Research article

Using deep learning in pathology image analysis: A novel active learning strategy based on latent representation

Yixin Sun[1,2]**, Lei Wu**[1,2,*]**, Peng Chen**[3]**, Feng Zhang**[4] **and Lifeng Xu**[4]

[1] School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China

[2] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 314099, China

[3] School of Computer and Software Engineering, Xihua University, Chengdu 611731, China

[4] The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou 324000, China

* **Correspondence:** Email: wulei@uestc.edu.cn; Tel: +8613084443881.

Abstract: Most countries worldwide continue to encounter a pathologist shortage, significantly impeding the timely diagnosis and effective treatment of cancer patients. Deep learning techniques have performed remarkably well in pathology image analysis; however, they require expert pathologists to annotate substantial pathology image data. This study aims to minimize the need for data annotation to analyze pathology images. Active learning (AL) is an iterative approach to search for a few high-quality samples to train a model. We propose our active learning framework, which first learns latent representations of all pathology images by an auto-encoder to train a binary classification model, and then selects samples through a novel ALHS (Active Learning Hybrid Sampling) strategy. This strategy can effectively alleviate the sample redundancy problem and allows for more informative and diverse examples to be selected. We validate the effectiveness of our method by undertaking classification tasks on two cancer pathology image datasets. We achieve the target performance of 90% accuracy using 25% labeled samples in Kather's dataset and reach 88% accuracy using 65% labeled data in BreakHis dataset, which means our method can save 75% and 35% of the annotation budget in the two datasets, respectively.

Keywords: active learning; pathology images classification; latent representation; samples selection; ALHS strategy

1. Introduction

Cancer affects a large population [1] and is a leading cause of death worldwide [2]. An accurate and timely cancer diagnosis is a critical first step in treating every cancer patient. Pathological diagnosis is generally made by a human pathologist using a microscope to observe the stained specimens on a slide glass. Due to the rapid development of digital microscopy in recent years, it is possible to digitize histology slides quickly and at a high resolution [3]. Pathological images contain a wealth of phenotypic information essential in diagnosing and classifying cancer. However, most countries worldwide continue to encounter a pathologist shortage [4, 5], significantly hindering the timely diagnosis and effective treatment of cancer patients. Each whole slide image (WSI) contains a large amount of information and may contain tens of thousands of image patches. Relying solely on the pathologist's visual inspection for analyses such as cancer detection, tumor staging, and grading would take a lot of time and effort, since it is very time-consuming for pathologists to accurately annotate an abundance of image patches in a large WSI dataset. For cancer patients, waiting for diagnosis and treatment results can also take a long time, which may lead to missing the best time for treatment and not receiving proper treatment.

Fortunately, advances in artificial intelligence technology offer several practical tools to automate or assist in diagnosing pathology, such as advanced tools proposed by [6–10], promising to improve the current dilemma of the lack of pathologists. Machine learning, especially deep learning technology [11], has achieved extraordinary performance in many fields, and its success relies heavily on large-scale annotated training examples. One of the bottlenecks of deep learning methods in computer vision is "the need for large annotated datasets" [12]. If we want to replicate the success of machine learning in the medical field, the problem of insufficiently labeled samples is the first thing to be solved. However, generating high-quality training datasets is challenging, as it is a labor-intensive manual process that requires the input of domain experts, who have limited time and high costs. How can we reduce the burden on pathologists and save the research budget? We work to solve these problems by finding solutions in the domain of active learning.

Active learning [13], which is a branch of machine learning, attempts to achieve the best possible performance of the model using as few, high-quality sample annotations as possible. A typical active learning procedure is shown on Figure 1. The model is randomly initialized at the beginning. At each training step, it selects some data from the unlabeled dataset for labeling and then retrains the model to obtain better performance. Coreset [14], Bayesian AL [15], and VAAL [16] are several active learning state-of-the-art methods in artificial intelligence.

In the medical field, a small number of scholars are also engaged in active learning research. Halder and Kumar [17] proposed an active learning method using gene expression data utilizing a rough-fuzzy classifier for cancer sample classification. Mahapatra et al. [18] proposed an active learning framework to select most informative samples using conditional generative adversarial networks (cGANs) to generate realistic chest X-ray images. Furthermore, similar to our work, [19] proposed an AL acquisition method that uses data grouping based on imaging features and model prediction uncertainty and then used it in the TIL binary classification task and applied it to pathology images. By contrast, our approach was validated not only in a binary classification task, but also in a multi-classification task.

In this study, we propose our active learning framework in Figure 1, which first learns latent

representations of all pathology images by an auto-encoder and trains a binary classification model to distinguish latent representations from labeled and unlabeled samples, then serves as the basis for sample selection. We find that in the case of good feature mapping, the closer the inputs of the approximate points into the neural network, the more similar the probability outputs are. Taking advantage of this attribute, we introduce a novel sample selection method called ALHS (Active Learning Hybrid Sampling). First, the dataset needs be sorted by the probability values given by the discriminative model and then divided into K (budget size) parts in order, and one sample is selected for expert labeling in each part of the dataset. We have four methods to select samples for each category to better select valuable samples: random, least confidence [20], margin sampling [21], and entropy [22]. Through our selection method, we avoid choosing more samples in similar regions, thus effectively avoiding the sample redundancy problem. Furthermore, we adopt a series of dimensionality reduction methods to visualize the samples selection. To the best of our knowledge, our study is the first attempt to apply active learning to two pathology image datasets, while explicitly considering the sample redundancy problem as well as the visualization of selection result based on latent representations.

The specific goal of this study is to propose an active learning framework applied to the pathology image dataset to reduce the heavy burden on pathologists. We validate the effectiveness of our method on two pathology image datasets. Savings of 75% and 55% of the sample labeling budget were achieved while reaching the target performance in the two datasets, respectively.
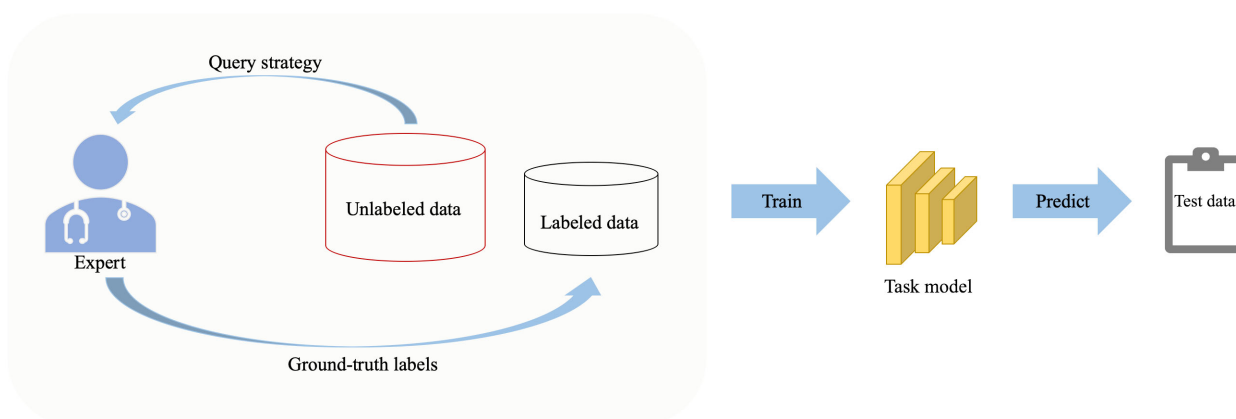

Figure 1. A typical active learning procedure. There is a situation in which unlabeled data is abundant, but manual labeling is expensive. In active learning, the algorithm selects a subset of examples to be labeled by human annotators instead of labeling an entire dataset.

2. Materials and methods

2.1. Data acquisition

2.1.1. Collection of textures in colorectal cancer histology (Kather's)

The multi-class colorectal histology image dataset [23] proposed by Kather et al. was used. The dataset contains 5000 images of 625 images for each class of fixed dimension 150 px × 150 px, which

is approximately 74 μm × 74 μm. All images are RGB, 0.495 μm per pixel, digitized with an Aperio ScanScope (Aperio/Leica biosystems), magnification 20x. Histological samples are obtained from the pathology archive at the University Medical Center Mannheim (Heidelberg University, Mannheim, Germany). The dataset contains the following eight classes of histology images: adipose, debris, lymphoma, mucosa, complex, stroma, tumor, and empty. Figure 2 shows a sample from each type of colorectal cancer histology. We divided this dataset into a training set, a validation set and a test set with the numbers of 3600, 400 and 1000, respectively.
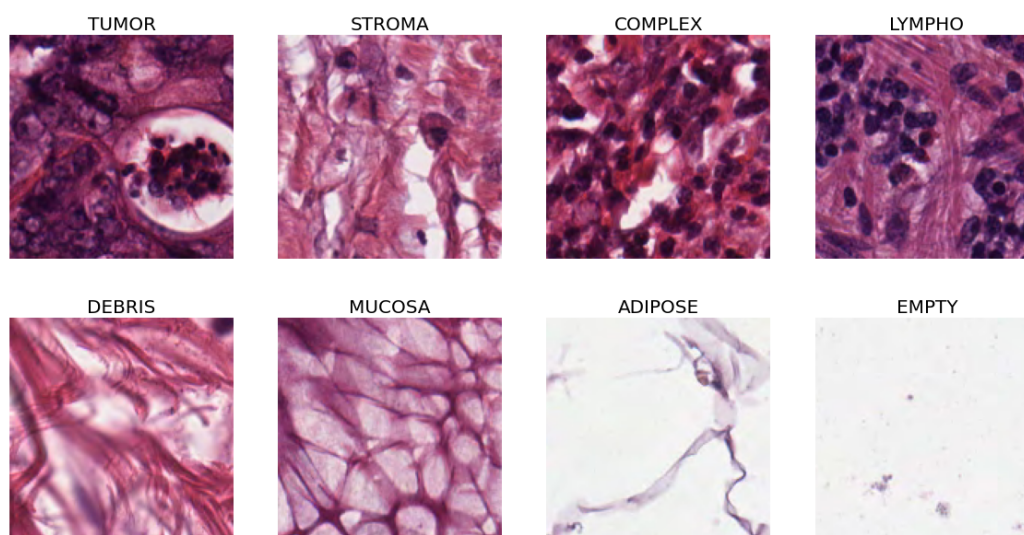

