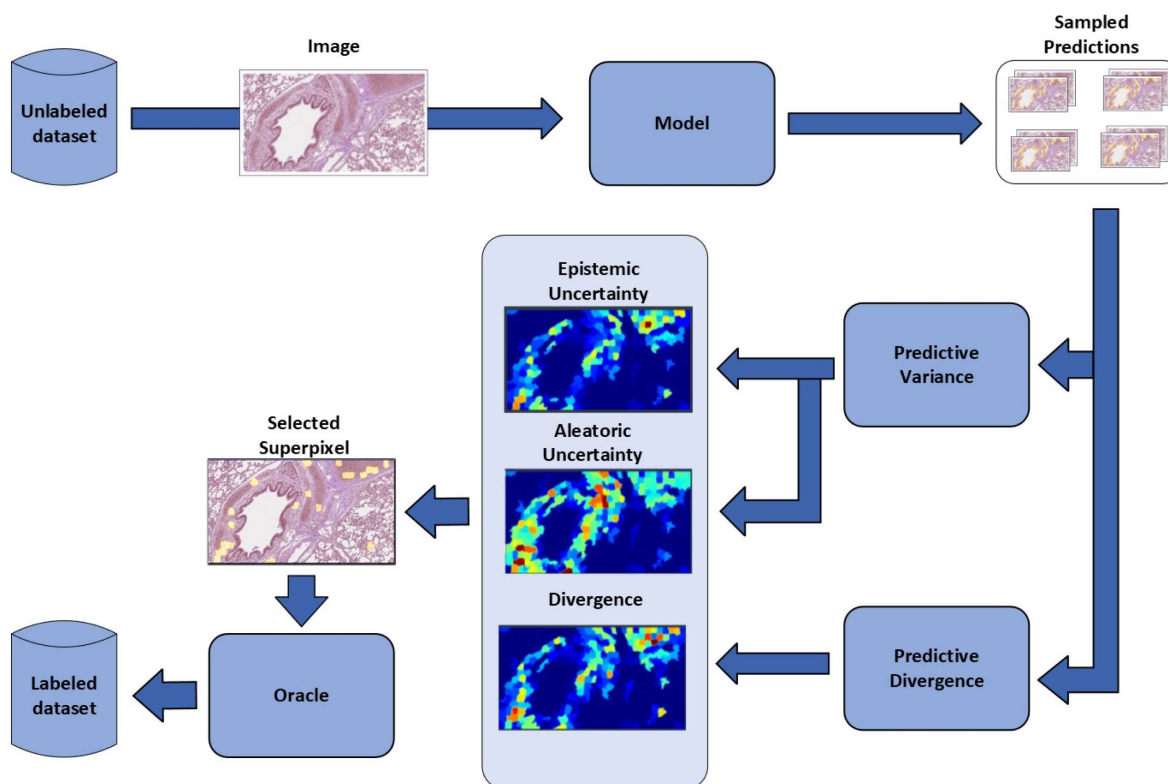


Fig 1. Active Learning Pipeline. Starting from the top left, an image is randomly selected from the pool of unlabeled images. The learner then attempts to segment the image, returning a confidence score on each image pixel. The confidence scores' Predictive Variance and Predictive Divergence are then computed and used to find the Epistemic uncertainty, Aleatoric uncertainty, and Divergence in prediction. These metrics find the most informative superpixels in the image and provide them to the oracle. The oracle annotates the image and adds the image to the pool of labeled images.

<https://doi.org/10.1371/journal.pcbi.1013611.g001>

Selecting regions for annotation. The regions of uncertainty are often sparsely found throughout a single image. They appear near the boundaries of objects and only in small, dense hotspots. Small sections spread out over a large image make annotating the whole image unnecessarily cumbersome, as relabeling an entire image is slow. Popular approaches tackle this problem by using superpixeling, a clustering heuristic that groups pixels into unique partitions based on local similarity. Using superpixels allows human annotators to minimize labeling costs by annotating the image by uniform regions instead of individual pixels. We employ SLIC [36] to find the superpixels of an entire image. We empirically identified that 1000 segments with a compactness of 28 performed well for all datasets tested. These default settings are likely to provide good results for many users and datasets but may require refinement to generate better results.

The pipeline trains a BNN to perform semantic segmentation on a given corpus following Algorithm 2. Each iteration, defined here as an *active epoch*, begins with the initialization of a BNN, whose starting parameters are fixed for an active training loop and is trained on the most recently updated pool of labeled images. The trained learner then samples the unlabelled pool of images for the most informative samples. The chosen samples are provided to the annotator with clear markings indicating where annotations are required [53], reducing the cost of labeling significantly as annotation is only necessary for highlighted regions.

Algorithm 2: SMART

Input: labelled pool of images L_0 , unlabelled pool of images U_0 , validation pool of images V_0 , BNN $\mathcal{M}$, number of active iterations $\mathbf{A}$, active sample size t

Output: Trained Model

```

1: for  $\mathbf{a}$  in  $\mathbf{A}$  do
2:    $\mathcal{M}_t \leftarrow \text{TrainModel}(\mathcal{M}, L_0)$ 
3:    $A_0 \leftarrow \text{SamplePool}(\mathcal{M}_t, U_0, t)$ 
4:    $L_0 \leftarrow L_0 \cup A_0$ 
5:    $U_0 \leftarrow U_0 \setminus A_0$ 

```

Results

This section presents the experimental results of the proposed AL paradigm and outlines the testing methodology employed. Section Datasets details the protocols used in the curation of the benchmark datasets and the custom pulmonary histology dataset. Subsequently, we describe image augmentation techniques used to introduce more variation on the training data. In Sections Experimental Settings and Comparison Study we describe the computational hardware used to test a model using SmartHisto and SOTA methods it was compared against, and the standardized benchmark datasets used in the comparison, respectively. The source code for the pipeline is available in the project repository at https://github.com/sriram98v/histology_segmentation.

Experimental settings

The AL paradigm and other comparative methods were tested under the proposed sampling heuristic with a modified Dice Coefficient loss layer. Images of different sizes were rigidly registered to a reference image size of 256×256 . A Bayesian-UNet was trained end-to-end, with cost minimization on 100 epochs performed using an Adam optimizer [54] with an initial learning rate of 0.01. An initial label pool size of five percent was used for the first active epoch, with increments of five percent for each active epoch. The training time for this network was approximately 2 hours on a workstation with NVIDIA GeForce RTX 3090 GPU. The output of the last convolutional layer with sigmoid non-linearity consisted of a probability map for each class. Pixels with computed probabilities of 0.6 or more were considered to belong to the respective output class of the corresponding dataset. This threshold was selected to be deliberately conservative but may be amended for other datasets.

