UCLA
UCLA Previously Published Works

Title
PathAL: An Active Learning Framework for Histopathology Image Analysis

Permalink
https://escholarship.org/uc/item/5439t7wx

Journal
IEEE Transactions on Medical Imaging, 41(5)

ISSN
0278-0062

Authors
Li, Wenyuan
Li, Jiayun
Wang, Zichen
et al.

Publication Date
2022-05-01

DOI
10.1109/tmi.2021.3135002

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE Trans Med Imaging. Author manuscript; available in PMC 2023 May 02.

Published in final edited form as:

IEEE Trans Med Imaging. 2022 May ; 41(5): 1176–1187. doi:10.1109/TMI.2021.3135002.

PathAL: An Active Learning Framework for Histopathology Image Analysis

Wenyuan Li,
Computational Diagnostics Lab, Departments of Radiological Sciences and Pathology and Laboratory Medicine, UCLA, 924 Westwood Blvd, Los Angeles, CA 90024 USA

Jiayun Li,
Computational Diagnostics Lab and Medical Imaging Informatics, Departments of Radiological Sciences and Pathology and Laboratory Medicine, UCLA, 924 Westwood Blvd., Los Angeles, CA 90024 USA.

Zichen Wang,
Computational Diagnostics Lab and Medical Imaging Informatics, Departments of Radiological Sciences and Pathology and Laboratory Medicine, UCLA, 924 Westwood Blvd., Los Angeles, CA 90024 USA.

Jennifer Polson,
Computational Diagnostics Lab and Medical Imaging Informatics, Departments of Radiological Sciences and Pathology and Laboratory Medicine, UCLA, 924 Westwood Blvd., Los Angeles, CA 90024 USA.

Anthony E. Sisk,
Department of Pathology and Laboratory Medicine, UCLA, Los Angeles, CA 90024 USA.

Dipti P. Sajed,
Department of Pathology and Laboratory Medicine, UCLA, Los Angeles, CA 90024 USA.

William Speier,
Computational Diagnostics Lab, the Department of Radiological Sciences and Bioinformatics, UCLA, CA 90024 USA.

Corey W. Arnold
Computational Diagnostics Lab and Medical Imaging Informatics, Departments of Radiological Sciences and Pathology and Laboratory Medicine, UCLA, CA 90024 USA

Abstract

Deep neural networks, in particular convolutional networks, have rapidly become a popular choice for analyzing histopathology images. However, training these models relies heavily on a large number of samples manually annotated by experts, which is cumbersome and expensive. In

Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

liwenyuan.zju@gmail.com .

Colour figures are available online. Supplementary materials are available in the supplementary files/multimedia tab on IEEEXplore.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

addition, it is difficult to obtain a perfect set of labels due to the variability between expert annotations. This paper presents a novel active learning (AL) framework for histopathology image analysis, named PathAL. To reduce the required number of expert annotations, PathAL selects two groups of unlabeled data in each training iteration: one “informative” sample that requires additional expert annotation, and one “confident predictive” sample that is automatically added to the training set using the model’s pseudo-labels. To reduce the impact of the noisy-labeled samples in the training set, PathAL systematically identifies noisy samples and excludes them to improve the generalization of the model. Our model advances the existing AL method for medical image analysis in two ways. First, we present a selection strategy to improve classification performance with fewer manual annotations. Unlike traditional methods focusing only on finding the most uncertain samples with low prediction confidence, we discover a large number of high confidence samples from the unlabeled set and automatically add them for training with assigned pseudo-labels. Second, we design a method to distinguish between noisy samples and hard samples using a heuristic approach. We exclude the noisy samples while preserving the hard samples to improve model performance. Extensive experiments demonstrate that our proposed PathAL framework achieves promising results on a prostate cancer Gleason grading task, obtaining similar performance with 40% fewer annotations compared to the fully supervised learning scenario. An ablation study is provided to analyze the effectiveness of each component in PathAL, and a pathologist reader study is conducted to validate our proposed algorithm.

Keywords

Histopathology Image Analysis; Active Learning; Noisy Label Detection; Curriculum Learning

I. INTRODUCTION

DEEP neural networks (DNNs) have achieved great success in a wide variety of medical image analysis tasks [1]. However, noise-free expert annotations are crucial to achieve high performance. Unfortunately, in medical image analysis, obtaining enough annotations can be expensive and time-consuming for many tasks. In histopathology images analysis, the size of the collected dataset can be large, and performing annotations requires years of professional training and domain knowledge. In addition, the labels provided by different pathologists can demonstrate high inter-reader variability. For example, in prostate cancer grading using Gleason scoring, the concordance rate of multiple pathologists can be as low as 57.9% [2], which results in noisy annotations. DNNs are capable of fitting to noisy annotations, but they may not generalize to unseen data, which is an important component of clinical applications. Furthermore, it is challenging to distinguish mislabeled samples from hard samples. Mislabeled samples are samples with incorrect annotations, while hard samples have the correct label, but the samples themselves are not “typical.” The lack of large and noise-free annotation sets is a significant challenge in histopathology image analysis, preventing DNNs to scale to the size of collected data.

Recent studies have investigated methods for dealing with annotation challenges in medical imaging, such as semi-supervised learning, multi-instance learning, and transfer learning [3]. One of these solutions that is of particular interests is to use active learning (AL) [4]. AL

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

aims to reduce the amount of labeled data necessary for the learning task. It employs various methods to select samples from an unlabeled set. The selected samples are then annotated by experts and used to train the model. A carefully designed sampling method can reduce the overall number of labeled data points required to train the model and make the model robust to class imbalances. However, traditional AL methods do not address the noisy label issue.

A few studies have also sought to detect noisy labels in training data and enhance the performance of DNNs in medical image analysis. Specifically, addressing the issue of noisy labels remains an ongoing challenge for the medical imaging analysis community. Dgani *et al.* [5] adopted a noisy channel in neural networks, which models the stochastic relation between the correct label and the observed noisy label. Xue *et al.* [6] proposed an online uncertainty sample mining strategy to suppress the noisy samples. However, these methods do not distinguish mislabeled samples from hard samples. Making this distinction could greatly improve histopathology images analysis tasks with noisy labels.

In this work, we present a histopathology AL framework (PathAL) that is able to dynamically identify noisy labels and sample images that need to be annotated. Our goal is to provide a solution that is able to reduce annotations required from experts and to simultaneously handle noisy labels. For each iteration of PathAL, we first train the network using annotated images. We then force the network to modulate between overfitting and underfitting by adjusting the hyper-parameters. In this process, we monitor and rank the normalized average loss of every labeled sample and the normalized average predictive entropy of every unlabeled sample. We also measure the complexity of data points using their distribution density in the feature space and rank their complexity in an unsupervised manner. By doing so, the noisy labeled samples can be identified and discarded, while the hard and minority samples can be preserved. The unlabeled images that are most informative to the model are selected for annotations and added for training for next iteration. In addition, the typical unlabeled samples with the highest predictive confidence are added to the training pool with pseudo annotations generated by the model itself. This cost-effective sample selection strategy is able to improve the classification performance with far fewer manual annotations. Our proposed method is a tailor-made strategy for histopathology image analysis. The main contributions of this paper include: 1) an AL framework (PathAL) that is able to dynamically identify important samples to annotate and to distinguish noisy from hard samples in the training set, 2) extensive experiments that demonstrate model improvement with less annotation effort and noisy samples; and 3) a reader study performed by a domain expert to validate our algorithm.[1]

II. RELATED WORK

A. Active Learning

A typical AL framework consists of a method to evaluate the *informativeness* of each unannotated data point x_u given $f'(x|L')$, where f' is a model trained on a labeled dataset L'. In literature, methods to evaluate *informativeness* can be generally classified into two types: 1) calculate the uncertainty, and 2) calculate the representativeness. In