Figure 2. Samples from each type of colorectal cancer histology on Kather's dataset: Adipose, debris, lymphoma, mucosa, complex, stroma, tumor, and empty.

2.1.2. Breast cancer histopathological database (BreakHis)

Moreover, the BreakHis [24] dataset was used in this work, which is composed of 7909 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40 X, 100 X, 200 X, and 400 X). It contains 2480 benign and 5429 malignant samples. Each slide of breast tumors is stained with hematoxylin and eosin(HE) and collected by SOB. The dataset comprises eight types of benign and malignant tumors. The four benign tumors types are as follows: Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), and Tubular Adenona (TA). The four malignant tumors types are as follows: Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC) and Papillary Carcinoma (PC). The number of samples for each magnification factor is provided in Table 1. We divided this dataset into a training set, a validation set and a test set with the full images of 62, 8 and 12 patients, respectively.

Table 1. The number of breast cancer histology images samples per class on BreakHis dataset.

CLASS	SUB-CLASS	Magnification factor				Total
		40 X	100 X	200 X	400 X	
Benign	Adenosis	114	113	111	106	444
	Fibroadenoma	253	260	264	237	1014
	Phyllodes Tumor	109	121	108	115	453
	Tubular Adenona	149	150	140	130	569
Malignant	Ductal Carcinoma	864	903	896	788	3451
	Lobular Carcinoma	156	170	163	137	626
	Mucinous Carcinoma	205	222	196	169	792
	Papillary Carcinoma	145	142	135	138	560
Total		1995	2081	2013	1820	**7909**

2.2. Basic methods of active learning

- **Random:** Select samples randomly.
- **Least Confidence:** The least confident method is to select those samples with the minimal maximum probability for labeling, as described by the following mathematical formula [20]:

$$x_{LC}^* = \arg\max_{x \in U} \left(1 - P_\theta(\hat{y} \mid x)\right) = \arg\min_{x \in U} \left(P_\theta(\hat{y} \mid x)\right), \tag{2.1}$$

where $\hat{y} = \arg\max_y P_\theta(y \mid x)$, θ denotes the set of parameters of a trained machine learning model, and $\hat{y}$ is the class with the highest probability of model prediction for x.

- **Margin Sampling:** Margin sampling refers to the selection of sample data that are either highly susceptible to being judged into one or two categories or that have a slight difference in the probability of being considered into two types. Margin sampling is the selection of the sample with the most minor difference between the largest and second-largest probabilities of model prediction, as described by the following mathematical formula [21]:

$$x_{Margin}^* = \arg\max_{x \in U} \left(P_\theta(\hat{y}_1 \mid x) - P_\theta(\hat{y}_2 \mid x)\right), \tag{2.2}$$

where $\hat{y}_1$ and $\hat{y}_2$ denote that for x the model predicts the maximum possible class and the second most probable class, respectively.

- **Entropy:** A general uncertainty sampling strategy uses Shannon entropy of the prediction probability as an uncertainty measure. Entropy is an information-theoretic measure that represents the amount of information needed to "encode" a distribution. As such, it is often thought of as a measure of uncertainty or impurity in machine learning, as described by the following mathematical formula [22]:

$$x_{Entropy}^* = \arg\max_{x \in U} \left(-\sum_i P_\theta(y_i \mid x) \cdot \ln P_\theta(y_i \mid x)\right). \tag{2.3}$$

2.3. Our active learning framework

Our active learning strategy is divided into three steps: the first step is to learn the latent representation using an auto-encoder; the second step is to identify whether the low-dimensional representation of a sample is from a labeled or unlabeled dataset using a discriminative network; and the third step is to use the probabilistic output of the discriminative network to select samples with value using the ALHS method.
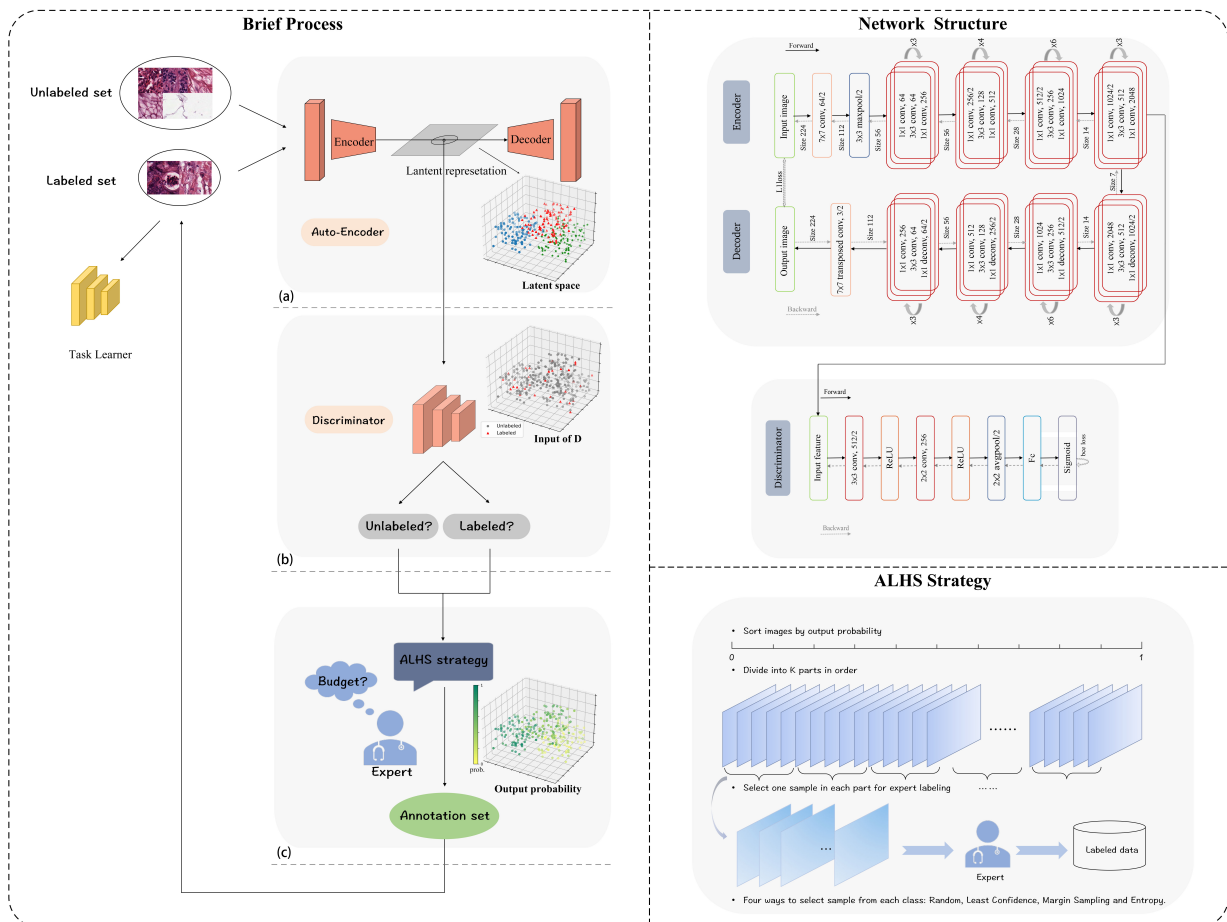

Figure 3. Our Active Learning Framework. We first learn latent representations of all pathology images by an auto-encoder to train a binary classification model and then select samples through a novel ALHS(Active Learning Hybrid Sampling) strategy. (a) The auto-encoder network learns the low-dimensional latent spatial representation of all pathology images. (b) The discriminant model learns whether the latent representation of the pathology image is from the labeled or unlabeled dataset. (c) Using ALHS strategy to select samples for experts labeling. ALHS Strategy: sort data by output probability; divide into K(= budget) parts in order; select one sample in each part.

