

Uncertainty Awareness Enables Efficient Labeling for Cancer Subtyping in Digital Pathology

Nirhoshan Sivaroopan^{1,†} Chamuditha Jayanga Galappaththige¹ Chalani Ekanayake¹
Hasindri Watawana¹ Ranga Rodrigo¹ Chamira U. S. Edussooriya¹
Dushan N. Wadduwage^{2,3,‡,*}

¹University of Moratuwa ²Harvard University ³Old Dominion University

[†]180428t@uom.lk, [‡]dwadduwa@odu.edu

Abstract

Machine-learning-assisted cancer subtyping is a promising avenue in digital pathology. Cancer subtyping models however require careful training using expert annotations, so that they can be inferred with a degree of known certainty (or uncertainty). To this end, we introduce the concept of uncertainty awareness into a self-supervised contrastive learning model. This is achieved by computing an evidence vector at every epoch, which assesses the model's confidence in its predictions. The derived uncertainty score is then utilized as a metric to selectively label the most crucial images that require further annotation, thus iteratively refining the training process. With just 1-10% of strategically selected annotations, we attain state-of-the-art performance in cancer subtyping on benchmark datasets. Our method not only strategically guides the annotation process to minimize the need for extensive labeled datasets, but also improve the precision and efficiency of classifications. This development is particularly beneficial in settings where the availability of labeled data is limited, offering a promising direction for future research and application in digital pathology. Our code is available at <https://github.com/Nirhoshan/AI-for-histopathology>

1. Introduction

Integration of deep learning into computer-assisted digital pathology has revolutionized cancer diagnostics, offering a powerful tool to streamline the complex, labor-intensive, and error-prone processes associated with image-based detection. Despite these advancements, the field faces a significant challenge: the exhaustive and costly process of image annotation. Histopathological analysis demands pre-

cise labeling by expert pathologists, a procedure that is not only time-consuming but also heavily resource-dependent. Addressing this issue, the community initially turned to self-supervised learning (SSL) as a solution [7, 23], which, while effective in some respects, often lacked in providing explainable model predictions, a critical requirement for potential clinical use. These models, focused mainly on accuracy, frequently underperformed on datasets with limited domain similarity to the training data [26].

Another approach to mitigate the annotation burden is active learning (AL). AL introduces a human-in-the-loop querying strategy, offering a degree of interpretability and reduction of labeling effort [27]. Research such as that conducted by [2] investigated different querying strategies in AL, finding that random sampling frequently surpasses strategic label selection in patch-based machine learning. This observation led to the adoption of a method where multiple patches are combined for AL, albeit at the expense of higher computational requirements.

A key enhancement to AL lies in incorporating uncertainty into the querying process [17], as it directly influences the explainability and reliability [14] of model predictions.

Previous works, such as [17], have employed uncertainty metrics at the Whole Slide Image (WSI) level. However, these approaches often lacked interpretability, as they relied on indirect uncertainty parameters such as dropouts or model weights.

Our work synergizes SSL and AL, addressing the shortcomings of each method when used individually. We begin by evaluating various models across different labeling scenarios to identify the most suitable SSL framework. This led us to select SimCLRv2 [4] for its exceptional performance. We then enhance this framework by integrating a novel strategy of modeling uncertainty within the architecture itself [25], an approach not previously employed

*Corresponding author

in histopathology. This addition not only boosts the interpretability of the model predictions—a vital aspect in critical domains like histopathology—but also sets a new standard in the field.

Leveraging this enhanced SSL framework, we then apply the uncertainty score as a querying strategy in AL. This approach is benchmarked against traditional random sampling methods. Our combined SSL and AL framework excels in patch-level classification for binary and multi-class cancer types. It adds explainability to model predictions and significantly reduces annotation efforts. The results are compelling: our model achieves parity with state-of-the-art (SOTA) outcomes using only 2-3% of labels and surpasses them at the 9% label mark. The subsequent sections of this paper detail our process in achieving these results, from the selection and enhancement of the SSL framework to the application of uncertainty-aware querying in AL for patch level classification.