[1]The related code is available at https://github.com/Wenyuan-Vincent-Li/PathAL

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

uncertainty-related methods, it is assumed that the more uncertain a prediction, the more information we can gain by including the ground truth for that sample in the training set. Wen *et al.* [7] proposed an AL method that uses uncertainty sampling to support quality control of nucleus segmentation in pathology images. Gal *et al.* [8] introduced Bayesian CNNs to measure the uncertainty of predictions (mc-dropout). They demonstrated their approach for skin cancer diagnosis to show significant performance improvements over uniform sampling using the method for sample selection [9], which sought to maximize the mutual information between predictions and model posterior. Kuo *et al.* [10] measured the uncertainty through the lens of query-by-committee (QBC). They used Jensen-Shannon(JS) divergence between multiple models and demonstrated that it works best for intracranial hemorrhage detection. Konyushkova *et al.* [11] proposed to exploit geometric smoothness priors in the image space to aid the segmentation process in AL. They demonstrated state-of-the-art performance on mitochondria segmentation from electron microscopy (EM) images and on an magnetic resonance imaging (MRI) tumor segmentation task for both binary and multi-class segmentation. Another area of work focuses on the measure of representativeness in addition to uncertainty measures. This research uses the idea that methods only concerned with uncertainty have the potential to focus only on small regions of the distribution, and that training on samples from the same area of the distribution will introduce redundancy to the selection strategy or may skew the model towards a particular area of the distribution. Therefore, the selection method should also cover a large range of the data distribution in order to increase sample representativeness. Yang *et al.* [12] presented Suggestive Annotation, a deep AL framework for medical image segmentation, which uses an alternative formulation of uncertainty sampling combined with a form of representativeness density weighting. They demonstrated state-of-the-art performance using 50% of the available data on the MICCAI gland segmentation challenge and a lymph node segmentation task. Smailagic *et al.* [13] proposed MedAL, an AL framework for medical image segmentation. They proposed a sampling method that combines uncertainty and distance between feature descriptors to extract the most informative samples from an unlabeled dataset. Ozdemir *et al.* [14] proposed a Borda-count based combination of an uncertainty and representativeness measure to select the next batch of samples. They introduced new representativeness measures such as “Content Distance,” defined as the mean squared error between layer activation responses of a pre-trained classification network. Sourati *et al.* [15] proposed a method for ensuring diversity among queried samples by calculating the Fisher Information. They demonstrated the performance of their approach improved after labelling a small percentage of voxels, outperformed random sampling, and achieved higher accuracy than entropy based querying.

Our proposed PathAL model combines both uncertainty and representativeness measures in the data selection algorithm. Unlike the methods discussed above, our AL framework also involves a *complementary sampling* strategy, in which the framework selects from an unlabeled dataset with: 1) a set of most uncertain samples to be annotated by an oracle, and 2) a set of highly certain samples that are “pseudo-labeled” by the framework. A similar idea has been proposed by [16] in natural images, but it has never been used in histopathology images. Furthermore, PathAL also considers the noisy label issue, which can deteriorate the performance of the AL framework in histopathology analysis. To the best of our knowledge,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

joint modeling of uncertainty and representation has not been explored in the previous literature in histopathology image analysis.

B. Noisy Label Detection

Addressing noisy labels in machine learning is an ongoing challenge. Several attempts have been made in natural image and medical image tasks [17]. In general, there are two types of solutions to deal with noisy labels in a training dataset: 1) train models to detect the noisy labels and then clean or remove them to reduce their impact in the model training; and 2) directly train a noise-robust model with noisy labels. In line with the first approach, Koh and Liang [18] proposed an influence function to measure samples that were "harmful" to model training. Lee *et al.* [19] proposed CleanNet, which was a joint neural embedding network. This approach summarized the knowledge of label noise from a fraction of manually verified classes. Transfer learning was then conducted to transfer the knowledge to other classes to handle label noise. Han *et al.* [20] proposed co-teaching, in which two deep networks were trained simultaneously. Each network selected which samples the other network used for training. Each of the networks taught the other to identify noisy labels. In [21], Guo *et al.* proposed CurriculumNet, in which training data were divided into several subsets by ranking their distribution density as a measure of complexity. The subsets were formed as a curriculum to teach the model to understand label noise gradually. A similar idea was proposed in [22]. In this work, a MentorNet was trained to identify potential noisy labels. The network then provided a data-driven curriculum for StudentNet, which was trained on the less noisy data samples. Huang *et al.* [23] proposed O2U-Net to make the network transition from overfitting to underfitting (O2U) automatically. By monitoring the training loss variation, they could detect and remove noisy labels from the original dataset.

On the other hand, several other approaches that directly train a noise-robust model with noisy labels have been proposed. Goldberger and Ben-Reuven [24] proposed to model label noise by adding softmax layers to estimate the transition between correct labels and noisy labels. Xiao *et al.* [25] proposed a probabilistic model to describe the relations among images, true labels, noisy labels, and noise types. This probabilistic model required a small set of verified labels without noise. Reed and Lee [26] proposed the idea of *consistency* to model noisy labels. Sample reconstruction errors were applied as the consistency objective to estimate the noise distribution. There are a few studies that have addressed issues of noisy labels in medical imaging. Dgani *et al.* [5] used a noise adaptation layer similar to [24] on a mammography classification task and outperformed standard training methods. Xue *et al.* [6] proposed an online uncertainty sample mining method (OUSM) to detect the noisy labels and iteratively re-weight sample losses. We refer interested readers to Karimi *et al.* [17] for a detailed and comprehensive review on recent research works on noisy labels.

To the best of our knowledge, we are the first to incorporate a noisy sample detector in an AL framework. In our proposed PathAL, we adopt O2U-Net as a noisy label detector. It enhances our AL framework for the following reasons: 1) O2U-Net is a noise-cleansing method, so AL can be conducted after noisy label detection and removal to reduce the need for human annotations and improve the generalization capacity of the model; 2) other noise-cleansing methods require either particular assumptions on noise distribution estimation or

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

extra specifically designed loss functions or networks (e.g. Co-teaching and MentorNet), while O2U-Net only requires adjusting the hyper-parameters of deep networks; and 3) by leveraging curriculum learning, in which images are divided into several subsets by ranking their distribution density in deep feature space, we can distinguish between the noisy labeled samples and the hard samples, which is a challenging task in histopathology image analysis.

III. METHODS

In this section, we first formally define our problem and the notations we use in this study. We then introduce curriculum sample classification and noisy sample detection methods, two key components of our proposed PathAL model. Finally we describe our proposed PathAL method in detail.

A. Problem Definition

In traditional AL, we assume there is a large pool of unlabeled data U available and an oracle to help with labeling for every unlabeled data point x_u to add to labeled set L. We consider the whole training set to be $T = L \cup U = L_1 \cup U_1 = \cdots = L_k \cup U_k$, where L_i, U_i represents the labeled and unlabeled sets in ith iteration. AL starts from a small labeled set L_1 and tries to find the most informative samples $x^*_{1,j} \in U_1$. All the informative samples selected by the algorithm form a set I_1 and will be annotated by domain experts and added to the labeled set for model training in the next iteration. Thus, we have $L_2 = L_1 + I_1$ and in general $L_{i+1} = L_i + I_i$.

In contrast to the traditional AL model, our proposed PathAL considers three groups of samples in the data pool: 1) annotated samples that are in the training dataset that have a high probability of incorrect label assignment (noisy samples), denoted as the noisy set N_i; 2) unlabeled samples that are most informative to the current model (informative samples), denoted as the informative set I_i; and 3) unlabeled samples for which the current model is confident in its predictions (confident samples), denoted as the confident set C_i. PathAL will discard the noisy samples, require experts to annotate the informative samples and add them to the training pool, and add confident samples to the data pool with their own, model-assigned annotations, simultaneously. Thus we have $L_{i+1} = L_i - N_i + I_i + C_i$, where $N_i \subseteq L_i$, $I_i \subseteq U_i$, and $C_i \subseteq U_i$, and the training set $T = L_i + U_i$. The general process of PathAL is illustrated in Figure 1(a). The core goals for PathAL are: 1) detect the noisy samples and distinguish them from hard samples, and 2) detect the informative samples to be annotated and add confident samples automatically. To meet these goals, we first briefly discuss our curriculum sample classification method, inspired by CurriculumNet [21] and O2U-Net [23] noisy sample detection, upon which these two questions are answered in PathAL.