Datasets

The comparative study uses the publicly available histology dataset GlaS [4]. Further, we also include a pulmonary histology dataset used in the ablation study of the proposed approach, as the benchmark datasets are binary segmentation datasets, not multi-class datasets, better reflecting semantic image segmentation in pathology.

Glas dataset. The GlaS dataset included 167 images from 16 H&E stained histological sections of stage T3 or T42 colorectal adenocarcinoma. Each section belonged to a different patient, and sections were processed in the laboratory on different occasions. Thus, the dataset exhibits high inter-subject variability in stain distribution and tissue architecture. The digitization of these histological sections into whole-slide images (WSIs) was accomplished using a Zeiss MIRAX MIDI Slide Scanner with a pixel resolution of $0.465 \mu\text{m}$. Each image had a dimension of 775×552 ; for the comparative study, each image was uniformly registered to 256×256 .

Pulmonary histopathology dataset. The Pulmonary Histology dataset included 235 training and 129 testing tissue images, all of resolution 512×512 pixels containing manually annotated structures and cell types using LabelMe [55]. Histologically normal lung tissue originated from clinically healthy, negative control pigs housed at the National Animal Disease Center (NADC) following animal use protocols approved by the NADC. Each image in this dataset was extracted using Aperio ImageScope from one of four individual WSI (scanned at 400x) scanned at a $\times 40$ equivalent resolution on a

Leica Aperio AT2. Four individual WSI were considered sufficient due to the low variability observed across histologically normal lung tissue. The labeled data included respiratory epithelium, alveolar septa, cartilage, and smooth muscle. The images were collected from different experiments with and without image color management within Aperio ImageScope, which introduced sources of variation due to the differences in the staining practices and image scanners across slides. Representative 512×512 sub-images from nuclei dense areas were extracted from WSIs to reduce the computational burden of processing multiple WSIs. The dataset includes one crop per WSI and animal to ensure diversity. Both epithelial and stromal nuclei were manually annotated in the 1000×1000 sub-images using Aperio ImageScope. In order to use the WSIs in the proposed pipeline with limited memory, ten sub-images, and their corresponding annotations were randomly sampled from each of the images at a resolution of 250×250 pixels, resulting in 233 training images and 100 testing images.

TIGER ROI dataset. The TIGER ROI dataset [56] consists of 153 region-of-interest (ROI) tiles extracted from whole-slide images of breast cancer tissue, annotated as part of the Tumor Infiltrating lymphocytes in breast cancer (TIGER) challenge. Each tile contains pixel-level annotations for six tissue classes: tumor epithelium, stroma, necrosis, normal mucosa, adipose tissue, and muscle. Masks were provided as RGB color-coded images and converted to per-class binary channels during loading. All images were resized to 256×256 pixels for training. The dataset exhibits significant class imbalance: tumor and stroma collectively account for the majority of annotated pixels, while necrosis and muscle are present in fewer than 5% of pixels each. To address this imbalance, median-frequency class weighting [57] was applied to the DiceBCE loss, with per-class weights recomputed at each active learning iteration from the current labeled pool. The dataset was partitioned into a fixed validation split of 20% (31 images) and an active learning pool of 80% (122 images). Starting from an initial labeled set of 24 images ($\approx 20\%$ of the AL pool), 8 images were queried per active learning iteration until the pool was exhausted, yielding 13 iterations in total.

Image augmentation. Each image was subject to a series of independent random transforms. All samples from the training dataset underwent a set of transforms randomly applied to the image with a probability of 0.5. Variable-degree flips, Gaussian noise, and Gaussian blur were employed to augment the images, applied in series to each image before training. Traditionally, color channel shifts and swaps are also employed in image augmentation. However, we do not employ any color alterations in image augmentation, as such alterations to input images may obscure the impact of histopathology tissue staining, hindering the performance of the final model. As SmartHisto samples superpixels by uncertainty, stain variability in image training is captured as informative areas for the model, thereby generating a more robust and generalizable model. Alterations to color channels would negatively impact generalization by obscuring staining variations, thereby affecting downstream sampling of uninformative regions.

Comparison study

The learner performance on each of the previously listed datasets are presented along with the performance of other benchmark models. The masks predicted by our learner on each of the datasets is displayed in their corresponding sub-sections. Our method was compared to some benchmark methods such as softmax entropy sampling (**ENT**), Margin sampling (**MAR**) and Random Sampling (**RAND**), and SOTA methods such as **QBC** [37] and **DEAL** [38]. The performance of each of the sampling methods was evaluated as the average performance over 10 full runs of the active sampling method over a full dataset, where for each run, the starting parameters of the learner were drawn from independent and identical distributions. The peak mIoU, peak memory utilization, and average runtime of a single active epoch for each method is summarized in [Tables 1–3](#).

Performance on glas dataset. The comparative results of our method against benchmark methods using the Glas dataset is plotted in [Fig 2](#). Our learner reaches a better performance than the benchmarks. The learner trained using our sampling approach outperforms the benchmark methods starting from the fifth active epoch, achieving an mIoU of ≈ 0.75 at the end of training. [Fig 3](#) shows the predictions, uncertainty maps, and selected superpixels computed per label class. A

Table 1. Runtime, Peak Memory, and Peak mIoU of each method on the Pulmonary dataset.

Method	Runtime	Peak Memory	Peak mIoU
SmartHisto	4.43 min	16.89 Gb	0.6287
DEAL	5.02 min	17.59 Gb	0.3974
QBC	4.53 min	18.01 Gb	0.4107
ENT	4.53 min	18.01 Gb	0.3130
MAR	4.53 min	18.01 Gb	0.1633
RAND	4.53 min	18.01 Gb	0.4539

ENT, MAR, QBC, and RAND share the same underlying architecture and therefore have identical runtime and memory usage.

<https://doi.org/10.1371/journal.pcbi.1013611.t001>

Table 2. Runtime, Peak Memory, and Peak mIoU of each method on the GLaS dataset.

Method	Runtime	Peak Memory	Peak mIoU
SmartHisto	2.36 min	10.62 Gb	0.7466
DEAL	3.14 min	11.04 Gb	0.6472
QBC	2.89 min	11.46 Gb	0.6418
ENT	2.89 min	11.46 Gb	0.6193
MAR	2.89 min	11.46 Gb	0.6036
RAND	2.89 min	11.46 Gb	0.5593

ENT, MAR, QBC, and RAND share the same underlying architecture and therefore have identical runtime and memory usage.

<https://doi.org/10.1371/journal.pcbi.1013611.t002>

Table 3. Runtime, Peak Memory, and Peak mIoU of each method on the TIGER ROI dataset.

Method	Runtime	Peak Memory	Peak mIoU
SmartHisto	27.3 min	11.9 Gb	0.4181
DEAL	2.9 min	10.95 Gb	0.3175
QBC	27.9 min	11.9 Gb	0.3859
ENT	28.1 min	11.9 Gb	0.3754
MAR	26.5 min	11.9 Gb	0.3438
RAND	28.0 min	11.9 Gb	0.3964

ENT, MAR, QBC, and RAND share the same underlying architecture and therefore have identical runtime and memory usage.