Three key components make up my active learning framework: an auto-encoder, a discriminator model, and an ALHS selection strategy. Figure 3 displays the entire process of our active learning framework.

2.3.1. Feature extraction with latent representation of pathological images

Pathological images contain a wealth of phenotypic information. Therefore, we utilize an auto-encoder model to extract latent representations from the pathology images. Auto-encoder is an unsupervised deep learning algorithm that learns unlabeled data. We use auto-encoder for representation learning, where the encoder learns to embed the medical dataset in the low-dimensional space, and the decoder reconstructs the input image based on the low-dimensional representation. Mapping the data distribution into the latent space can effectively improve the performance with less computational cost.

The feature extraction portion of the Resnet50 [25] structure is chosen to implement the encoder in the auto-encoder network, and the decoder portion is the counterpart encoder transposed. The network structure diagram of the auto-encoder model is shown in the first grey block on the right of Figure 3. We set the batch size to 128, the number of iterations to 50,000 and the optimizer to Adam. Specifically, we set the learning rate of the backbone to 0.01 and a weight decay rate of 1/50,000 is set to inhibit overfitting, which can keep the weights of the neural network from becoming too large.

Formally, let $\Psi : \mathcal{X} \rightarrow Z$ be the mapping from the original input space to the learned latent representation (i.e., $z = \Psi(x)$). Let $\Phi : \mathcal{Z} \rightarrow \widetilde{\mathcal{X}}$ be the mapping from the latent feature representation to the reconstructed image, i.e., $\hat{x} = \Phi(z)$, and the model reconstructs the input image based on the low-dimensional representation z. The objective function of AE is the MAE loss function, i.e., L1 loss:

$$L_{AE} = L_1(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i| \tag{2.4}$$

where x denotes the original image, $\hat{x}$ denotes the reconstructed image and N represents the number of images. We use a stochastic gradient descent to optimize the objective function and save the model with the lowest loss. We expect our trained auto-encoder model to achieve excellent image reconstruction and its latent representation to provide a good representation input for the next discriminative model to distinguish between labeled and unlabeled datasets.

2.3.2. Discriminative approach prepares for sample selection

After learning latent representations and dimensionality reduction by the auto-encoder network, the pathology images are used as the input of a discriminative model. The discriminative model tries to distinguish the labeled samples from the unlabeled samples, where the label of the labeled data set is recorded as one, and the label of the unlabeled data is recorded as zero. Formally, with $\Psi : \mathcal{X} \rightarrow \hat{\mathcal{X}}$ being a mapping from the original input space to the learned latent representation, we define a binary classification problem with $\hat{\mathcal{X}}$ as our input space and $\mathcal{Y} = \{l, u\}$ as our label space, where l is the label for a sample being in the labeled set and u is the label for the unlabeled set.

The network structure diagram of the discriminative model is shown in the second grey block on the right of Figure 3. We set the batch size to 64, the number of epochs to 100 and the optimizer to Adam. Specifically, we set the learning rate to 0.01 and adopted equal-interval adjustment of learning rate measures to inhibit overfitting.

Formally, the labeled sample set is denoted as $\mathcal{L}$ and the unlabeled sample set is denoted as $\mathcal{U}$. All pathological images are used as the input of the discriminator D, and the labels of the labeled samples

are set to 1 and the labels of the unlabeled samples are set to 0. Let discriminator $D : \mathcal{Z} \rightarrow \mathcal{T}$, $\mathcal{Z}$ is the input space, which is the latent representation of the images learned by the encoder in auto-encoder, and $\mathcal{T} = \{l, u\}$ is the label space where l is the label of the labeled sample set, (i.e., the set of all 1 s,) and u is the label of the unlabeled sample set, (i.e., the set of all 0 s). Then, the objective function of the discriminator D is the cross-entropy loss function:

$$L_D = L_{CE} = \frac{1}{N} \sum_{i=1}^{N} - [t_i \cdot \log(p_i) + (1 - t_i) \cdot \log(1 - p_i)] \tag{2.5}$$

where t_i is the label of the sample 1 or 0 and p_i is the predicted probability that the model predicts the label of the sample to be t_i. We solve the classification problem over $\mathcal{U} \cup \mathcal{L}$ for every iteration of the active learning process by minimizing the objective function using a stochastic gradient descent and obtaining the best model $\hat{P}(t \mid \Psi(x))$ with the lowest training loss.

By observing the distribution of the results of the discriminative model, we find that the inputs of the approximate points into the neural network have similar probability outputs. If we select the top-K samples that satisfy

$$x = \underset{x \in \mathcal{U}}{\arg\max} \hat{P}(t = u \mid \Psi(x)) \tag{2.6}$$

as with other discriminative active learning methods [26], this may make the selected samples redundant and thus result in an ineffective task model. To overcome the sample redundancy problem, we designed a novel sampling method that considers the attribute of the discriminative model outputs.

2.3.3. Selecting samples by ALHS strategy for experts labeling

We take the latent representation of the pathology image dataset through the already trained AE and then distinguish the labeled and unlabeled datasets by a discriminator to obtain a probability value output. Obviously, the close input points have similar probability outputs in the discriminant neural networks. Thus, we cannot refuse some close input points to be selected for a single selection indicator at the same time, which leads to sample redundancy. Naturally, we thought that we could divide the unlabeled samples into K parts after arranging them in a certain order, in which K is the budget size, and select one of the samples in each sample for expert annotation, thus avoiding the selection of some redundant samples at similar output probabilities.

Based on the above, we propose a novel sample selection strategy - ALHS (Active Learning Hybrid Sampling). Our initial procedure for selecting samples is shown at the bottom right of Figure 3. The specific process is as follows: we first sort the dataset by probability values, then divide it into K parts in order, and select one sample in each part of the dataset for experts labeling. To better select valuable and informative samples, we have four ways to select samples from each class: random, least confidence [20], margin sampling [21] and entropy [22].

In summary, our sample selection strategy relies on an auto-encoder and a discriminator, which work together to perform image reconstruction and label prediction tasks, respectively. By leveraging the discriminative power of the discriminator, we are able to select representative samples from the unlabeled set and use them to improve the performance of our model. The way we divide the unlabeled samples into K parts based on the probability values of the discriminator's output is a novel point that enables us to balance the diversity and representativeness of the selected samples. With such a sample

selection method, we select samples that take into account the representativeness and uncertainty of the samples, while improving the sample redundancy problem.

2.4. Dimension reduction via t_SNE to visualize

To better visualize the generated results and the subsequent selection results, we use the t_SNE method to reduce the dimensionality of the features again. The t-SNE (t-distributed stochastic neighbor embedding) is an unsupervised non-linear dimension reduction technique which was introduced by [27]. This dimensionality reduction technique is a variant of the random neighborhood embedding introduced by [28], whose main purpose is to construct probability distributions from pairwise distances such that larger distances correspond to smaller probabilities and vice versa. t-SNE is currently a robust technique for visualization quality in dimension reduction [29].

3. Results

3.1. Performance of multi-classification task for Kather's dataset

One of our experiments to test multi-classification performance starts with an initial pool of labels in which 10% of the training set is labeled. The budget size of each batch is equal to 5% of the training dataset. The other experiment has 5% of the training set labeled. The budget size of each batch is equal to 2.5% of the training dataset. The pool of unlabeled data contains the rest of the training set from which samples are selected to be annotated by experts. Once labeled, they will be added into the initial training set and training is repeated on the new training set. To reduce the variance from the stochastic training process of neural networks, we average our results over 10 experiments for every method.

We chose random, entropy, and the DAL [26] methods as baselines for comparison. Random and entropy methods are described in Section 2.2, followed by a detailed description of the DAL method. The intuitive motivation for the DAL method is that if we can say with high probability that an unlabeled example came from the unlabeled set, then it is different from our current labeled examples and labeling it should be informative. The DAL method poses active learning as a binary classification task, attempting to choose examples to label in such a way as to make the labeled set and the unlabeled pool indistinguishable.