B. Curriculum Sample Classification

A key component of our PathAL framework is to leverage curriculum learning to classify each example in a training set to be easy, medium, or hard based on its complexity. We extend CurriculumNet [21] for AL scenarios and use it in a fully unsupervised fashion. In each iteration, we use a trained model to compute a deep representation for each image in the training set T. This step aims to roughly map all training images into a feature

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

space where the underlying structure and the complexity of the images can be discovered. We then classify each sample into different complexity levels, ranging from easy samples with high confidence labels to difficult samples whose labels may contain noise. To do so, we first reduce the dimension of the deep features using t-distributed Stochastic Neighbor Embedding (t-SNE) [27]. With this set of reduced features, we use the K-means algorithm to cluster the images into different groups. Each group will ideally contain images with similar diagnoses. This step aims to help the following process select representative samples covering the whole training sample space. Next, we calculate a Euclidean distance matrix $D \subseteq \mathbb{R}^{n \times n}$ as,

$$D_{i,j} = \left\| f(I_i) - f(I_j) \right\|^2 \tag{1}$$

where n is the number of images in the same group, I_i, I_j are two images in this group, $f(I_i)$, $f(I_j)$ are the feature vectors of the two images in deep feature space. $D_{i,j}$ indicates a similarity value between I_i and I_j. Then we calculate a local density (ρ_i) for each image,

$$\rho_i = \sum_j X\left(D_{i,j} - d_c\right) \tag{2}$$

where

$$X(d) = \begin{cases} 1 & d < 0 \\ 0 & \text{other} \end{cases} \tag{3}$$

d_c in the above equation is a distance threshold we select for determining the local density. It is selected by first sorting n^2 distances from small to large values, and choosing the top k%. Following the practice in [21], we set $k = 60$ in all our experiments. The local density ρ_i counts how many samples are closer to image I_i than the threshold d_c in the deep feature space. Finally, we use a K-means clustering method to classify each sample as easy, medium, or hard based on the local density for each group. To this end, we assume that a group of easy images with correct labels will often have similar visual characteristics, project closely to each other in the feature space, and therefore have a high ρ_i. By contrast, hard images often have more visual diversity, resulting in a sparse distribution with a smaller ρ_i. Figure 1(b) illustrates the workflow of our curriculum classification (CC) algorithm. We also summarize the CC in Algorithm 1.

Algorithm 1

Curriculum Classification

Require: trained DNN (f), images (x_i) in training set T;

(1) Generate deep image features $f(x_i)$ for each image x_i;

(2) Reduce dimensionality using t-SNE, then use K-means cluster algorithm to cluster these features into k different groups $g_1, g_2, \ldots, g_k$;

(3) Calculate a Euclidean distance matrix $D \subseteq R^{n \times n}$ as $D_{i,j} = \|f(x_i) - f(x_j)\|^2$;

(4) Calculate a local density function for each image $\rho_i = \sum_j X\left(D_{i,j} - d_c\right)$;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(5) Use K-means cluster algorithm to classify each image x_i to easy, medium, or hard based on their local density ρ_i in each group g_r.

Note that although our curriculum classification algorithm is inspired by CurriculumNet [21], it is substantially different from it in the following aspects. CurriculumNet is performed in weakly supervised learning settings, where the authors have access to all the labels of the samples and are able to use the subgroup with the same label for curriculum classification. In this study, our curriculum classification is performed in each iteration of AL, where we do not have full access to the annotations of training samples. Therefore, we have to use unsupervised K-means clustering to first group the training images, and then classify the samples in each group based on the local density. In this way, we are able to sample the unlabeled images evenly in the deep feature space.

As our model evolves during each AL iteration, it is hard to distinguish whether a sample is a noisy sample that has a wrong label or is a complex sample that the model has not learned yet. Accordingly, we introduce another key component of PathAL, with which we are able to distinguish noisy samples from hard ones, and discover the most informative samples to be annotated.

C. Noisy Sample Detection

It is challenging to determine whether an incorrectly classified sample is a noisy one with the wrong label or a complex one that is inherently hard to learn for deep learning models. The CC algorithm in Section III-B only considers the visual complexity of the training samples, but does not provide much information about how well the current AL model learns these samples. To help with this, we introduce a noisy sample detector by using O2U-Net [23].

The key observation from O2U-Net is that noisy-labeled samples are usually memorized at the late stages of training, as is the case with hard samples. At the beginning of training, when the network is still underfitting, the losses of noisy and hard samples are larger than those of easy samples because the model quickly fits to easy samples. Conversely, during the late stages of training, the network usually overfits to the training set. It memorizes both the noisy/hard samples and easy samples, so that the losses generated from them are indistinguishable. Therefore, by tracking the variation of loss for every sample at different stages of training, it is possible to detect noisy and hard samples. Based on this idea, the O2U-Net attempts to cycle training between underfitting and overfitting by tuning the learning rate, while observing the variation of loss for every sample in L_i. Specifically, at the beginning of training, a large learning rate is set. The learning rate gradually decreases to some extent during training and is then reset to the original learning rate. This process repeats for multiple rounds until enough loss statistics are gathered. When the network almost converges to some minimum (nearly overfitting), a large learning rate can make the network jump out of the minimum. As a result, the network will quickly start underfitting the data. By monitoring the training loss for each sample, we can expect samples with larger average loss after cyclical training to have higher probability of being a mislabeled or hard sample. We apply the same network to detect noisy labels and to train the final classifier

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

using EfficentNet-B0 [28] (see Section IV-C for more training details). For a more detailed description of O2U-Net, please refer to [23].

The original O2U-Net only monitors the training loss for each sample in L_i. We extend it to monitor the predictive entropy for every sample in the unlabeled dataset U_i. Specifically, we do inference after each epoch in the O2U training cycles. We record the predictive entropy for each sample in U_i and find the samples with highest average predictive entropy. These samples are the most "informative" samples to the current model because they cannot be predicted confidently and may not be represented in the feature space of the current labeled set L_i. We summarize the O2U training workflow in Algorithm 2. We point out that the O2U-Net alone cannot distinguish between noisy samples and hard samples. With the help of curriculum classification, however, we are able to heuristically separate these two types of samples, which we will discuss in the next section.

Algorithm 2

Training O2U

Require: trained DNN (f), labeled image x_{l_i}, unlabeled image x_{u_i};

for Each epoch **do**

Adjust learning rate via Equation (7).

for Each labeled image x_{l_i} **do**

(1) Compute and record training loss lss_i;

(2) Update the network f;

end for

for Each unlabeled image x_{u_i} **do**

(1) Compute and record predictive entropy ent_i;

end for

end for

(1) Compute the normalized average loss $\overline{lss_i}$ of every labeled sample among all the epochs;

(2) Compute the normalized average predictive entropy $\overline{ent_i}$ of every unlabeled sample among all the epochs;

(3) Obtain the order by ranking all the labeled samples by $\overline{lss_i}$ and all the unlabeled samples by $\overline{ent_i}$.

D. PathAL

After introducing the CC algorithm and the O2U process, we now specify the core goals of PathAL: 1) detect noisy samples and distinguish them from hard samples, and 2) detect informative samples to be annotated and add confident samples using "pseudo-labeles" assigned by the model itself.

At the ith iteration of PathAL, we first train a network using the current labeled dataset L_i until it converges. Then we apply the O2U process to continue the training. We monitor the loss variation of each labeled sample and predictive entropy of each unlabeled sample. Simultaneously, we apply the CC algorithm on the samples in training set T and classify them as easy, medium, or hard samples based on their local density in the feature space.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To detect noisy samples, we find those that have large loss variations in L_i and are also classified as easy by the the CC algorithm. On one hand, these samples have large loss variations, which means they are hard to learn by the current network. On the other hand, the samples must have a high local density in the deep feature space in order to be labeled as "easy", *i.e.* they are typical samples that are visually similar to other samples in T. Thus, there is a higher probability that the pathologist annotations for these samples contain noise. To prevent them from impacting the model's training and performance, we discard these samples in the next training iteration. To detect the informative samples that require additional expert annotations, we select the samples with the highest average predictive entropy during O2U training. As discussed in Section III-C, these samples are most informative because they cannot be predicted confidently by the current model. In addition, we add unlabeled samples that have the lowest predictive entropy and are classified as "easy" or "medium" by the CC algorithm. Our model is confident in these predictions, and they are "typical" samples in the deep feature space, so there is a high probability that the model predictions are correct. Therefore, it is cost-effective to add them automatically into L_{i+1} with self-assigned "pseudo-labels." Algorithm 3 illustrates the workflow of our PathAL algorithm.