<https://doi.org/10.1371/journal.pcbi.1013611.t003>

key behavior to be noted here is that in the Bayesian settings, the learners consistently identify the parts of the image that are always found to be a part of the gland (the outer ring) with both high accuracy and low uncertainty. The inner segment of each gland is highly variable and is detected as such by the learner as an uncertain area of the image. This uncertainty is highlighted in the uncertainty map produced by the learner. Such identification of the variable and consistent patterns in the input images is a common trend seen in all of the following experiments.

Performance on the pulmonary histology dataset. The comparative results of our method against the previously listed benchmark methods is plotted in Fig 4. Our method reaches a better performance than the other benchmarks. The predictions by each of the learners trained with each of the sampling techniques is displayed in Fig 5. The predictions for each target class by each of the methods is listed with their corresponding class name. Fig 6 shows the uncertainty maps computed per label class. Note that there is a high concentration of uncertainty on the border of each class map rather than the center. This is likely due to the variability in annotation accuracy from a human annotator.

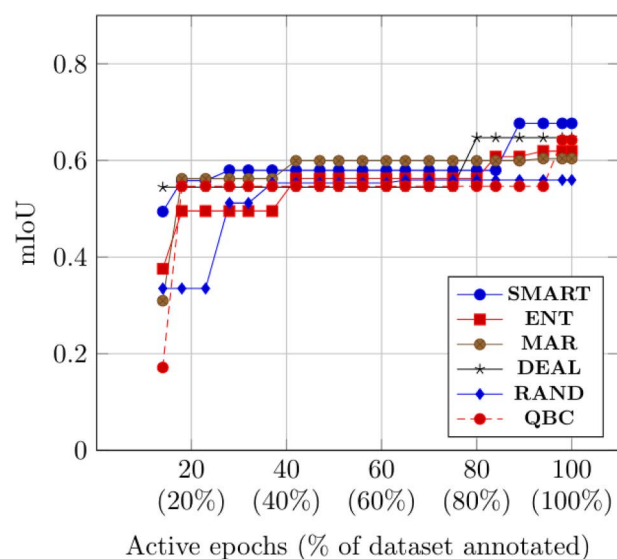


Fig 2. Model Performance after training on each of the sampling methods on the GLaS dataset. Performance of trained using each of the benchmark methods. Our proposed method outperforms the state-of-the-art methods in each consecutive active iteration.

<https://doi.org/10.1371/journal.pcbi.1013611.g002>

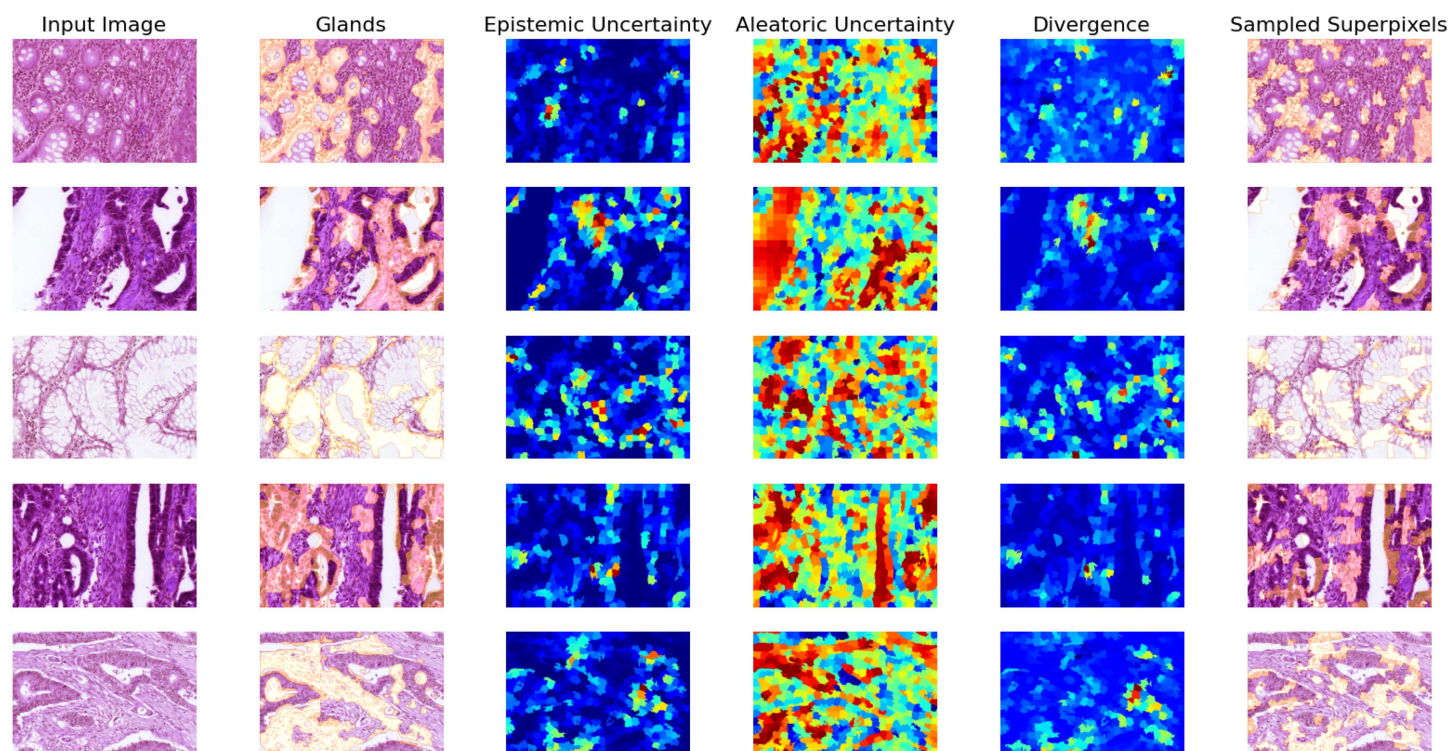


Fig 3. Predictions, uncertainty maps, and sampled superpixels computed by the learner trained with the proposed sampling technique on the GLaS dataset. Starting from the leftmost column, each row represents the input image, predicted mask, superpixel-wise epistemic uncertainty, superpixel-wise aleatoric uncertainty, and superpixel-wise divergence by SmartHisto.