Figure 4 shows the multi-classification performance of our framework compared to prior works on Kather's dataset. Figure 4(a) plots the results of experiments in which 10% of the initial training set is labeled and the budget size of each batch is equal to 5% of the training dataset. It shows that our strategy is the first to exceed the average accuracy of 90%, getting 90.6% by using 25% of the data, ahead of other methods, whereas using the entire dataset yields accuracy 93.3%. Figure 5 shows the confusion matrices obtained after applying 25% and 100% of the labeled dataset to train the task model. The accuracy of 90% is set as the target performance to be achieved by the model on Kather's dataset. Figure 5 shows that the model trained with 25% of the labeled dataset achieves the target performance and is very close to the final version of the model with 100% of the labeled dataset. Comparing the mean accuracy values for data ratios above 15% shows that our method evidently outperforms random sampling, entropy and DAL method. The maximum achievable mean accuracy is 93.3% on Kather's dataset using 100% of the data, while ALHS achieves 92.4%, by only using 60% of it.
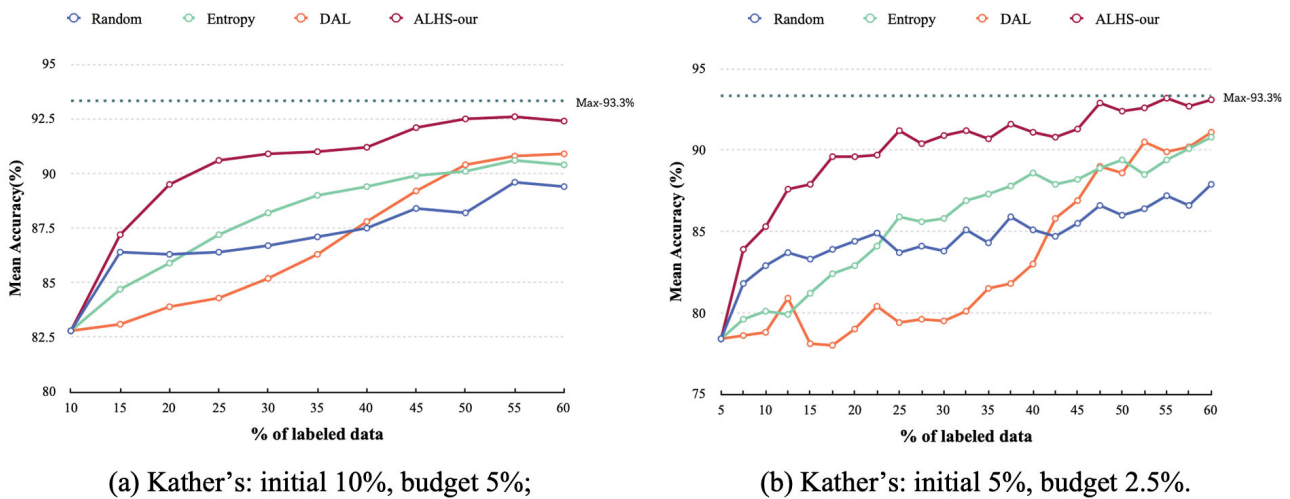
(a) Kather's: initial 10%, budget 5%; (b) Kather's: initial 5%, budget 2.5%.

Figure 4. Muti-classification results of Kather's dataset. Compared to other methods, our strategy is the first to exceed the average accuracy of 90% by using 25% of the data. ALHS achieves a mean accuracy of 92.4% by only using 60%, which is close to the bottleneck performance of 93.3%.
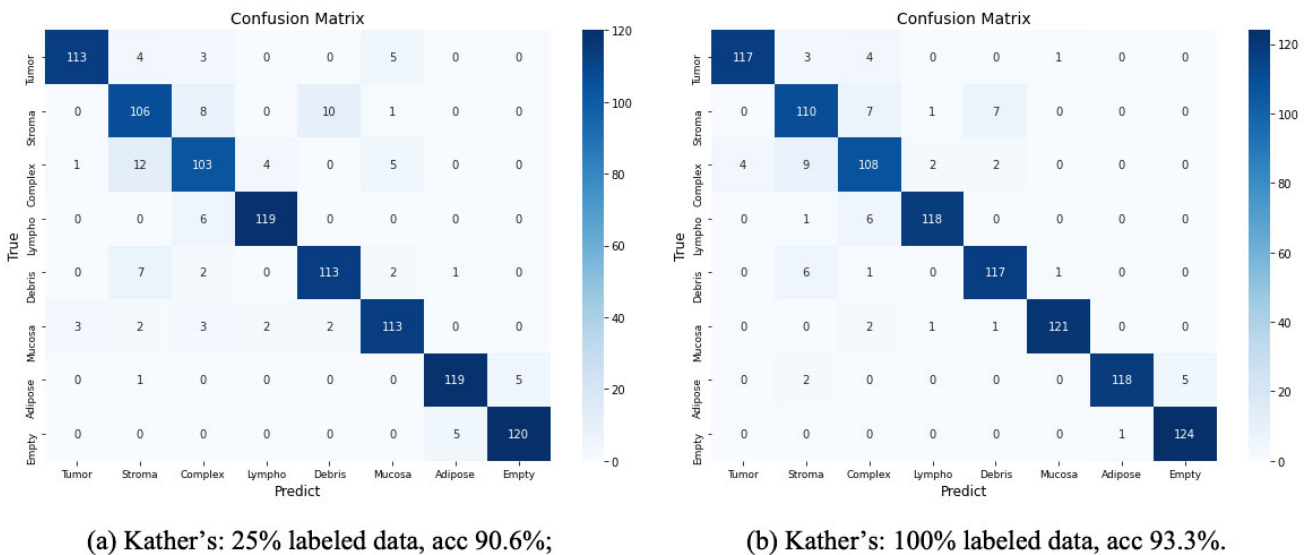

(a) Kather's: 25% labeled data, acc 90.6%; (b) Kather's: 100% labeled data, acc 93.3%.

Figure 5. The confusion matrices obtained after applying 25% and 100% of the labeled dataset to train the task model. The performance of the model trained with 25% of the labeled dataset is close to the final performance of the model with 100% of the labeled dataset.

Moreover, Figure 4(b) plots the results of experiments in which 5% of the initial training set is labeled and the budget size of each batch is equal to 2.5% of the training dataset. Our approach has more advantages over other methods. Similar to the previous experiment, our strategy is the first to exceed the average accuracy of 90%, getting 91.2% by using 25% of the data. Additionally, our method achieves 93.1% by using 60% of data, which is closer to the bottleneck performance of 93.3%.

Moreover, from the data ratios above 25% of the labeled data, entropy method outperforms random selection. The DAL method outperforms the random method after the ratios exceed 40%, but it rises quickly afterwards, outperforming the entropy method after the ratios exceed 50%. However, our approach performs much better.

3.2. Performance of binary classification task for BreakHis dataset

Figure 6 shows binary classification performance of our framework compared to prior works on BreakHis dataset. We also chose random, entropy, and the DAL methods as baselines for comparison.
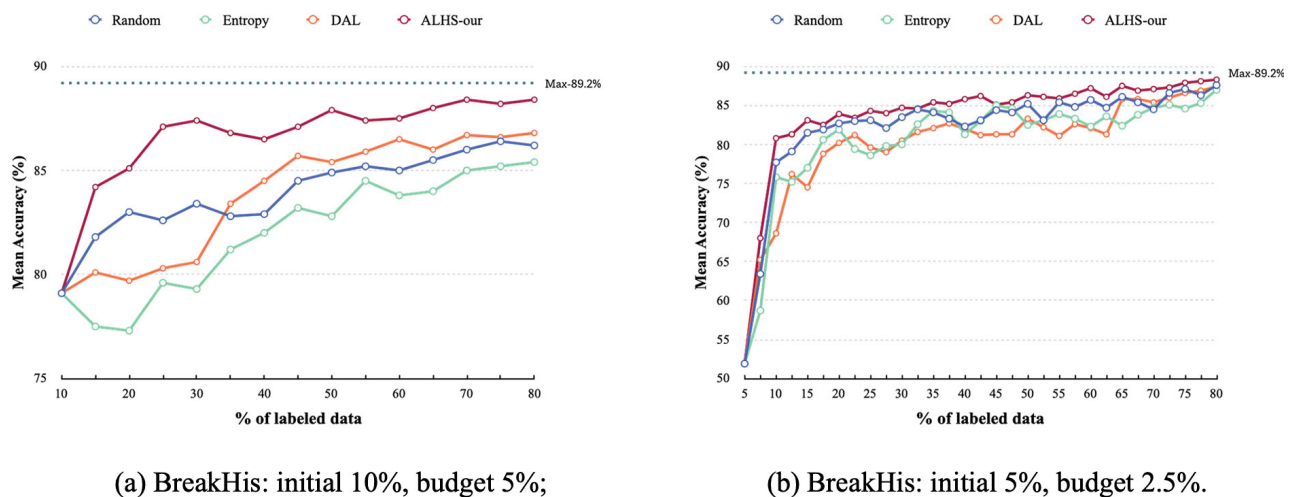

(a) BreakHis: initial 10%, budget 5%; (b) BreakHis: initial 5%, budget 2.5%.

Figure 6. Binary classification results of BreakHis dataset. Compared to other methods, our strategy is the first to reach the average accuracy of 88% by using 65% of the data, which is close to the bottleneck performance of 89.2%.