Algorithm 3

PathAL

Require: a DNN (f), training set T, initial selected labeled image set L_1 and the rest unlabeled image set U_1;

for Each iteration of PathAL **do**

(1) Train DNN f based on the current labeled image set L_i until it is converged;

(2) Perform CC and O2U training in training set T;

(3) Select noisy samples N_i, most informative samples I_i, and confident predicted samples C_i based on CC and O2U results;

(4) Update the training set for next iteration as $L_{i+1} = L_i - N_i + I_i + C_i$;

end for

As the model evolves during the training process, both the CC and O2U results change. Therefore, we do not discard the noisy samples completely. Instead, we keep them in a pool and examine if they need to be added back throughout the training process. In the later stage, if noisy samples have been identified as informative, we add them back with their original annotated labels. If noisy samples are identified as easy, we add them back with model-assigned "pseudo-labels". In doing so, we build a mechanism for the model to correct errors made at the beginning of the AL process. We performed a detailed analysis on this mechanism in Section IV-E5. We summarize how to combine the CC and O2U results based on their relationship in Figure 1(c).

IV. DATASETS, EXPERIMENTS AND RESULTS

In this section, we first introduce the dataset and evaluation metrics we used in our experiments. We then discuss the implementation details of our model, followed by the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

description of several baseline models. Finally, we demonstrate and discuss the experimental results.

A. Dataset and Pre-processing

To demonstrate the effectiveness of our PathAL technique, we use the Kaggle dataset from the "Prostate cANcer graDe Assessment using the Gleason grading system" (PANDA) challenge to simulate the AL scenario. The dataset consists of over 11,000 whole-slide images of digitized H&E-stained biopsies originating from two centers (Karolinska Institute and Radboud University Medical Center). Different slide scanners with slightly different maximum microscope resolutions were used for digitization and labels were generated from different pathologists. The Karolinska dataset was labeled by a single experienced pathologist. Label noise may exist in this dataset due to the lack of label validation by another pathologist. The Radboud dataset was read by trained students. For this dataset, some minor label noise may also exist in the training set due to mistakes in the annotation process or inconclusive results. Though the label noise presents a modeling challenge, it resembles many real-world scenarios. As mentioned above, even experts in the field with years of experience do not always agree on how to interpret prostate histology.

Each sample image in the PANDA dataset is a large, which requires an efficient algorithm to locate areas of concern on which to focus. We used our previously developed tiling algorithm with a blue-ratio selection criteria to identify the most informative tissue areas [29]. Specifically, the algorithm consists of four steps. First, a binary mask of the tissue on the slide is created by setting a threshold for the average intensity. This threshold is set empirically to 90% of the maximum image intensity value. Second, the mask is smoothed using morphological closing and the skeleton of the smoothed mask is then found and branches are removed by finding the endpoints with the maximum geodesic distance. Third, the mid-line is partitioned based on the patch size and overlap, tangent lines are found at each of these locations by looking at the neighborhood of nine pixels along the mid-line and the perpendicular line is drawn until intersection with the mask boundary. Finally, a set of patches that intersect with more than 60% with the mask are chosen to calculate their blue ratio, and the top k blue-ratio patches are selected. In this work, a patch size of 256 × 256 pixels was used, and 36 patches were selected for each slide. Figure 2 illustrates the pre-processing steps of the PANDA dataset.

B. Evaluation Metrics

The task of the PANDA challenge is to predict the ISUP grade on a 0–5 scale for each biopsy image based on Gleason grading system. Gleason grading is a subjective task, with high inter- and intra-observer variability. Agreement between pathologists is often measured using Cohen's kappa. Therefore we used the quadratic weighted kappa (QWK) to evaluate our model's performance. QWK measures the agreement between two outcomes. It typically varies from 0 (random agreement) to 1 (complete agreement), though it may be negative if there is less agreement than expected by chance.

The QWK is calculated as follows. First, an $N \times N$ histogram matrix O is constructed, such that $O_{i,j}$ corresponds to the number of ISUP grade i (actual) that received a predicted value

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

j. An $N \times N$ matrix of weights, w, is calculated based on the difference between actual and predicted values as,

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{4}$$

After that, an $N \times N$ histogram matrix of expected outcomes, E, is calculated assuming that there is no correlation between values. This is calculated as the outer product between the actual histogram vector of outcomes and the predicted histogram vector, normalized such that E and O have the same sum. From these three matrices, the QWK is calculated as,

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{5}$$

C. Network Backbone, Loss Function, and Other Training Details

In this study, we used EfficientNet-B0 [28] for all of our experiments as it achieved competitive results in the PANDA challenge without high computational cost compared to other network architectures, such as ResNeXt34 and ResNeXt50. However, we note that PathAL does not require a specific network backbone and can be easily adapted to use other networks to best fit various scenarios. We used normal Adam optimization in all the experiments. The model is trained on one single Tesla V100S GPU in PyTorch.

To predict an ISUP grade on a 0–5 scale, we used the ordinal regression loss function at the final layer of our network. This loss can better capture the ordinal relationship between grade and severity in the training set compared to multi-class classification or the mean square error loss function. Specifically, we used binary cross entropy loss with binning labels. For example, label Y = [r0, 0, 0, 0, 0] means ISUP grade 0, Y = [r1, 0, 0, 0, 0] means ISUP grade 1, and Y = [r1, 1, 1, 1, 1] means ISUP grade 5. The output of our model is a vector P with the same dimension of label Y. Assume p_i and y_i represent i th value of P and Y, the loss function will be,

$$\mathscr{L} = \sum_{i=1}^{N} -(y_i \log(p_i) + (1-y_i)\log(1-p_i)), \tag{6}$$

where N is our ISUP grade scale, *i.e.* 5 in our case.

We performed four-fold stratified cross-validation to demonstrate the effectiveness and robustness of PathAL. In each fold, we used a hold-out set as the testing set, and the rest as the training set. For comparison, we asked an expert pathologist to annotate 10% of the training samples each time. We compared PathAL performance with other baseline models alongside the pathologist annotations. In each iteration of PathAL, we excluded 1% of the whole training set $|T|$ as noisy samples, added the other 10% * $|T|$ annotated data and 5% * $|T|$ confidently predicted samples with their “pseudo-labels”.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

D. Baselines

We compared our model with state-of-the-art AL baselines including MC-Dropout [30], QBC [10], VAAL [31], and Core-Set [32]. The comparison is conducted using the identical four-fold cross-validation data split. In this section, we describe the acquisition methods used by these baseline models briefly.

- Monte Carlo Dropout to obtain model uncertainty for each sample in the unlabeled dataset and detect the top uncertain samples (***MC-Dropout***) [30]. Specifically, in MC-Dropout we simply collect the results of stochastic forward passes through drop out and measure the entropy of the predictions. Samples with the highest entropy are chosen for annotation.
- Query-by-committee to choose the data points for the next iteration (***QBC***). Similar to [10], we use Jensen-Shannon (JS) divergence to measure the uncertainty of unlabeled samples and select the top ones.
- Variational Adversarial Active Learning method to select unlabeled samples for annotation (***VAAL***) [31]. This method learns a low dimensional latent space from labeled and unlabeled data using Variational Autoencoder (VAE) and selects instances for labeling from the unlabeled pool that are sufficiently different in the latent space through a discriminator.
- Core-Set to select unlabeled samples to be annotated for the next round (***Core-Set***) [32]. Specifically, in the Core-Set method, we choose k data points to be labeled so that the largest distance between a data point and its nearest center is minimized.

E. Experimental Results

We first conducted a toy example run to show the effectiveness of our methods on identifying hard and noisy samples. Then we performed experiments in various settings and compared PathAL with other AL baselines. Note that the ISUP grade for samples in U_i is not available in the real AL scenario. However, we used the label in the dataset as ground truth to provide a quick sanity check and demonstrate that PathAL worked as expected.

1) Illustration of Toy Example: To illustrate the effectiveness of our method, we created a toy example for 2-class classification. Figure 4(a) demonstrates the sample-distributions of the classification problem. The dotted circle indicates the decision boundary while we have class 0 (C=0) inside the circle and class 1 (C=1) outside the circle. Samples that are far away from the decision boundary are considered as easy samples (blue dots for C=0 and orange dots for C=1), while samples that are close to the decision boundary are considered as hard samples. We also randomly inserted noisy samples (indicated by larger purple dots) that have the wrong label. We tackled this problem with a simple NN that had one hidden layer with four neurons. We applied CC and O2U after the model converged with 100 training epochs. Figure 4(b) shows a heat map of averaged predictive entropy of each samples. As expected, the easy samples have smaller averaged predictive entropy during the O2U process, while the harder samples have larger averaged predictive entropy. The noisy samples have the largest predictive entropy since whenever the model went from

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

overfitting to underfitting, their predictive entropy spiked. Using CC and the strategy of Figure 1(c), we classified samples into easy, hard, and noisy, and compared them with their original categories (easy, hard, and noisy). Figure 4(c) demonstrates a confusion matrix with horizontal axis representing the classified results and vertical axis representing the original setting. As can be seen, our proposed method was able to classify most of the samples correctly. 98% easy samples were identified correctly, and for hard and noisy samples, the percentages are 87% and 82%, demonstrating the effectiveness of the proposed method.