<https://doi.org/10.1371/journal.pcbi.1013611.g003>

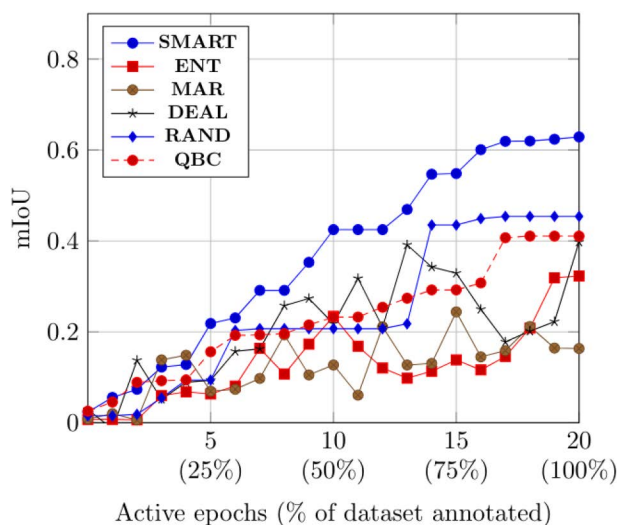


Fig 4. Model Performance after training on each of the sampling methods on the Pulmonary Histology dataset. Performance of trained using each of the benchmark methods. Our proposed method outperforms the state-of-the-art methods in each consecutive active iteration.

<https://doi.org/10.1371/journal.pcbi.1013611.g004>

Performance on the TIGER ROI dataset. The comparative results of our method against the benchmark methods on the TIGER ROI dataset are plotted in Fig 7. The dataset presents a substantially harder segmentation problem than GlaS or the pulmonary histology dataset, owing to its six-class multi-tissue structure and pronounced class imbalance (tumor and stroma together account for the majority of annotated pixels, while necrosis and muscle each contribute fewer than 5%). Median-frequency class weights were applied to the DiceBCE loss and recomputed at each active learning iteration from the current labeled pool.

SmartHisto achieves the highest peak mIoU of 0.4181 at the end of training (13 active iterations, 120 labeled images). Qualitative predictions on the three highest-scoring validation images are shown in Fig 8. RAND and QBC follow with peak mIoU values of 0.3964 and 0.3859, respectively, while ENT and MAR trail at 0.3754 and 0.3438. Notably, SmartHisto reaches competitive performance (>0.38 mIoU) earlier in the active learning curve than the baselines, demonstrating more sample-efficient label acquisition under severe class imbalance. The uncertainty-based single-model baselines ENT and MAR exhibit high variance across iterations, suggesting that entropy- and margin-based scores are less reliable query signals when class distributions are highly skewed.

Practical annotation efficiency. A key motivation for active learning in digital pathology is reducing the annotation burden on pathologists working with WSIs. To quantify this, we analyzed how many ROI tiles each method must have annotated before reaching a target segmentation quality of $\text{mIoU} \geq 0.35$. Results are summarized in Table 4.

SmartHisto reaches $\text{mIoU} \geq 0.35$ after annotating only 53 tiles (44% of the 122-tile active learning pool), compared with 60 tiles for RAND (+13% more annotations) and 65 tiles for ENT (+23%). MAR and DEAL never reach this threshold within the available budget. Critically, SmartHisto's peak mIoU of 0.4181 is not matched by any baseline at any annotation level.

Ablation study

In this section, the results of an ablation study on the parameters of our sampling method using the pulmonary histology dataset described previously in Section Pulmonary Histopathology Dataset are provided with a comparison against the baseline sampling methods **RAND**, **ENT**, and **MAR**.

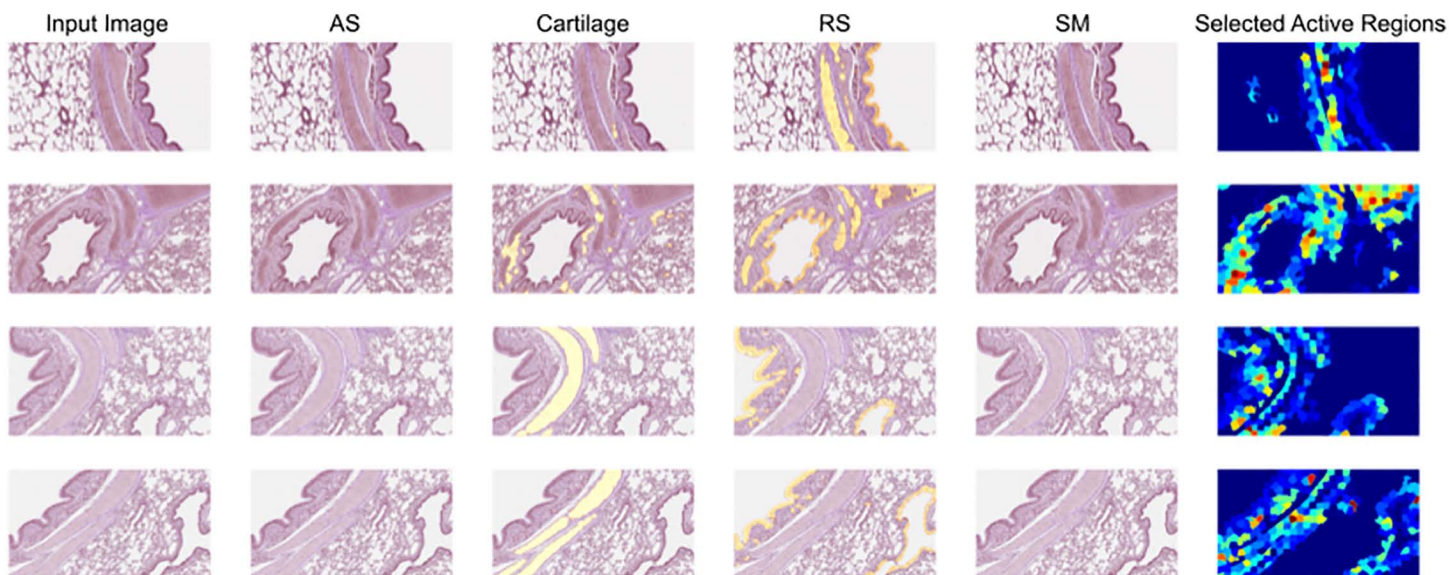


Fig 5. Model Predictions after training on each of the sampling methods on the Pulmonary Histology dataset. Starting from the leftmost column, each row represents the input image, Ground Truth (GT) binary mask, predicted mask by each of sampling methods as indicated by the column title.