2. Related Work

2.1. Self-supervised Representation Learning

SSL has emerged as a powerful approach, especially for pre-training large models using unlabeled data [5, 12, 13, 24, 28, 33]. In the realm of digital pathology, SSL frameworks like SimCLR [3] and its enhanced version, SimCLRv2 [4], have shown promise by learning rich representations from relatively large amount of unlabeled data. SimCLRv2, in particular, improves upon its predecessor through larger backbone networks, an expanded projection head, and the application of knowledge distillation [16]. Another notable SSL method, Masked Auto-Encoding (MAE) [15], reconstructs images from partially masked inputs, demonstrating its utility in tasks like image classification. While these SSL methods, including adaptations for digital pathology like those by [7] and [23], have proven effective, SimCLRv2’s adoption in digital pathology remains unexplored. Moreover, the potential of these models, particularly in terms of the uncertainty in their predictions, has yet to be fully investigated.

2.2. Uncertainty Quantification

Estimating uncertainty in deep learning models is a crucial yet challenging aspect of machine learning, particularly in clinical applications like histopathology. Traditional methods, such as Monte Carlo dropout [11, 30], deep ensembles [22, 32], and test-time augmentation [9], generate variability in predictions have been used to estimate uncertainty. However, these approaches often rely on inherent ambiguities in model parameters, lacking precise mathematical quantification of uncertainty. Recognizing this limitation, our work enhances SSL framework with a Bayesian approach to uncertainty estimation, as proposed by [25].

This approach, grounded in the theory of evidence, excels in task-agnostic learning across different domains and it aligns perfectly with our objective of harnessing public datasets in digital pathology to acquire extensive pre-trained domain knowledge.

In our exploration of uncertainty, it is important to acknowledge the two primary types: aleatoric and epistemic uncertainty. Aleatoric uncertainty refers to the inherent noise in the data that cannot be reduced through model training such as image quality, while epistemic uncertainty is related to the model’s lack of knowledge, which can be mitigated through better training and data representation. In this work, we focus on reducing epistemic uncertainty by quantifying the aspects of uncertainty that are within our control through careful training strategies. By addressing epistemic uncertainty, we aim to enhance model reliability and improve performance in clinical applications.

2.3. Active Learning

AL is the machine learning method which actively queries the most informative labels to consistently improve the model training. AL is a well adapted method in histopathology to reduce annotation cost [17], [2]. In AL the network training will be initiated by labelling a limited number of randomly selected images. Then, the key problem in AL is how the querying strategy is defined to select the most valuable samples to gather the most information for model training. Researches conducted for AL were circling around the challenge of identifying the most effective querying strategy to find images that yield the highest entropy. In [2], the researchers compared random sampling with different querying strategies. In [17], authors turned into quantifying uncertainty to find the images with highest entropy. However, in [10], authors tried to improve the AL by incorporating both samples with high entropy values and low entropy values to emphasize the confidence boosting. Though all these work introduced different querying strategies to reduce the annotation cost, the accuracy was on par or slightly less [10] compared to that with random selection of images. This may be due to the querying strategy failing to select the most informative images for the next iteration. We leverage the uncertainty estimation method introduced in [25], that proved to be performing better in uncertainty quantification compared to other uncertainty estimating methods [25], to develop the querying strategy, which resulted in reduced annotation cost and significant improvement in accuracy compared to the random sampling of labels.