2) Illustration of Curriculum Classification: We first illustrate the process of the CC algorithm to help explain its effectiveness. As training proceeds, the deep image features should be more separable according to their ISUP grades in feature space. In other words, if we use k-means to group them in a fully unsupervised fashion, the label diversity within each group should decrease. We defined a metric called “grade concentration” to measure the ISUP diversity for each cluster group at each iteration in PathAL. The “grade concentration” was calculated as an average negative entropy of the ISUP grade distribution of each group. Figure 3(a) demonstrates the t-SNE plot in the deep feature space. Each point in the figure represents a slide whose color indicates its ISUP grade. As expected, in the early iterations of PathAL, different ISUP gardes were not well separated in the feature space, indicating the model was not able to achieve high accuracy in its predictions. As training progressed, ISUP images became more separable according to their grades. As a result, the subsets clustered by k-means methods would have higher “grade concentration”. Figure 3(b) depicts the trend of “grade concentraion.” The insets of the figure demonstrate a typical ISUP distribution for the clusters. At the beginning of the training iterations, the ISUP grades were more spread, while at the end of the training, each cluster had lower label diversity.

3) Illustration of O2U Cyclic Training: In this illustration, we demonstrate the cyclic training in the O2U process. After the model converged in each iteration, we adjusted the learning rate periodically so that the network could transition from overfitting to underfitting cyclically. The learning rate was adjusted based on a cosine annealing function in each cyclic round as,

$$lr = lr_{min} + \frac{1}{2}(lr_{max} - lr_{min})(1 + \cos(\frac{T_{cur}}{T_{max}}\pi)) \tag{7}$$

where T_{max} was the epoch for one cycle, lr_{min} and lr_{max} were the minimum and maximum learning rates in one cycle.

We monitored the training loss for every sample in L_i, and the predictive entropy for every sample in U_i. This process is illustrated in Figure 3(c)–(d). After the cyclic training, the samples were ranked according to their losses and predictive entropy variation. The samples were plotted in terms of three groups: top 0%–40% ranked samples, top 40%–80% ranked samples, and the rest of the samples. It was observed that the training losses and predictive entropy fluctuated with the cyclical adjustment of the learning rate. The training losses of the top 40% of samples fluctuated drastically during the cyclical training when compared to the rest of the samples, which may indicate they were noisy or hard samples

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(see Figure 3(c)). It is also observed in Figure 3(d) that the top 40% samples ranked for predictive entropy did not change during early iterations, which implies that the samples in L_i contained limited information for the model to classify these samples. As the training proceeded, we observed that the predictive entropy of the top 40% started to fluctuate, as the samples in L_i contained more information about the samples in U_i, even for the most uncertain group.

4) PathAL Performance: We compared our proposed PathAL with the four AL baselines mentioned in Section IV-D. Figure 5(a) demonstrates the comparison results of QWK on ISUP grade prediction between PathAL and the baselines. The QWK was calculated as the average performance of the four folds. The dotted black line shows the fully supervised learning performance. As shown in the figure, PathAL improved the QWK performance with fewer required annotations. It achieved a higher QWK compared with the full training set supervision baseline with only 60% of the annotations required from the expert. The VAAL method performed the second best with the PANDA dataset, achieving comparable performance with PathAL at the early stage of AL (when the annotated samples were limited). However, it only achieved the fully-supervised baseline with 70% of annotated data. While PathAL did not outperform the fully-supervised baseline by a large margin, we argue that it achieved slightly better performance because it discarded the noisy labels. Here, we report the mean value across the four cross-validation runs. However, our model that achieves 89.5% QWK on the training data produces a 93.1% QWK on the final private leaderboard, which is among top 10 results in the challenge. As our work aims to demonstrate that we can achieve similar performance to state-of-the-art methods with only 60% of labeled data, we argue that our results can achieve the same state-of-the-art performance with much less data annotation burden. To illustrate the effectiveness of each component in PathAL, we performed an ablation study in Section IV-E7.

To determine whether: 1) the samples we discarded in each iteration were noisy with low QWK, 2) the samples we asked the expert to annotate were “informative” with low QWK, and 3) the samples we added with their “pseudo-label” were correct with high QWK, we plotted the QWK for each group during the training process in Figure 5(b). It is observed that the noisy sample group N_i had a relatively low QWK even though their labels were used for training, while the confident predicted samples in U_i have a much higher QWK. Samples in I_i that were re-annotated by the expert had low QWK, indicating that those samples were most “informative” to the current models, and would improve the model’s performance by a large margin if added to L_i with annotations. Both QWK of I_i and N_i increases throughout the training and the curve for Noisy N_i becomes higher than Informative I_i at the later stage of the training. We argue that this is because N_i are samples from the labeled dataset L_i and they are directly used to train the model in the *ith* iteration, while I_i, C_i are samples from the unlabeled dataset U_i and they are not used to train the model. At the later stage of the training, the model became well-trained on easy samples and started to memorize the label of noisy samples (*i.e.*, overfitting), which is why we observed that Noisy N_i become higher than Informative I_i in the figure.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5) Keeping Noisy Samples in a Pool: As mentioned in Section III-D, during the PathAL process, we do not discard the noisy samples completely. Instead, we keep them in a pool and examine if they need to be added back throughout the training process. To analyze how this mechanism works, we calculated the percentage of the noisy samples that were returned to training in later iterations. Specifically, assume we have N_i samples that are identified as noisy samples in the ith iteration, among them we have n_{i+1} samples returned back for training in the $(i+1)th$ iteration, n_{i+2} samples in the $(i+2)th$ iteration, …, then the percentage is defined as,

$$P_i = \sum_{m=i+1} n_m / N_i . \tag{8}$$

We plot the average percentage across the four runs in Figure 5(c). As can be seen in the figure, there were around 25% noisy samples in the initial iteration that would return to training during the whole process, indicating the necessity of keeping them in a pool instead of discarding them. As the iteration continued, the percentage dropped quickly, indicating that the model became more capable of distinguishing noisy samples. The shaded area indicates the standard deviation among four runs. As expected, it became smaller as we went through the training process.

6) Noisy vs. Expert Pathologists: The labels provided by the pathologists might vary depending on their experience. In order to simulate the impact of acquiring labels from an inexperienced annotator, we applied random noise to the labels in each iteration. We randomly change the ground truth labels for 15% and 30% of the training set to have an incorrect label. In our experiments, we did not constrain the distance between the noisy label and ground truth label, instead we uniformly sample from the ISUP grades. Figure 5 (d) shows how a noisy annotator affects the performance of PathAL and random sampling. We observed that PathAL constantly outperformed random sampling, although the relative performance with noisy labels fell compared to experts. As the percentage of noisy labels increased, PathAL performance was closer to the random sampling, as expected.

7) Ablation Study: To illustrate the effectiveness of each component in PathAL, we performed an ablation study with varying amounts of labeled data until 60% of expert annotations were added to L_i, when PathAL would have access to all sample labels either through expert's annotations or pseudo-assigned labels. Table I demonstrates the results when only parts of PathAL components were used. The QWK is shown as an average for four different folds with a calculated standard deviation.