<https://doi.org/10.1371/journal.pcbi.1013611.g005>

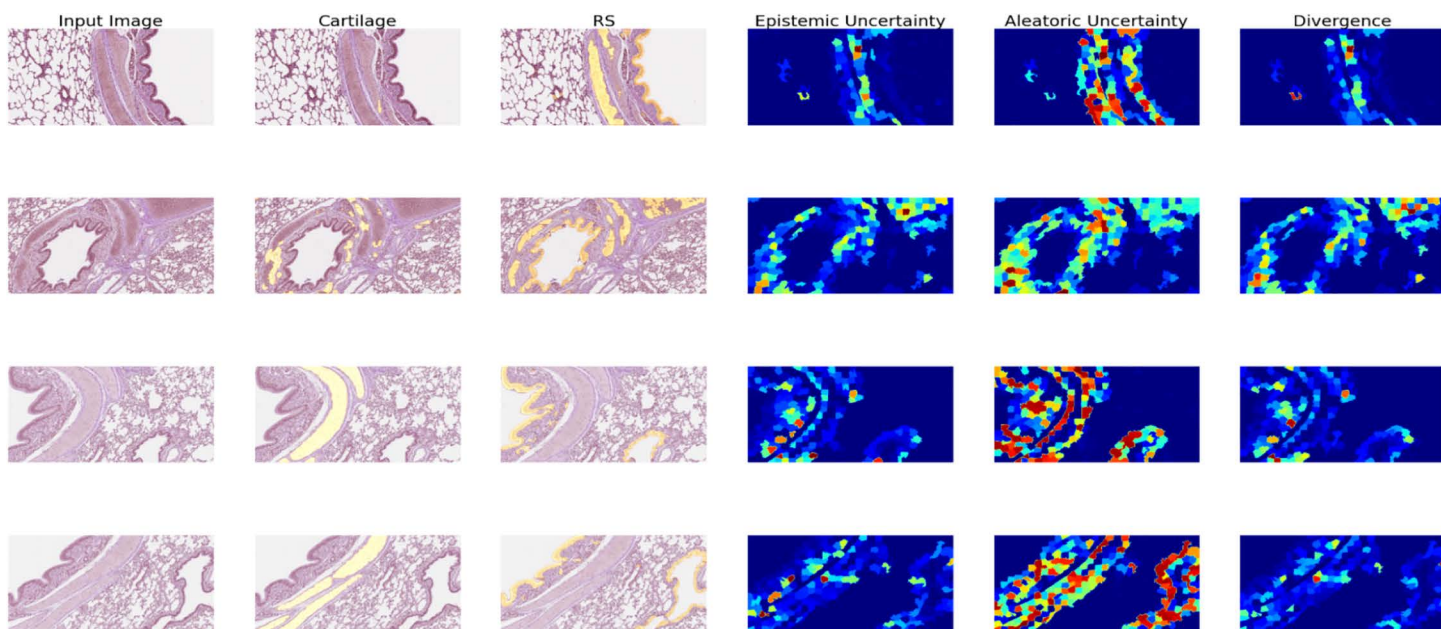


Fig 6. Uncertainty maps computed by the learner trained with the proposed sampling technique on the Pulmonary Histology dataset. The uncertainty maps for predicted class labels as computed by the SmartHisto approach. Note that the borders of each of the classes have a higher uncertainty as they are more variable in each of the input training images.

<https://doi.org/10.1371/journal.pcbi.1013611.g006>

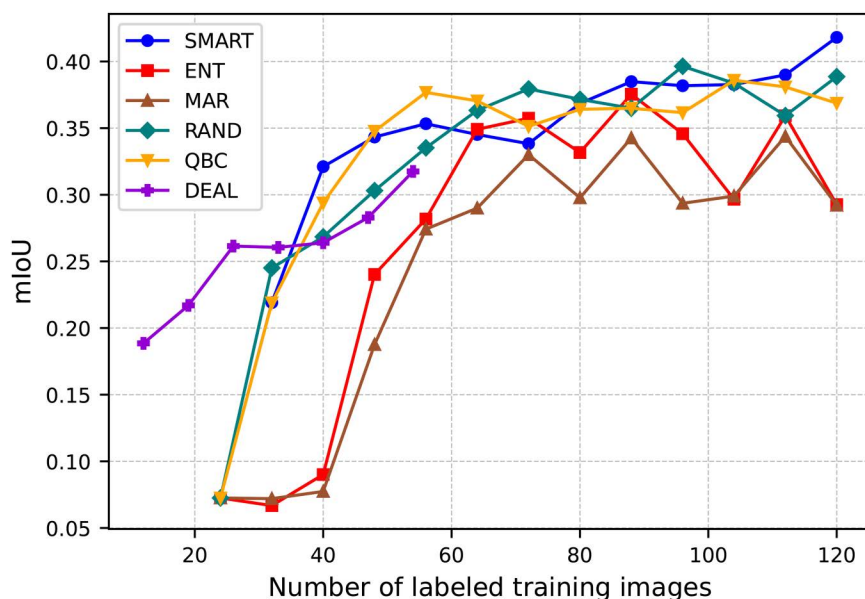


Fig 7. Model Performance after training on each of the sampling methods on the TIGER ROI dataset. Peak mIoU as a function of the number of labeled training images for each benchmark method. SmartHisto consistently achieves higher mIoU than the baseline sampling strategies across active learning iterations.

<https://doi.org/10.1371/journal.pcbi.1013611.g007>

Effect of KL weight. In order to show the effect of KL Divergence on the overall performance of the learner training within each active epoch, the learner performance was evaluated with a range of values for κ in Eq (11) in the loss function. A fixed training set and a distinct validation set are used to evaluate the learner performance and compare the loss of the learner during each training epoch along with the learner accuracy on the validation set. As shown in S1 Fig (a), the learner performances on four different values of $\kappa \in \{0, 1\}$ is plotted at each active epoch, with $\kappa = 1$ providing the best overall performance.

Effect of aleatoric uncertainty weight. In order to evaluate the contribution of aleatoric uncertainty to the active learning sampling scheme, the performance of the learner was evaluated with a range of different values for α in Eq (16) of the loss function. Full active training loops were used for each of 4 different values of $\alpha \in \{0, 1\}$ without including the divergence term described in Eq (14). The learner performance was evaluated at each active epoch and plotted for each value of α . S1 Fig (b) shows that $\alpha = 1$ displays the best overall learner performance.

Effect of divergence. To evaluate the contribution of divergence to the active learning sampling scheme, learner performance with and without the divergence term in Eq (16) were compared while setting $\alpha = 0$. The learner was run once on each setting for a full active learning cycle and was evaluated at the end of each active learning epoch. S1 Fig (c) shows the best learner performance with all the sampling terms.