Figure 6(a) plots the results of experiments in which 10% of the initial training set is labeled and the budget size of each batch is equal to 5% of the training dataset. It shows that our strategy is the first to reach the average accuracy of 88% by using 65% of the data, ahead of other methods, whereas using the entire dataset yields accuracy 89.2%. Figure 7 shows the confusion matrices obtained after applying 65% and 100% of the labeled dataset to train the task model. The accuracy of 88% is set as the target performance to be achieved by the model on BreakHis dataset. Figure 7 shows that the model trained with 65% of the labeled dataset achieves the target performance and is very close to the final version of the model with 100% of the labeled dataset. From Figure 6(a), comparing the mean accuracy values for data ratios above 20% shows that our method evidently outperforms random sampling, entropy and DAL method. The maximum achievable mean accuracy is 89.2% on BreakHis dataset using 100% of the data while ALHS achieves 88.4% by using 80% of it. Moreover, Figure 6(b) plots the results of experiments in which 5% of the initial training set is labeled and the budget size of each batch is equal to 2.5% of the training dataset. Our approach shows some advantages over other methods. Our strategy is the first to exceed the average accuracy of 88%, getting 88.1% by using 75% of the data.
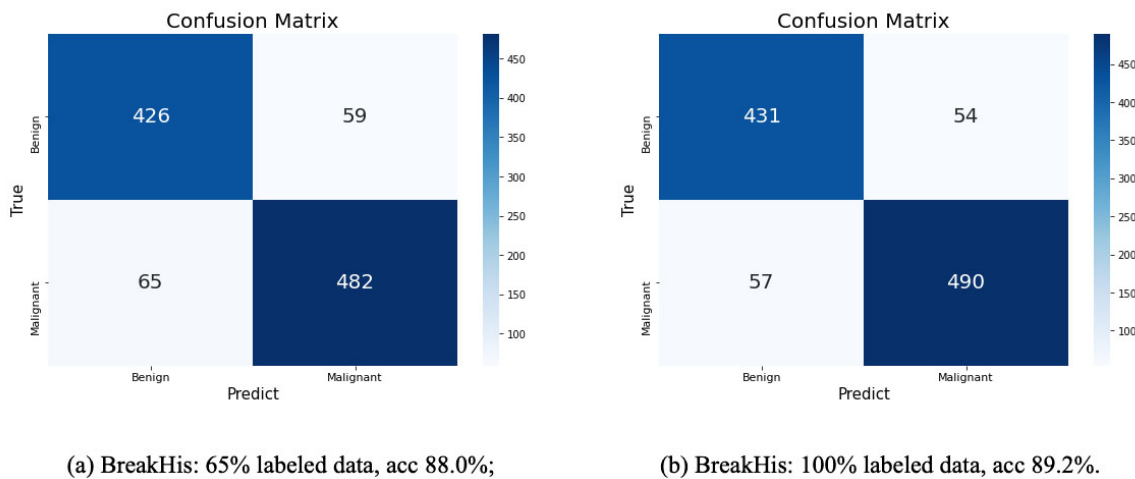
(a) BreakHis: 65% labeled data, acc 88.0%;   (b) BreakHis: 100% labeled data, acc 89.2%.

Figure 7. The confusion matrices obtained after applying 65% and 100% of the labeled dataset to train the task model. The performance of the model trained with 35% of the labeled dataset is close to the final performance of the model with 100% of the labeled dataset.

3.3. Visualization of latent representation and samples selection

To better visualize the latent representation of the auto-encoder model and the subsequent selection results, we use t_SNE to reduce the dimensionality to two and draw a two-dimensional scatter plot shown in Figure 8. Figure 8(a),(b) represent the results of latent spatial by t_SNE visualization for Kather's dataset and the BreakHis dataset, respectively. We can see from Figure 8 that the pathological images have some aggregation under the latent layer representation in the auto-encoder model. Although auto-encoder is a self-supervised model that can be trained without labels, the results of the representation of the feature space that it extracts are interpretable.
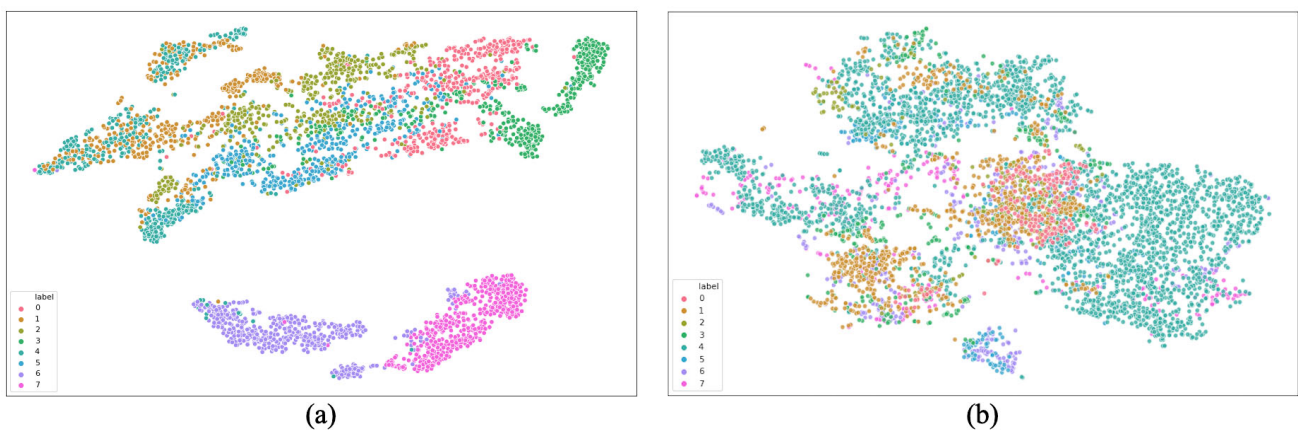

(a)   (b)

Figure 8. Visualization of latent representation in auto-encoder model. The pathological images have some aggregation of the same class of data under the latent representation in the AE, which indicates that the AE presents the latent representation of the images well.
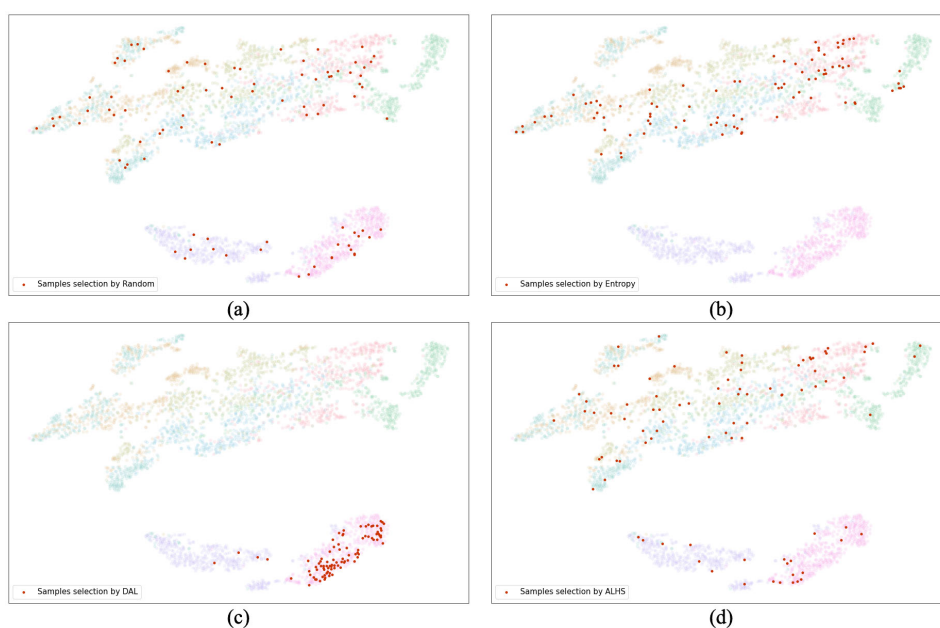
Figure 9. Results of samples selection on Kather's dataset at one time. (a)–(d) represent the sample selection results using random, entropy, DAL method and ALHS respectively. Our method selects more informative and diverse samples without the sample redundancy problem.
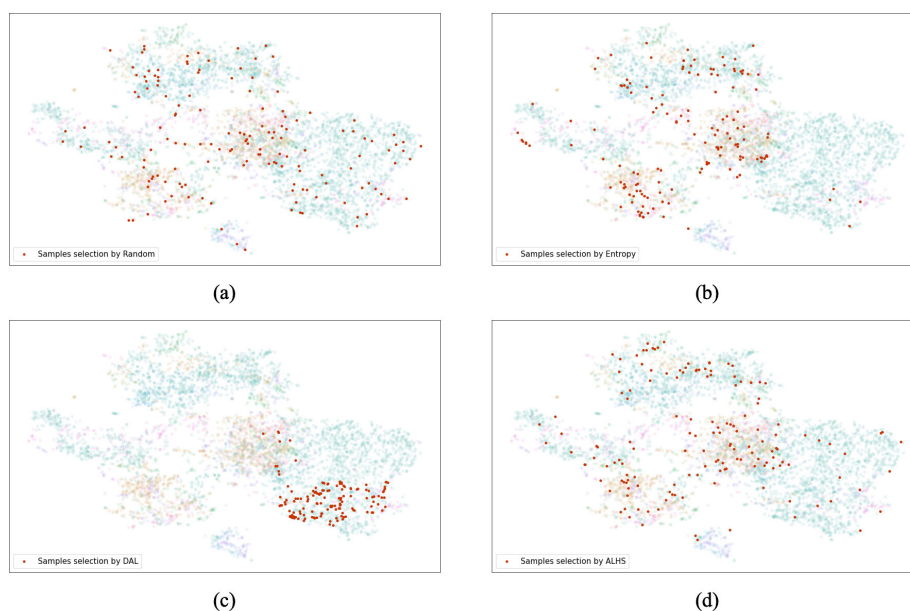

Figure 10. Results of samples selection on BreakHis dataset at one time. (a)–(d) represent the sample selection results using random, entropy, DAL method and ALHS respectively. Our method selects more informative and diverse samples without the sample redundancy problem.