From the table, we observed that using a simple predictive entropy measure by O2U in row 2 (Entropy (O2U)) to select the most "informative" samples improved the QWK by 1.3% compared with the random selection baseline, while PathAL improved the random baseline by 2.4%. When we excluded the noisy samples in row 3 (Entropy + Noisy (O2U)), we saw QWK improve by 0.3%, while adding the high-confidence predictive samples by self-assigned pseudo-labels (row four Entropy + Conf Preds) improved the QWK by 0.4%. To illustrate whether the O2U and CC component helped with the selection of N_i, I_i, C_i, we implemented PathAL with selection based on the predictive entropy (row 5 PathAL (w/o

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

O2U & CC)), *i.e.*, our selection was solely based on the model's predictive entropy in this experiment. Specifically, we selected the top ranked samples in L_i based on the model's predictive entropy as N_i (the high entropy of theses labeled samples indicates they are most likely to be noisy samples), the top ranked samples in U_i as I_i (the high entropy of these unlabeled samples indicates the model is uncertain about their prediction), and the bottom ranked samples as C_i (the low entropy of these unlabeled samples indicates the model is pretty certain about their prediction). We showed that using O2U and CC in combination as a sample selective strategy improved the QWK by 1.3%, indicating the effectiveness of their roles in PathAL.

8) Pathologist Validation: To validate our algorithm for detecting easy, noisy, and hard samples, two pathologists (AS and DS) with expertise in Gleason grading performed an independent reader study. Specifically, we provided three groups of slides that were labeled as easy, noisy, and hard by our algorithm. Each group consisted of 100 slides. The pathologists were asked to give final ISUP grades without knowing the ground truth labels provided by the dataset. As shown in Table II, we measured the QWK of each group, between the ground truth labels and each reader. As expected, the QWK of the "easy" group was very high, with highest value 1 that indicates 100% agreement between the pathologist and the ground-truth label for the 100 samples. Conversely, the QWK of the "noisy" group was −0.14 and −0.06 between the ground truth labels and the two pathologists, respectively. The QWK between the two pathologists was 0.7922, which was much higher compared to the ground truth labels. This demonstrates that the high variance between the dataset labels and pathologist readings can be explained by noise in the dataset labels. The QWK of the "hard" samples was 0.10 and 0.15 for the two pathologists, which was slightly higher than that of the "noisy" group. Comparing the readings between the two pathologists, we observed 0.58 QWK. This was lower than the 0.79 QWK of the "noisy" group, indicating the intrinsic hardness of grading these samples. Although it is not as low as comparing with the ground truth, we believe it can be partially explained by the fact that the two pathologists have gone through similar training and review slides using the same grading criteria. Furthermore, by visually inspecting the discordant slides in the "hard" group, we found that some of the slides were also likely to have incorrect labels. In other slides, the cancerous regions were either small or ambiguous, so it was difficult for the pathologists to spot the cancerous areas or reach a consensus. In general, we found that our algorithm robustly distinguished between "easy" and "hard & noisy" groups, and was able to tell "hard" and "noisy" samples apart in a reasonable way. However, there is still room for improvement in distinguishing between "hard" and "noisy" samples. For more details, please refer to the Appendix.

V. LIMITATIONS

There are many hyper-parameters in our approach, and to a degree, these parameters require task-sensitive tuning. A potential direction for future work is to develop a hyper-parameter robust version of PathAL. Another limitation of our approach is the extra training time needed for PathAL. The computational overhead compared to other AL baselines is dominated by the O2U process. In the PANDA dataset, the O2U process can take up to 20

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

minutes to collect the statistically meaningful averaged training loss and predictive entropy, while other sampling strategies of common AL baselines require several minutes. However, other AL methods are not able to distinguish noisy and hard samples, which is a key innovation of our work. Furthermore, our model does not impact inference time as the CC algorithm and O2U process are all performed during training.

VI. CONCLUSION

In this paper, we have proposed PathAL, a novel AL framework for histopathology image analysis. Unlike prior studies in medical image AL, which only consider the most "informative" samples to be added in each iteration, PathAL also heuristically excludes noisy samples and adds confident predictive samples with self-assigned pseudo-labels. Specifically, the combination of a CC algorithm and an O2U process was used to detect noisy, confident, and informative samples. Our proposed method achieved competitive performance while requiring only 60% of samples to be annotated, compared to the fully supervised learning baseline on the PANDA challenge. Extensive experiments conclude the effectiveness of each component in PathAL. We expect to apply PathAL to other histopathological image analysis scenarios in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

[1]. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, and Sánchez CI, "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017. [PubMed: 28778026]

[2]. Yang C-W, Lin T-P, Huang Y-H, Chung H-J, Kuo J-Y, Huang WJ, Wu HH, Chang Y-H, Lin AT, and Chen K-K, "Does extended prostate needle biopsy improve the concordance of gleason scores between biopsy and prostatectomy in the taiwanese population?" Journal of the Chinese Medical Association, vol. 75, no. 3, pp. 97–101, 2012. [PubMed: 22440266]

[3]. Cheplygina V, de Bruijne M, and Pluim JP, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," Medical image analysis, vol. 54, pp. 280–296, 2019. [PubMed: 30959445]

[4]. Budd S, Robinson EC, and Kainz B, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," arXiv preprint arXiv:1910.02923, 2019.

[5]. Dgani Y, Greenspan H, and Goldberger J, "Training a neural network based on unreliable human annotation of medical images," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 2018, pp. 39–42.

[6]. Xue C, Dou Q, Shi X, Chen H, and Heng P-A, "Robust learning at noisy labeled medical images: Applied to skin lesion classification," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019, pp. 1280–1283.

[7]. Wen S, Kurc TM, Hou L, Saltz JH, Gupta RR, Batiste R, Zhao T, Nguyen V, Samaras D, and Zhu W, "Comparison of different classifiers with active learning to support quality control in nucleus segmentation in pathology images," AMIA Summits on Translational Science Proceedings, vol. 2018, p. 227, 2018.

[8]. Gal Y, Islam R, and Ghahramani Z, "Deep bayesian active learning with image data," arXiv preprint arXiv:1703.02910, 2017.

[9]. Houlsby N, Huszár F, Ghahramani Z, and Lengyel M, "Bayesian active learning for classification and preference learning," ArXiv, vol. abs/1112.5745, 2011.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[10]. Kuo W, Häne C, Yuh E, Mukherjee P, and Malik J, “Cost-sensitive active learning for intracranial hemorrhage detection,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 715–723.

[11]. Konyushkova K, Sznitman R, and Fua P, “Geometry in active learning for binary and multi-class image segmentation,” Computer vision and image understanding, vol. 182, pp. 1–16, 2019.

[12]. Yang L, Zhang Y, Chen J, Zhang S, and Chen DZ, “Suggestive annotation: A deep active learning framework for biomedical image segmentation,” in International conference on medical image computing and computer-assisted intervention. Springer, 2017, pp. 399–407.

[13]. Smailagic A, Noh HY, Costa P, Walawalkar D, Khandelwal K, Mirshekari M, Fagert J, Galdrán A, and Xu S, “Medal: Deep active learning sampling method for medical image analysis,” arXiv preprint arXiv:1809.09287, 2018.

[14]. Ozdemir F, Peng Z, Tanner C, Fuernstahl P, and Goksel O, “Active learning for segmentation by optimizing content information for maximal entropy,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2018, pp. 183–191.

[15]. Sourati J, Gholipour A, Dy JG, Kurugol S, and Warfield SK, “Active deep learning with fisher information for patch-wise semantic segmentation,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2018, pp. 83–91.

[16]. Wang K, Zhang D, Li Y, Zhang R, and Lin L, “Cost-effective active learning for deep image classification,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 12, pp. 2591–2600, 2016.

[17]. Karimi D, Dou H, Warfield SK, and Gholipour A, “Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis,” Medical Image Analysis, vol. 65, p. 101759, 2020. [PubMed: 32623277]

[18]. Koh PW and Liang P, “Understanding black-box predictions via influence functions,” arXiv preprint arXiv:1703.04730, 2017.

[19]. Lee K-H, He X, Zhang L, and Yang L, “Cleannet: Transfer learning for scalable image classifier training with label noise,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5447–5456.

[20]. Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, Tsang I, and Sugiyama M, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” in Advances in neural information processing systems, 2018, pp. 8527–8537.

[21]. Guo S, Huang W, Zhang H, Zhuang C, Dong D, Scott MR, and Huang D, “Curriculumnet: Weakly supervised learning from large-scale web images,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 135–150.

[22]. Jiang L, Zhou Z, Leung T, Li L-J, and Fei-Fei L, “Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels,” in International Conference on Machine Learning, 2018, pp. 2304–2313.

[23]. Huang J, Qu L, Jia R, and Zhao B, “O2u-net: A simple noisy label detection approach for deep neural networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3326–3334.