Discussion

This paper introduces a novel Active Learning paradigm that uses a Bayesian approach to reduce annotation costs in image segmentation. The approach employs a deep Bayesian active learning method for semantic segmentation, leveraging a Bayesian Neural Network to identify areas of high uncertainty in images. By using an uncertainty metric to segment images, we were able to distinguish between poor classifications caused by deficient model knowledge and those due to the inherently noisy nature of the data. Following the annotation of a small set of histopathology slides by a subject-matter expert, we demonstrated that image segmentation and classification could be performed on hundreds to thousands of

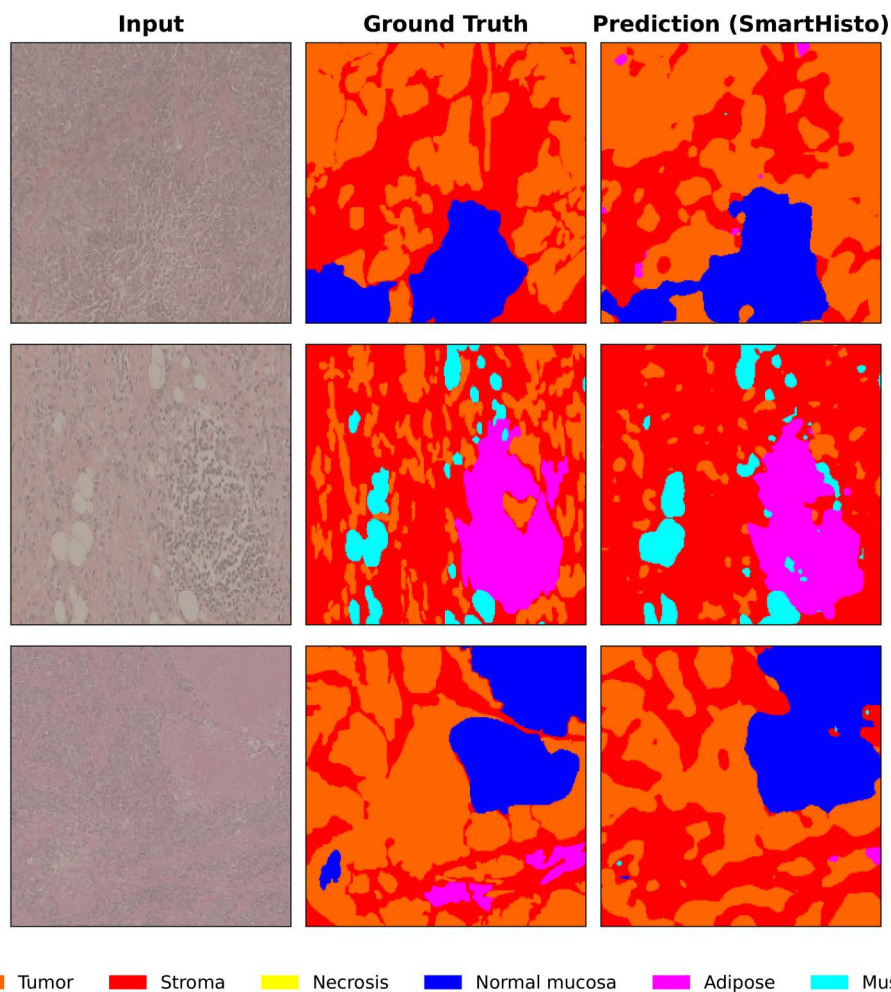


Fig 8. Predictions by the SmartHisto-trained model on the TIGER ROI dataset. Each row shows the input ROI tile (left), ground-truth annotation (centre), and the model prediction (right) for the four validation images with the highest per-image mIoU. Colours denote tissue class: ■ Tumor, ■ Stroma, ■ Necrosis, ■ Normal mucosa, ■ Adipose, ■ Muscle.

<https://doi.org/10.1371/journal.pcbi.1013611.g008>

Table 4. Annotation efficiency summary on the TIGER ROI dataset. Number of labeled tiles required to reach $\text{mIoU} \geq 0.35$ and annotation savings relative to SmartHisto. “-” indicates the method never reached the target within the 122-tile budget.

Method	Peak mIoU	Tiles	SmartHisto
SmartHisto	0.4181	53	0%
RAND	0.3964	60	+11.3%
ENT	0.3754	65	+17.7%
QBC	0.3859	49	−9.9%
MAR	0.3438	—	—
DEAL	0.3175	—	—

Savings are computed as $(n_{\text{other}} - n_{\text{SmartHisto}})/n_{\text{other}}$; positive values indicate SmartHisto needs fewer annotations. QBC reaches the 0.35 threshold marginally faster but plateaus below SmartHisto’s peak and never matches its final quality.

<https://doi.org/10.1371/journal.pcbi.1013611.t004>

images. Most pathology laboratories are equipped with the necessary digital pathology equipment [58,59], and integrating tools such as SmartHisto can facilitate automated analysis and identification of images of high utility for additional investigation and characterization.

We also developed an acquisition heuristic that reduces labeling effort for large distributed datasets. This approach has significant utility in histopathology, a field challenged by the high cost of accurately annotating training datasets. The method is robust to overfitting and enables faster model convergence through intelligent sample selection. Additionally, SmartHisto is sufficiently flexible to be applied to most neural network architectures for image segmentation, substantially reducing the number of samples that require accurate annotations.

We demonstrate consistent performance improvements across all datasets for semantic segmentation. More broadly, the proposed framework is applicable to a range of computer vision tasks, including instance segmentation, object localization, and pose estimation. A key advantage of our current approach is its ability to produce pixel-level uncertainty estimates. This contrasts with other state-of-the-art approaches, such as Deep ensembles [60] and Variational Adversarial Active learning [61], which provide uncertainty estimates only at the whole-image level. When evaluated on the Pulmonary dataset, this pixel-level uncertainty modeling led to clear performance gains over competing methods. On the GLaS dataset, improvements became evident only after the 80th training epoch. This discrepancy is likely attributable to annotations in the GLaS dataset that include regions of the glass slide absent in the Pulmonary dataset, introducing additional heterogeneity. Despite this, SmartHisto achieved a higher mIoU at the conclusion of active training, indicating its effectiveness when biologically relevant annotations are available. In practical terms, SmartHisto reached competitive performance using approximately 50% of the annotated data in the Pulmonary dataset, corresponding to a reduction in annotation requirements of about 50%. Similarly, competitive performance on the GLaS dataset was achieved using roughly 80% of the annotated data, yielding an annotation reduction of approximately 20%. In both cases, only small image fragments required annotation, substantially lowering the overall annotation burden. When deployed within large-scale slide archiving systems [58,59], this approach could translate into a dramatic reduction in expert effort by shifting from full-slide annotation to targeted, informative patches sampled across multiple WSIs, if deployed to analyze data within large-scale slide archiving, thereby avoiding redundant annotations and focusing labeling effort on regions exhibiting meaningful variability.

In future work, we will explore hierarchical losses and sampling functions for datasets with interdependent classes, varying image magnifications, and protocols that include different histologic stains. A current limitation in our approach is that the segmented classes are mutually exclusive, i.e., there is no overlap between any two classes. Relaxing this constraint may increase utility in biological datasets, where the distinction between classes may not always be perfect. Additionally, in our development, images from both the pulmonary and GLaS datasets were at relatively consistent magnifications. We adopted this constraint because datasets with varying magnifications may not generalize well. Further studies are needed to the impact of varying magnification. Finally, all training images were stained using an H&E protocol, which, while widely used in histopathology, represents only one staining modality. Understanding how alternative stains affect classification and SmartHisto's performance is necessary to broaden the applicability of our approach. Despite the current limitations, SmartHisto achieves competitive performance with substantially smaller annotated datasets, drastically reducing the annotation burden on domain specialists.