Figures 9 and 10 display the samples selection results of applying the four methods on Kather's dataset and BreakHis dataset, respectively. From the selection results, the DAL method does not perform well. It picks almost all samples on the lower right side of Kather's and BreakHis dataset in one selection, which have obvious sample redundancy and lead to poor results in the final target model. The next is the entropy method, which selects the samples in the upper-middle on Kather's dataset and in the center-left on BreakHis dataset. The sample selection results of the random method are as expected. Our method, ALHS, performs excellently. It picks samples that cover almost every aspect of the feature map on Kather's dataset. On the BreakHis dataset, we find that it selects few sample points on the right side in this pick. Seeing that the right side is almost always blue sample points representing the same category, only a small number of samples need to be selected. This is exactly what our method does, allowing expensive annotations to be used on other instances that are more valuable. Moreover, our approach do not produce sample redundancy, which is one of the main reasons for the superior performance of our method on the task model.

Figure 11 shows the comparison of the reconstructed images after auto-encoder modeling with the original images on Kather's and BreakHis datasets. The difference between the original and generated images is not apparent, at least not distinguishable by the human eye, showing that the auto-encoder network we used performs the image reconstruction task well.
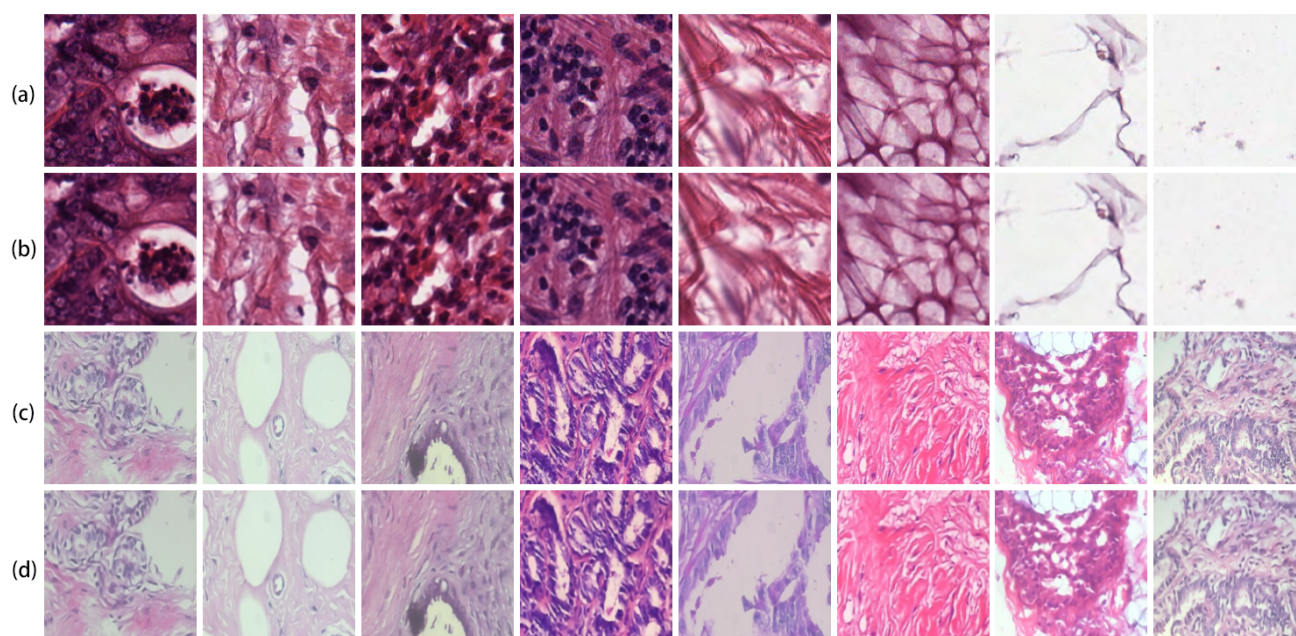

Figure 11. Comparison of the reconstructed images after auto-encoder modeling with the original images on Kather's and BreakHis datasets. (a) Original images of Kather's dataset; (b) Reconstructed images of Kather's dataset. (c) Original images of BreakHis dataset; (d) Reconstructed images of BreakHis dataset. There is little difference between the original and reconstructed images.

3.4. Ablation study

Figure 12 presents our ablation study to inspect the contribution of the key modules in our framework, including auto-encoder and the ALHS strategy. We perform ablation on the multi-classification task, which is more challenging than the binary classification task, and we perform an ablation experiment with Kather's dataset.
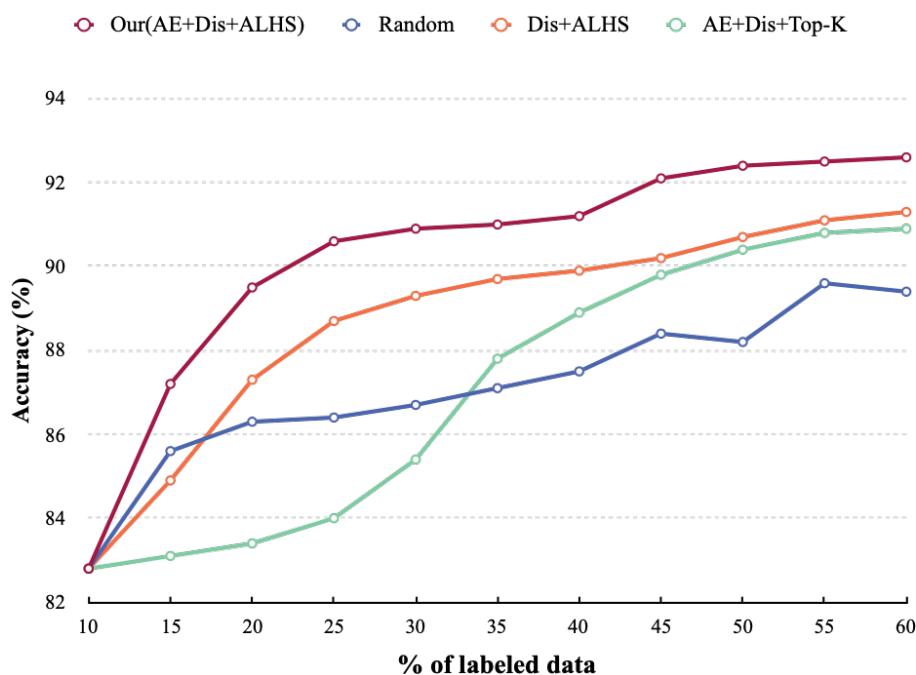

Figure 12. Ablation results on analyzing the effect of the Auto-Encoder and the discriminator denoted as Dis here. Random: just select samples randomly. Dis+ALHS: remove Auto-Encoder, just use discriminator and the ALHS strategy. AE+Dis+Top-K: use Auto-Encoder and discriminator, and replace the ALHS strategy with top-K strategy. Our approach outperforms other cases because we learn latent representations to train discriminative models and select samples with a novel ALHS strategy.

The variants of ablation we consider are as follows: 1) removing auto-encoder, just using discriminator and the ALHS strategy, and 2) removing the ALHS strategy and replacing it with top-K. In the first ablation, we only use a discriminator to memorize the data, which yields better performance than random; however, this performs worse that our method. Moreover, it reveals the key role of the auto-encoder is not only learning a rich latent representation, but also downscaling the data so that the discriminator can better learn the data representations. In the second ablation scenario, we use an auto-encoder to the previous setting to learn a lower dimensional space for training the discriminator. However, instead of using the ALSH strategy to select samples, we choose the top-K strategy, as shown in Eq (2.6). This ablation performed even less well than random at first, accumulating 35% of the labeled sample before surpassing the random method. However, our approach seems better than all these cases because we learn latent representations to train discriminative models and select samples with a novel ALHS strategy.

4. Discussion

Active learning, which is a branch of deep learning, holds great promise in the medical field. Supervised deep learning techniques traditionally relied on a large number of uniformly distributed, accurately annotated sample points [30]. Although an increasing number of medical image datasets are available, the time, cost, and effort required to annotate medical image datasets remain significant. To date, even well-resourced large national consortia have been challenged by the task of acquiring enough expert-validated labeled data [28]. Active learning methods are precisely proposed to alleviate the shortage of annotated labeled samples.

In this study, we propose a novel active learning framework applied to the pathology image dataset to reduce the heavy burden on pathologists. We use various feature extraction and dimensionality reduction methods to improve the efficiency and visualize the selection results. We also use the ALHS method to select samples to alleviate the redundancy problem. Experimental results demonstrate the effectiveness and robustness of our approach. Savings of 75% and 55% of the sample labeling budget were achieved while reaching the target performance in the two datasets, respectively.