[24]. Goldberger J and Ben-Reuven E, “Training deep neural-networks using a noise adaptation layer,” 2016.

[25]. Xiao T, Xia T, Yang Y, Huang C, and Wang X, “Learning from massive noisy labeled data for image classification,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2691–2699.

[26]. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, and Rabinovich A, “Training deep neural networks on noisy labels with bootstrapping,” arXiv preprint arXiv:1412.6596, 2014.

[27]. Maaten LVD and Hinton GE, “Visualizing data using t-sne,” Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.

[28]. Tan M and Le QV, “Efficientnet: Rethinking model scaling for convolutional neural networks,” arXiv preprint arXiv:1905.11946, 2019.

[29]. Speier W, Li J, Li W, Sarma K, and Arnold C, “Image-based patch selection for deep learning to improve automated gleason grading in histopathological slides,” bioRxiv, 2020.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[30]. Gal Y and Ghahramani Z, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in Proceedings of The 33rd International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, Balcan MF and Weinberger KQ, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059. [Online]. Available: http://proceedings.mlr.press/v48/gal16.html

[31]. Sinha S, Ebrahimi S, and Darrell T, “Variational adversarial active learning,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5972–5981.

[32]. Sener O and Savarese S, “Active learning for convolutional neural networks: A core-set approach,” arXiv preprint arXiv:1708.00489, 2017.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

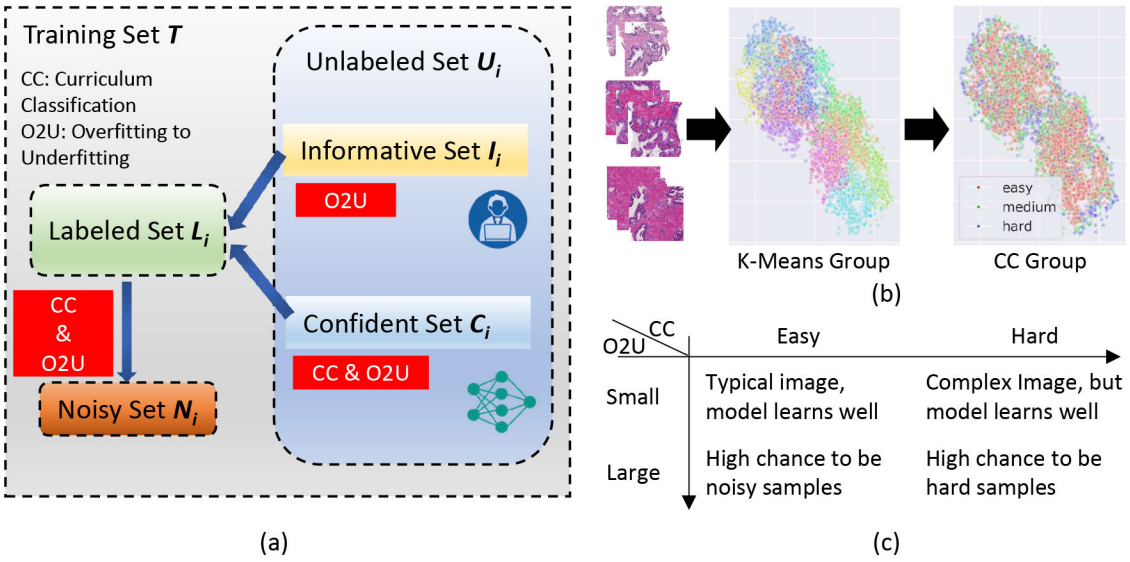

Fig. 1.

(a) Schematics of our proposed PathAL. The core algorithm of PathAL consists of three steps in the ith iteration: 1) discarding noisy samples N_i; 2) requesting human experts to annotate informative samples I_i and adding them to L_{i+1}; and 3) adding confident predictictive samples C_i with their "pseudo-labels" to L_{i+1}. The curriculum classification (CC) algorithm and overfitting to underfitting (O2U) monitor are used to select N_i, I_i, C_i. (b) Illustration of the CC algorithm. Tissues from one slide are mapping to one single point in deep feature space, where K-Means Clustering is used to group them in subsets. The CC algorithm is applied to each subset and the image complexity is classified as "easy", "medium" or "hard" based on their local density. (c) Principles on how to determine N_i and C_i based on CC and O2U results. A sample that is classified as "easy" based on its complexity but has large training loss variation is more likely to be incorrectly annotated. If it is classified as "hard" for its complexity, it is more likely to be a difficult sample. If a sample's complexity is classified as "easy" and the variation of its predictive entropy is low by the current model, we will have a higher confidence that the current prediction is correct.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

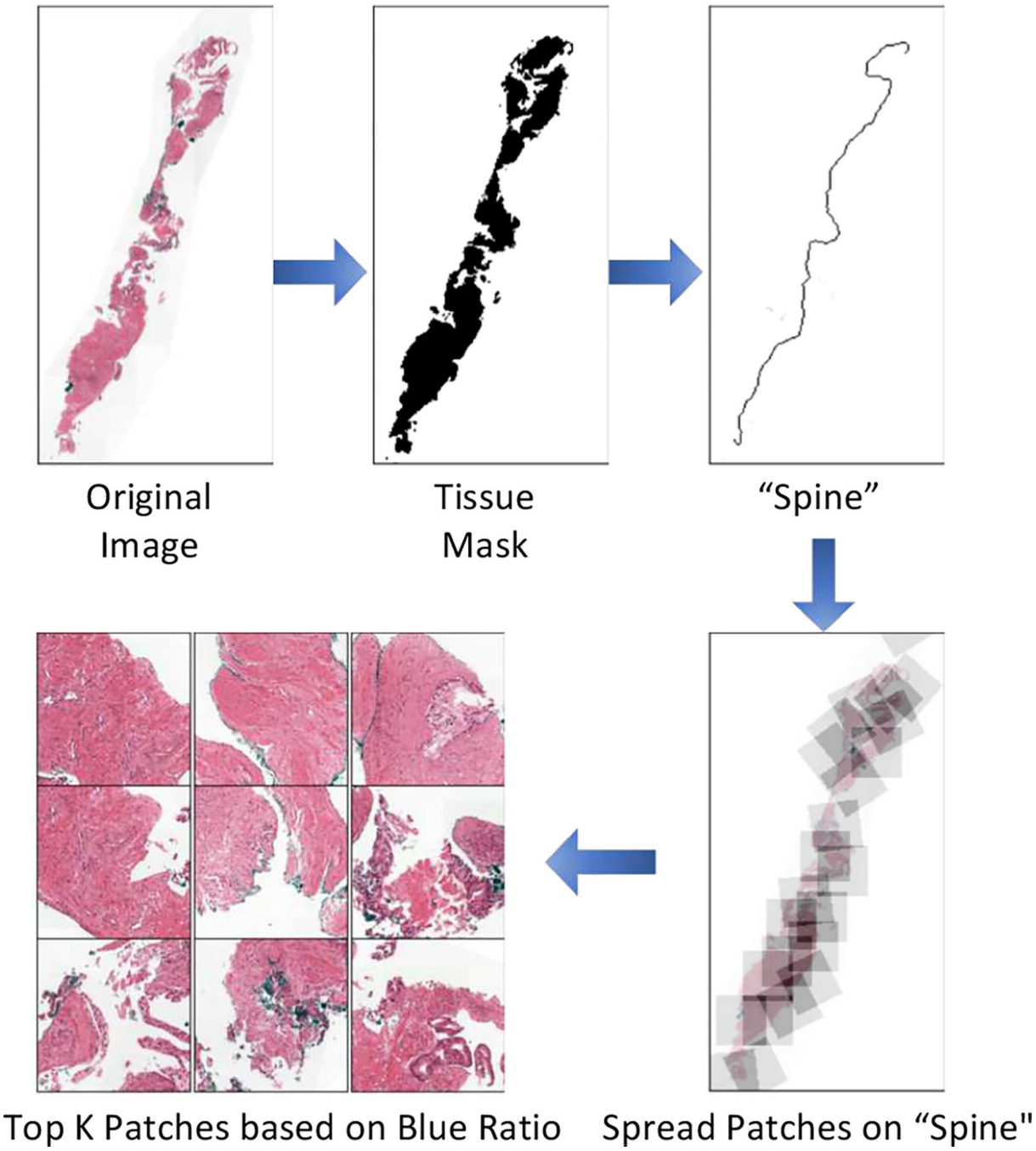

Fig. 2.