Supporting information

S1 Fig. Ablation study of the proposed method using the Pulmonary Histology dataset. (a) Ablation study of our loss function. A fixed training set and a distinct validation set were used to evaluate learner performance and compare the loss of the learner during each training epoch along with learner accuracy on the validation set. As shown in the plots, the tested and compared learner performances used four different values of $\kappa \in \{0, 1\}$ which was plotted at each active epoch, with $\kappa = 1$ providing the best overall performance. (b) Ablation study of our uncertainty metric: the learner's performance was evaluated with different values for α in Eq (16) of the loss function. Full active training loops were used for

each of 4 different values of $\alpha \in \{0, 1\}$ without including the divergence term described in [Eq \(14\)](#). The model performance was evaluated at each active epoch and plotted for each value of α . The plot shows that $\alpha = 1$ provides the best performance of the model. (c) Ablation study of divergence: model performance with and without the divergence term in [Eq \(16\)](#) are compared while setting $\alpha = 0$. The model was run once on each setting for a full active learning cycle and was evaluated at the end of each active learning epoch. The plot shows the best overall model performance with all the sampling terms.

(PDF)

Acknowledgments

We sincerely thank Dr. Li for their valuable comments and insightful discussions, which greatly improved this work.

Author contributions

Conceptualization: Sriram Vijendran, Bailey Arruda, Oliver Eulenstein.

Data curation: Tavis K. Anderson.

Formal analysis: Sriram Vijendran, Bailey Arruda, Tavis K. Anderson, Oliver Eulenstein.

Funding acquisition: Tavis K. Anderson, Oliver Eulenstein.

Investigation: Bailey Arruda.

Methodology: Sriram Vijendran, Tavis K. Anderson.

Project administration: Sriram Vijendran, Tavis K. Anderson, Oliver Eulenstein.

Resources: Bailey Arruda, Tavis K. Anderson, Oliver Eulenstein.

Software: Sriram Vijendran, Tavis K. Anderson.

Supervision: Tavis K. Anderson, Oliver Eulenstein.

Validation: Oliver Eulenstein.

Visualization: Sriram Vijendran.

Writing – original draft: Sriram Vijendran, Bailey Arruda, Tavis K. Anderson, Oliver Eulenstein.

Writing – review & editing: Sriram Vijendran, Bailey Arruda, Tavis K. Anderson, Oliver Eulenstein.

References

1. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans Pattern Anal Mach Intell.* 2022;44(7):3523–42. <https://doi.org/10.1109/TPAMI.2021.3059968> PMID: [33596172](#)
2. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. p. 779–88.
3. Chen X, Ma H, Wan J, Li B, Xia T. Multi-view 3D Object Detection Network for Autonomous Driving. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. p. 6526–34.
4. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng P-A, Guo YB, et al. Gland segmentation in colon histology images: The glas challenge contest. *Med Image Anal.* 2017;35:489–502. <https://doi.org/10.1016/j.media.2016.08.008> PMID: [27614792](#)
5. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Modern Pathol.* 2022;35(1):23–32.
6. Yuan Y, Chao M, Lo Y-C. Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance. *IEEE Trans Med Imaging.* 2017;36(9):1876–86. <https://doi.org/10.1109/TMI.2017.2695227> PMID: [28436853](#)
7. Gevers T, Smeulders AWM. Image search engines: An overview. *Emerging Topics in Computer Vision.* 2004. p. 1–54.
8. Ranganathan H, Venkateswara H, Chakraborty S, Panchanathan S. Deep active learning for image classification. In: 2017 IEEE International Conference on Image Processing (ICIP). 2017. p. 3934–8.