There are three significant advantages to our study. Above all, our ALHS sample selection strategy is proposed explicitly for the sample redundancy problem. The sample redundancy phenomenon is reflected in the DAL method shown in Figures 9(c) and 10(c). We discover that the samples selected by the DAL method are concentrated in a particular region in one of the sample selection result plots, leading to a lack of diversity in the selected samples and a wasted annotation. We also find that under good mapping, there is a certain aggregation between the close data, and the output of the close data points will be near after the discriminative network. Based on this, we propose the ALHS method for samples selection. It avoids repeatedly drawing samples from a region overly but selects multiple regions and draws only a small number of samples from them, which can effectively avoid the sample redundancy problem and improve the quality of selected samples.

Furthermore, we visualize the sample selection results to indicate more directly the advantages of our method over other approaches. Visualization is a powerful tool for identifying problems and finding causes in the data processing. The visualized images can visually and clearly show that the samples selected by our method are more informative and representative. From each visualization of the sample selection results, it is also possible to infer the reasons for the poor performance of some methods, such as DAL, due to severe sample redundancy, and entropy methods, where some selection results are uncertain but lack diversity.

In addition, class imbalance [31] and noisy samples [32] are common problems in medical image datasets. The two datasets we selected are very representative, one of which is a balanced dataset Kather's and the other is an unbalanced dataset BreakHis. The validity of our active learning framework is verified on both datasets, showing that our approach works equally well for both balanced and unbalanced data. Moreover, in Kather's dataset, there is a class of 'empty' pathology images, approximating noisy samples as the images contain very little information and are almost substantially blank. The DAL method first picks the samples in the 'empty' class, indicating that the class differs from the labeled set. In contrast to the DAL method, which is strongly influenced by noisy samples, our method is less affected by noise and selects informative and diverse samples.

Despite our study's contribution in applying active learning methods to pathology image analysis, there are some potential shortcomings. First, we choose the AE network to accomplish the image

reconstruction task and extract its latent representations. However, the current more advanced image generation methods are VAE (Variational Auto-Encoders) [33] and GAN (Generative Adversarial Nets) [34], which take many times longer to train the network until fitting. Additionally, GAN has many variants [35–39] to adapt to different situations. With sufficient computational capacity, we can choose a more efficient network to find the latent representation of pathological images.

Second, while we use an active learning framework to reduce the need for labeled samples, there are state-of-the-art semi-supervised methods such as deep co-training [40], meta pseudo labels [41] and EnAEF [42] in machine learning. One of the advantages of semi-supervised learning over active learning as another approach to alleviating the shortage of labeled samples is that it can leverage the information of unlabeled samples without the additional cost of manual labeling. Improved results with a smaller budget could be obtained by combining our active learning framework with the advanced semi-supervised prerequisite methods.

Third, our active learning selection strategy picks informative and diverse samples. But the sets of samples to be labeled are independent between each selection, which may lead to close points in each selected sample set and result in wasted labeling. The current selection can be considered by combining the results of previous sample selections with the conditional probability method. Lastly, our selected pathology image datasets are not large enough to compare with deep learning datasets, such as MNIST [43], CIFAR [44], and ImageNet [45]. However, we can still verify the effectiveness of our framework.

Last but not least, the classification of images under the study may be affected by multiple uncertainties as well as inaccuracies. In this study, we choose Resnet50 and Inception v3, powerful convolutional neural networks, for image classification. We can also choose a fuzzy logic-based classifier [46, 47] to improve classification effect.

In conclusion, our proposed novel active learning framework in pathology image datasets can effectively alleviate the dilemma of insufficient annotated samples. Moreover, we hope our approach will soon reach real applications and practically relieve the pressure on pathologists, thus enabling cancer patients to receive timely pathology reports and proceed to diagnosis and treatment.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: Cancer J. Clin.*, **71** (2021), 209–249. https://doi.org/10.3322/caac.21660

2. J. Ferlay, M. Colombet, I. Soerjomataram, D. M. Parkin, M. Piñeros, A. Znaor, et al., Cancer statistics for the year 2020: An overview, *Int. J. Cancer*, **149** (2021), 778–789. https://doi.org/10.1002/ijc.33588

3. B. Acs, M. Rantalainen, J. Hartman, Artificial intelligence as the next step towards precision pathology, *J. Int. Med.*, **288** (2020), 62–81. https://doi.org/10.1111/joim.13030

4. E. J. Topol, High-performance medicine: The convergence of human and artificial intelligence, *Nat. Med.*, **25** (2019), 44–56. https://doi.org/10.1038/s41591-018-0300-7

5. D. M. Metter, T. J. Colgan, S. T. Leung, C. F. Timmons, J. Y. Park, Trends in the us and canadian pathologist workforces from 2007 to 2017, *JAMA Netw. Open*, **2** (2019), e194337. https://doi.org/10.1001/jamanetworkopen.2019.4337

6. Y. Song, R. Xin, P. Chen, R. Zhang, J. Chen, Z. Zhao, Identifying performance anomalies in fluctuating cloud environments: A robust correlative-gnn-based explainable approach, *Future Gener. Comput. Syst.*, **145** (2023), 77–86.

7. T. Xie, X. Cheng, X. Wang, M. Liu, J. Deng, T. Zhou, et al., Cut-thumbnail: A novel data augmentation for convolutional neural network, in *Proceedings of the 29th ACM International Conference on Multimedia*, (2021), 1627–1635.

8. H. Liu, P. Chen, X. Ouyang, G. Hui, Y. Bing, P. Grosso, et al., Robustness challenges in reinforcement learning based time-critical cloud resource scheduling: A meta-learning based solution, *Future Gener. Comput. Syst.*, **146** (2023), 18–33. https://doi.org/10.1016/j.future.2023.03.029

9. H. Lu, X. Cheng, W. Xia, P. Deng, M. Liu, T. Xie, et al., Cyclicshift: A data augmentation method for enriching data patterns, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 4921–4929.

10. P. Chen, H. Liu, R. Xin, T. Carval, J. Zhao, Y. Xia, et al., Effectively detecting operational anomalies in large-scale IoT data infrastructures by using a GAN-based predictive model, *Comput. J.*, **65** (2022), 2909–2925.

11. C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, *Electron. Mark.*, **31** (2021), 685–695. https://doi.org/10.1007/s12525-021-00475-2

12. A. L. Yuille, C. Liu, Deep nets: What have they ever done for vision, *Int. J. Comput. Vision*, **129** (2021), 781–802. https://doi.org/10.1007/s11263-020-01405-z

13. Z. H. Zhou, A brief introduction to weakly supervised learning, *Natl. Sci. Rev.*, **5** (2017), 44–53, https://doi.org/10.1093/nsr/nwx106

14. O. Sener, S. Savarese, Active learning for convolutional neural networks: A core-set approach, *arXiv preprint*, (2017), arXiv:1708.00489. https://doi.org/10.48550/arXiv.1708.00489

15. N. Houlsby, F. Huszár, Z. Ghahramani, M. Lengyel, Bayesian active learning for classification and preference learning, *arXiv preprint*, (2011), arXiv:1112.5745. https://doi.org/10.48550/arXiv.1112.5745

16. S. Sinha, S. Ebrahimi, T. Darrell, Variational adversarial active learning, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, (2019), 5971–5980. https://doi.org/10.1109/ICCV.2019.00607

17. A. Halder, A. Kumar, Active learning using rough fuzzy classifier for cancer prediction from microarray gene expression data, *J. Biomed. Inf.*, **92** (2019), 103136. https://doi.org/10.1016/j.jbi.2019.103136

18. D. Mahapatra, B. Bozorgtabar, J. P. Thiran, M. Reyes, Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer International Publishing, (2018), 580–588. https://doi.org/10.3917/perri.berli.2018.01.0580

19. A. L. Meirelles, T. Kurc, J. Saltz, G. Teodoro, Effective active learning in digital pathology: A case study in tumor infiltrating lymphocytes, *Comput. Methods Programs Biomed.*, **220** (2022), 106828.

20. A. Culotta, A. McCallum, Reducing labeling effort for structured prediction tasks, in *AAAI*, **5** (2005), 746–751.

21. T. Scheffer, C. Decomain, S. Wrobel, Active hidden markov models for information extraction, in *International Symposium on Intelligent Data Analysis (IDA)*, Springer, Cascais, Portugal, (2001), 309–318.

22. C. E. Shannon, A mathematical theory of communication, *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, **5** (2001), 3–55. https://doi.org/10.1145/584091.584093

23. J. N. Kather, C. A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, et al., Multi-class texture analysis in colorectal cancer histology, *Sci. Rep.*, **6** (2016), 1–11. https://doi.org/10.1038/srep27988

24. F. A. Spanhol, L. S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, *IEEE Trans. Biomed. Eng.*, **63** (2015), 1455–1462. https://doi.org/10.1109/TBME.2015.2496264

25. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, USA, (2016), 770–778.