Illustration of data pre-processing steps. A binary mask of tissue is first extracted; then the mid-line is found using morphological closing; after that, the mid-line is partitioned to form patches based on the batch size and overlap; finally, the blue ratios of patches are calculated and the top *k* patches are selected.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

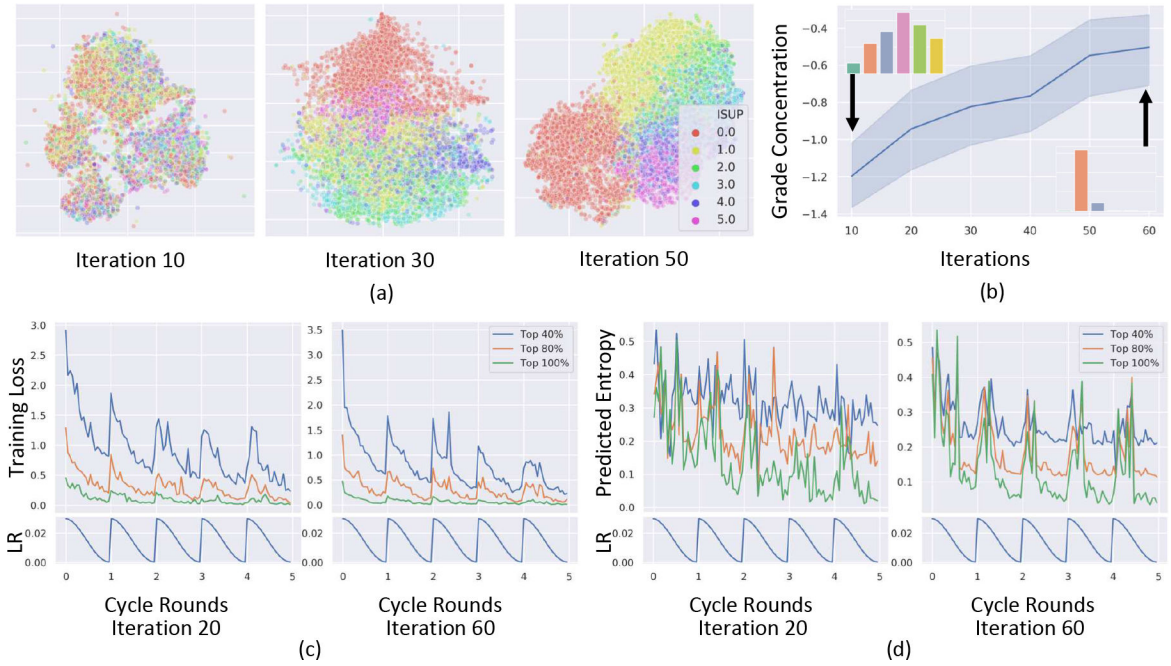

Fig. 3.

(a) t-SNE plot in deep feature space. Each point in the figure represents a slide whose color indicates its ISUP grade. As training progressed, different ISUP grades became more separable in the deep feature space, indicating the model captured more essential information to make correct predictions. (b) The trend of “grade concentraion” that measured the ISUP grade distribution within subsets clustered by k-means. The insets of the figure demonstrate typical ISUP distributions for the subsets. At the beginning of training, the ISUP grades were more diffuse, while at the end of the training, each cluster concentrated on fewer grades. (c)(d) The training loss for every sample in L_i, and predictive entropy for every sample in U_i during the O2U process.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

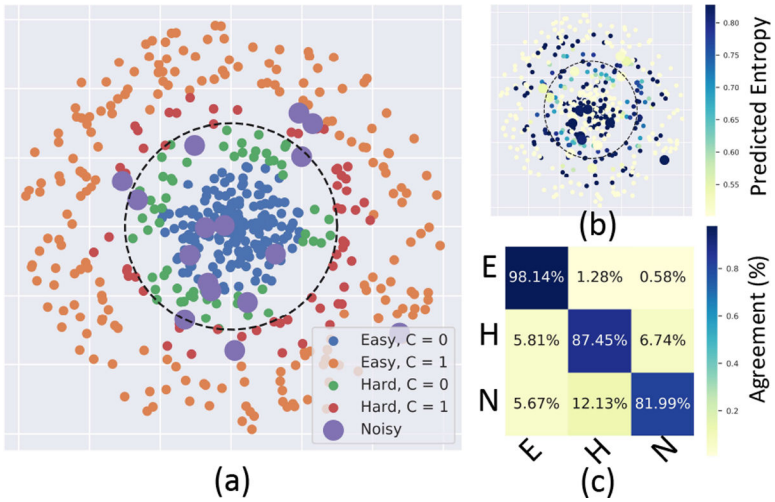

Fig. 4.

(a) Toy example for a 2-class classification. The dotted circle indicates the decision boundary while inside the circle we have class 0 (C=0) and class 1 (C=1) out side the circle. Samples that are far away from the decision boundary are considered as easy samples (blue for C=0 and orange for C=1), while samples that are closer to the decision boundary are considered as hard samples. We also randomly insert noisy samples (indicated by larger purple dots) that have wrong labels. (b) A heat map of averaged predictive entropy of each samples during the O2U process. (c) A confusion matrix of easy, hard, and noisy samples with horizontal axis representing the classified results and vertical axis representing the original categories.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

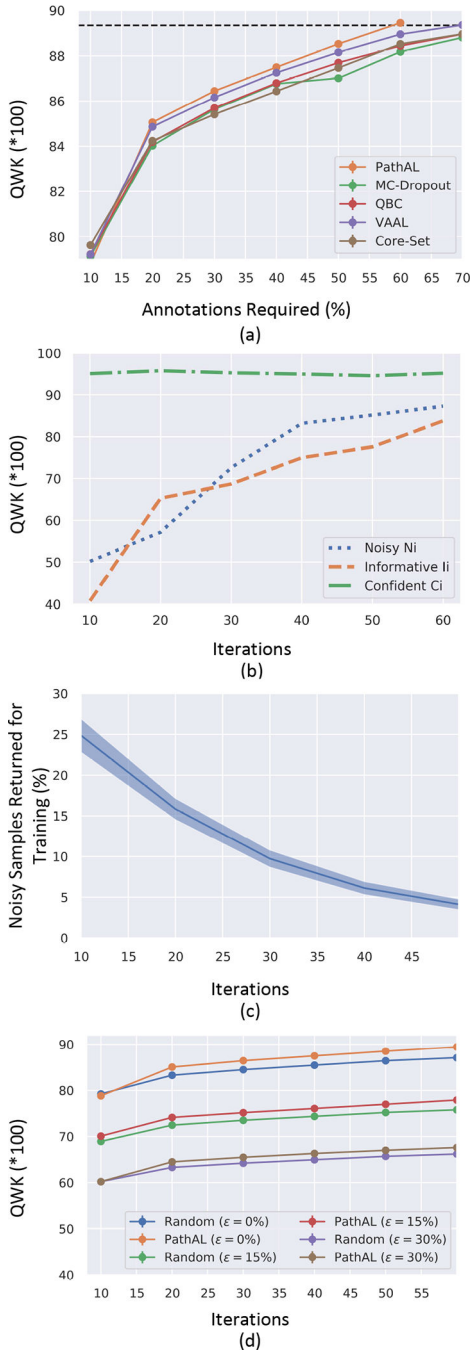

Fig. 5.

(a) Performance comparison between PathAL and other AL baselines. (b) QWK for each group (N_i, C_i, I_i) during the training process. (c) Percentage of noisy samples that returned for training in later iterations. (d)Performance comparison between PathAL and random sampling with noise samples injection.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I

Ablation Study of PathAL.

	QWK
Random	87.1 ± 0.6
Entropy (O2U)	88.4 ± 0.5
Entropy + Noisy (O2U & CC)	88.6 ± 0.3
Entropy + Conf Preds (O2U & CC)	88.8 ± 0.5
PathAL (w/o O2U & CC)	88.6 ± 0.4
PathAL	**89.5 ± 0.5**

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE II

Pathologist Reader Study.

	Easy	Hard	Noisy
Ground Truth vs. Reader1	1.0000	0.1074	−0.1438
Ground Truth vs. Reader2	0.9494	0.1488	−0.0591
Reader1 vs. Reader2	0.9494	0.5805	0.7922