9. Zhang Y, Lv C, Wang D, Mao W, Li J. A novel image detection method for internal cracks in corn seeds in an industrial inspection line. *Comput Electr Agric.* 2022;197:106930. <https://doi.org/10.1016/j.compag.2022.106930>
10. Wang L, Shen Q. Visual Inspection of Welding Zone by Boundary-Aware Semantic Segmentation Algorithm. *IEEE Trans Instrum Meas.* 2020;70:1–9. <https://doi.org/10.1109/tim.2020.3010665>
11. Riaz F, Kamal K, Zafar T, Qayyum R. An inspection approach for casting defects detection using image segmentation. In: 2017 International Conference on Mechanical, System and Control Engineering (ICMSC). IEEE; 2017. p. 101–5.
12. Wilkinson MD, Dumontier M, Aalbersberg IJ sbrand J, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data.* 2016;3(1):1–9.
13. Scheffler M, Aeschlimann M, Albrecht M, Bereau T, Bungartz H-J, Felser C, et al. FAIR data enabling new horizons for materials research. *Nature.* 2022;604(7907):635–42.
14. Ren P, Xiao Y, Chang X, Huang P-Y, Li Z, Gupta BB, et al. A Survey of Deep Active Learning. *ACM Comput Surv.* 2021;54(9). <https://doi.org/10.1145/3472291>
15. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal.* 2021;71:102062. <https://doi.org/10.1016/j.media.2021.102062> PMID: 33901992
16. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. 2015.
17. Kwon Y, Won J-H, Kim BJ, Paik MC. Uncertainty quantification using Bayesian neural networks in classification: application to biomedical image segmentation. *Computat Stat Data Anal.* 2020;142:106816. <https://doi.org/10.1016/j.csda.2019.106816>
18. Ebrahimi S, Elhoseiny M, Darrell T, Rohrbach M. Uncertainty-guided Continual Learning with Bayesian Neural Networks. 2019. Available from: <https://arxiv.org/abs/1906.02425>
19. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. ICML'16. 2016. p. 1050–9.
20. Li Y, Gal Y. Dropout Inference in Bayesian Neural Networks with Alpha-divergences. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. vol. 70 of Proceedings of Machine Learning Research. PMLR; 2017. p. 2052–61.
21. Bao Y, Velni JM, Shahbakhti M. Epistemic uncertainty quantification in state-space LPV model identification using Bayesian neural networks. *IEEE Control Syst Lett.* 2021;5(2):719–24. <https://doi.org/10.1109/LCSYS.2020.3005429>
22. Foong AYK, Li Y, Hernández-Lobato JM, Turner RE. 'In-Between' Uncertainty in Bayesian Neural Networks. 2019. Available from: <https://arxiv.org/abs/1906.11537>
23. An B, Wu W, Han H. Deep Active Learning for Text Classification. In: Proceedings of the 2nd International Conference on Vision, Image and Signal Processing. ICVIPSP 2018. New York (NY): Association for Computing Machinery; 2018. p. 1–6.
24. Beck N, Sivasubramanian D, Dani A, Ramakrishnan G, Iyer R. Effective evaluation of deep active learning on image classification tasks. 2021.
25. Yang L, Zhang Y, Chen J, Zhang S, Chen DZ. A Deep Active Learning Framework for Biomedical Image Segmentation. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins L, Duchesne L, editors. Medical Image Computing and Computer Assisted Intervention - MICCAI 2017. Cham: Springer International Publishing; 2017. p. 399–407.
26. Casanova A, O Pinheiro P, Rostamzadeh N, Pal CJ. Reinforced active learning for image segmentation. 2020.
27. Zhao Z, Zeng Z, Xu K, Chen C, Guan C. DSAL: Deeply Supervised Active Learning From Strong and Weak Labelers for Biomedical Image Segmentation. *IEEE J Biomed Health Inform.* 2021;25(10):3744–51. <https://doi.org/10.1109/JBHI.2021.3052320> PMID: 33460386
28. Siddiqui Y, Valentin J, Niessner M. ViewAL: Active Learning With Viewpoint Entropy for Semantic Segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. p. 9430–40.
29. Jospin LV, Laga H, Boussaid F, Buntine W, Bennamoun M. Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users. *IEEE Computat Intell Magazine.* 2022;17(2):29–48. <https://doi.org/10.1109/MCI.2022.3155327>
30. Mullachery V, Khera A, Husain A. Bayesian Neural Networks. 2018. Available from: <https://arxiv.org/abs/1801.07710>
31. Roy S, Unmesh A, Namboodiri VP. Deep active learning for object detection. In: BMVC. 2018. 91 p.
32. Sun Q, Laddha A n k i t, Batra D. Active learning for structured probabilistic models with histogram approximation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. 3612–21.
33. Cai L, Xu X, Liew JH, Sheng Foo C. Revisiting Superpixels for Active Learning in Semantic Segmentation with Realistic Annotation Costs. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. p. 10983–92.
34. Siméoni O, Budnik M, Avrithis Y, Gravier G. Rethinking deep active learning: Using unlabeled data at model training. In: 2020 25th International Conference on Pattern Recognition (ICPR). 2021. p. 1220–7.
35. Shi Z, He L, Suzuki K, Nakamura T, Itoh H. Survey on neural networks used for medical image processing. *Int J Computat Sci.* 2009;3(1):86.
36. Achanta R, Shaji A p p u, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans Pattern Analys Mach Intell.* 2012;34(11):2274–82. <https://doi.org/10.1109/TPAMI.2012.120>
37. Yu S, Luo X, He Z, Yan J, Lv K, Shi D. An Improved Sampling Strategy for QBC Algorithm and its Application on Gas Sensor Array Signal Processing. In: 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP). 2018. p. 224–8.

38. Xie S, Feng Z, Chen Y, Sun S, Ma C, Song M. DEAL: Difficulty-aware Active Learning for Semantic Segmentation. 2020.
39. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. 2015. p. 234–41.
40. Graves A I e x. Practical Variational Inference for Neural Networks. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, editors. *Advances in Neural Information Processing Systems*. vol. 24. Curran Associates, Inc.; 2011.
41. Kumar S, Felix L, Marcus L. A comprehensive guide to Bayesian convolutional neural network with variational inference. 2019.
42. Berman M, Triki AR, Blaschko MB. The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018. p. 4413–21.
43. Rezatofghi H, Tsoi N, Gwak JYO, Sadeghian A, Reid I, Savarese S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. p. 658–66.
44. Zhao R, Qian B, Zhang X, Li Y, Wei R, Liu Y, et al. Rethinking Dice Loss for Medical Image Segmentation. In: 2020 IEEE International Conference on Data Mining (ICDM). 2020. p. 851–60.
45. Bertels J, Eelbode T, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham: Springer International Publishing; 2019. p. 92–100.
46. Mohseni Salehi SS, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. 2017.
47. Abraham N, Khan NM. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. 2018.
48. Izmailov P, Vikram S, Hoffman MD, Wilson AGG. What Are Bayesian Neural Network Posteriors Really Like? In: Meila M, Zhang T, editors. *Proceedings of the 38th International Conference on Machine Learning*. vol. 139 of *Proceedings of Machine Learning Research*. PMLR; 2021. p. 4629–40.
49. Campbell T r e v o r, Broderick T a m a r a. Automated scalable Bayesian inference via Hilbert coresets. *J Mach Learn Res*. 2019;20(1):551–88.
50. Cobb AD, Roberts SJ, Gal Y. Loss-Calibrated Approximate Inference in Bayesian Neural Networks. 2018. Available from: <https://arxiv.org/abs/1805.03901>
51. Sun S, Zhang G, Shi J, Grosse R. Functional Variational Bayesian Neural Networks. 2019. Available from: <https://arxiv.org/abs/1903.05779>
52. Hinton GE, van Camp D. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In: *Proceedings of the Sixth Annual Conference on Computational Learning Theory*. COLT '93. New York (NY): Association for Computing Machinery; 1993. p. 5–13.
53. Wang K, Zhang D, Li Y, Zhang R, Lin L. Cost-Effective Active Learning for Deep Image Classification. *IEEE Trans Circuits Syst Video Technol*. 2017;27(12):2591–600. <https://doi.org/10.1109/tcsvt.2016.2589879>
54. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 2014.
55. Torralba A, Russell BC, Yuen J. LabelMe: Online Image Annotation and Applications. *Proc IEEE*. 2010;98(8):1467–84. <https://doi.org/10.1109/JPROC.2010.2050290>
56. Shephard A, Jahanifar M, Wang R, Dawood M, Graham S, Sidlauskas K, et al. Tiager: Tumor-infiltrating lymphocyte scoring in breast cancer for the tiger challenge. *arXiv preprint arXiv:2206.11943*. 2022.
57. Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *Proceedings of the IEEE international conference on computer vision*. 2015. p. 2650–8.
58. Huisman A, Looijen A, van den Brink SM, van Diest PJ. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. *Hum Pathol*. 2010;41(5):751–7. <https://doi.org/10.1016/j.humpath.2009.08.026> PMID: 20129646
59. Marcolini A, Bussola N, Arbitrio E, Amgad M, Jurman G, Furlanello C. histolab: A Python library for reproducible Digital Pathology preprocessing with automated testing. *SoftwareX*. 2022;20:101237.
60. Beluch WH, Genewein T, Nurnberger A, Kohler JM. The Power of Ensembles for Active Learning in Image Classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 9368–77.
61. Sinha S, Ebrahimi S, Darrell T. Variational adversarial active learning. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019. p. 5972–81.