26. D. Gissin, S. Shalev-Shwartz, Discriminative active learning, *arXiv preprint*, (2019), arXiv:1907.06347. https://doi.org/10.48550/arXiv.1907.06347

27. L. Van der Maaten, G. Hinton, Visualizing data using t-sne, *J. Mach. Learn. Res.*, **9** (2008), 2579–2605.

28. T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, et al., Opportunities and obstacles for deep learning in biology and medicine, *J. R. Soc. Interface*, **15** (2018), 20170387.

29. S. Nanga, A. T. Bawah, B. A. Acquaye, M. I. Billa, F. D. Baeta, N. A. Odai, et al., Review of dimension reduction methods, *J. Data Anal. Inf. Process.*, **9** (2021), 189–231. https://doi.org/10.4236/jdaip.2021.93013

30. A. L'Heureux, K. Grolinger, H. F. Elyamany, M. A. M. Capretz, Machine learning with big data: Challenges and approaches, *IEEE Access*, **5** (2017), 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

31. A. Bria, C. Marrocco, F. Tortorella, Addressing class imbalance in deep learning for small lesion detection on medical images, *Comput. Biol. Med.*, **120** (2020), 103735. https://doi.org/10.1016/j.compbiomed.2020.103735

32. M. Outtas, *Compression Oriented Enhancement of Noisy Images: Application to Ultrasound Images*, USTHB-Alger, 2019.

33. C. Doersch, Tutorial on variational autoencoders, *arXiv preprint*, (2016), arXiv:1606.05908. https://doi.org/10.48550/arXiv.1606.05908

34. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial networks, *Commun. ACM*, **63** (2020), 139–144. https://doi.org/10.1145/3422622

35. M. Mirza, S. Osindero, Conditional generative adversarial nets, *arXiv preprint*, (2014), arXiv:1411.1784. https://doi.org/10.48550/arXiv.1411.1784

36. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, in *Advances in Neural Information Processing Systems*, (2017), 5769–5779.

37. J. Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, Italy, (2017), 2242–2251.

38. A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, *arXiv preprint*, (2018), arXiv:1809.11096. https://doi.org/10.48550/arXiv.1809.11096

39. J. Zhao, M. Mathieu, Y. LeCun, Energy-based generative adversarial network, *arXiv preprint*, (2016), arXiv:1609.03126. https://doi.org/10.48550/arXiv.1609.03126

40. S. Qiao, W. Shen, Z. Zhang, B. Wang, A. Yuille, Deep co-training for semi-supervised image recognition, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, (2018), 142–159. https://doi.org/10.1787/qna-v2018-2-12-en

41. H. Pham, Z. Dai, Q. Xie, Q. V. Le, Meta pseudo labels, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, (2021), 11557–11568.

42. X. Wang, D. Kihara, J. Luo, G. J. Qi, Enaet: A self-trained framework for semi-supervised and supervised learning with ensemble transformations, *IEEE Trans. Image Process.*, **30** (2021), 1639–1647. https://doi.org/10.1109/TIP.2020.3044220

43. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE*, **86** (1998), 2278–2324. https://doi.org/10.1109/5.726791

44. A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, **2009** (2009), 1–58.

45. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos, USA, (2009), 248–255.

46. M. Versaci, G. Angiulli, P. Crucitti, D. De Carlo, F. Laganà, D. Pellicanò, et al., A fuzzy similarity-based approach to classify numerically simulated and experimentally detected carbon fiber-reinforced polymer plate defects, *Sensors*, **22**, (2022), 4232. https://doi.org/10.3390/s22114232

47. A. T. Azar, A. E. Hassanien, Dimensionality reduction of medical big data using neural-fuzzy classifier, *Soft comput.*, **19** (2015), 1115–1127. https://doi.org/10.1007/s00500-014-1327-4

48. N. Lei, Y. Guo, D. An, X. Qi, Z. Luo, S. T. Yau, et al., Mode collapse and regularity of optimal transportation maps, *arXiv preprint*, (2019), arXiv:1902.02934. https://doi.org/10.48550/arXiv.1902.02934

49. M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, in *International Conference on Learning Representations(ICLR)*, Toulon, France, (2017), 1–17.

Appendix

Latent representation for discrimination

The discriminator of our active learning framework is trained to distinguish latent representations from labeled and unlabeled samples, and the AE network is trained to observe the latent representations. This means that the encoder network in AE can be considered a generator of the discriminator network. Following [48, 49], we take a theoretical step forward in understanding the training dynamics of this procedure.

According to previous work [48], training the AE model is equivalent to computing the encoding map f_θ and decoding map g_ξ:

$$(\nu_{gt}, \chi) \xrightarrow{f_\theta} (\mu_{gt}, \Omega) \xrightarrow{g_\xi} (\nu_{gt}, \chi) \tag{A1}$$

where f_θ and g_ξ are parameterized by CNNs, and ν_{gt} is a probability measure modeled from the data distribution, which is supported on a k-dimensional manifold χ embedded in a Euclidean space with N dimensions. Given a dense sampling from the image manifold and ideal optimization, $f_\theta \emptyset g_\xi$ coincides with the identity map. Then, f_θ will be continuous and convertible, namely homeomorphism, and g_ξ will be the inverse homeomorphism during the training. That means, $f_\theta : \chi \to \Omega$ is an embedding, and pushes forward ν_{gt} to the latent data distribution $\mu_{gt} := f_\theta \# g_\xi$.

Obviously, since $k << N$, the latent representations in μ_{gt} is a lower-dimensional representation relative to the data distribution but can indicate ν_{gt} over the full dimensions. Therefore, we can say that mapping the data distribution into the latent space can effectively improve the performance with less computational cost.

Then we recall the theory of the perfect discriminator according to [49].

Definition 1. *Let $\mathcal{M}$ and $\mathcal{P}$ be two boundary free regular submanifolds of $\mathbb{R}^d$. Let $x \in \mathcal{M} \cap \mathcal{P}$ be an intersection point of the two manifolds. $\mathcal{M}$ and $\mathcal{P}$ intersect transversally in x if $T_x \mathcal{M} + T_x \mathcal{P} = T_x \mathbb{R}^d$, where $T_x \mathcal{M}$ means the tangent space of $\mathcal{M}$ around x. Accordingly, $\mathcal{M}$ and $\mathcal{P}$ perfectly align if there is an $x \in \mathcal{M} \cap \mathcal{P}$ such that $\mathcal{M}$ and $\mathcal{P}$ don't intersect transversally in x.*

Lemma 1. *Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that don't perfectly align and don't have full dimension. Let $\mathcal{L} = \mathcal{M} \cap \mathcal{P}$. If $\mathcal{M}$ and $\mathcal{P}$ don't have boundary. then $\mathcal{L}$ is also a manifold, and has strictly lower dimension than both the one of $\mathcal{M}$ and the one of $\mathcal{P}$. If they have boundary, $\mathcal{L}$ is a union of at most 4 strictly lower dimensional manifolds. In both cases, $\mathcal{L}$ has measure 0 in both $\mathcal{M}$ and $\mathcal{P}$.*

Lemma 2. *Let $\mathbb{P}_r$(data distribution) and $\mathbb{P}_g$(generated distribution) be two distributions that have support contained in two closed manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't perfectly align and don't have full dimension. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds, meaning that if there is a set A with measure 0 in $\mathcal{M}$, then $\mathbb{P}_r(A) = 0$(and analogously for $\mathbb{P}_g$). Then, there exists and optimal discriminator $D^* : \mathcal{X} \to [0, 1]$ that has accuracy 1 and for almost any x in $\mathcal{M}$ or $\mathcal{P}$, D^* is smooth in a neighbourhood for x and $\nabla D^*(x) = 0$.*

As mentioned in Lemmas 1 and 2, we can assume that there is a perfect discriminator D that is smooth and constant almost everywhere in $\mathcal{M}$ and $\mathcal{P}$. At the same time, their supports are disjoint or lie on low-dimensional manifolds. In this case, the discriminator and generator updates will be stopped completely.

Let M_1 and M_2 be two regular submanifolds come from the space of the data distribution and the latent representation, respectively. Analogously for P_1 and P_2, by Lemma 1, $L_1 = M_1 \cap P_1$ and $L_2 = M_2 \cap P_2$ is strictly lower dimensional than their supports, respectively.

Let $x \in M_1 \setminus L_1$, we say that $x \in P^c$ (the complement of P) which is an open set that there exists a ball of radius ϵ_x such that $B(x, \epsilon_x) \cap P = \emptyset$. Let

$$\hat{\mathcal{M}}_\infty = \bigcup_{x \in \mathcal{M}_\infty \setminus \mathcal{L}_\infty} B(x, \varepsilon_x \setminus 3) \tag{A2}$$

Define $\hat{P}_1$ analogously. The discriminator is perfect while $\hat{M} \cap \hat{P} = \emptyset$. We can say that the discriminator used to observe the latent space is more difficult to perfect because of the high correlation between each x in the latent space. The ball in the latent space is smoother and relatively close, indicating less tendency to have the perfect discriminator problem, consequently improving classification performance.