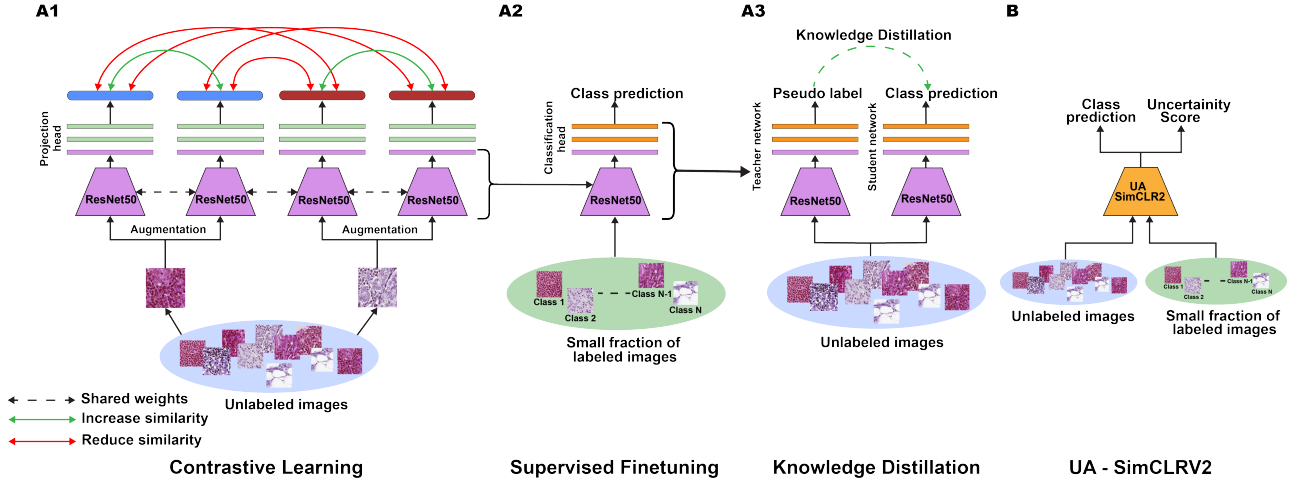


Figure 1. The SimCLRv2 framework comprises three steps: (A1) Pre-training employs contrastive learning on unlabelled images. (A2) Supervised fine-tuning adds a classification head to the pre-trained encoder and fine-tunes using labeled images. (A3) Knowledge distillation involves using the fine-tuned model as a teacher network to generate pseudo labels for unlabeled images, then training a student network. (B) The proposed UA-SimCLRv2 model extends this with an additional output for the uncertainty score, enhancing model prediction explainability.

3. Methodology

3.1. Datasets

In this work, we used two datasets, the Patch Camelyon (PCam) dataset [29], and the NCT-CRC-HE-100K (NCT100k) dataset [20]. Table 1 details the PCam dataset and Table 2 details the NCT100k dataset.

Table 1. Description of attributes and characteristics of the PCam dataset extracted from histopathology scans of lymph node sections from CAMELYON16.

Attribute	Description
Source	CAMELYON16 WSI
Patch Count	327,680
Patch Size	96x96 pixels (resized to 224x224)
Label	Binary indicating metastatic tissue presence
Data Splits	Training (75%), Validation (12.5%), Test (12.5%)

3.2. Patch Level Classification

We first utilized SimCLRv2 as a patch-level classifier (see the model architecture in Fig. 1). For both datasets, we benchmarked our model against other models that have been used previously for the same task in digital pathology. Our experimental framework assessed model performance based on two criteria: the proportion of training set annotations used for fine-tuning and the context of pre-training

Table 2. Description of attributes and characteristics of the NCT100k dataset, consisting of H&E stained histological images annotated into nine classes.

Attribute	Description
Source	Human colorectal cancer and normal tissues
Patch Count	100,000
Patch Size	224x224 pixels
Label	Adipose (ADI), Background (BACK), Debris (DEB), Lymphocytes (LYM), Mucus (MUC), Smooth Muscle (MUS), Normal Colon Mucosa (NORM), Cancer-Associated Stroma (STR), Colorectal Adenocarcinoma Epithelium (TUM)
Data Splits	Training (NCT100k), Validation (CRC-VAL-HE-7K (CRC7k))

data (in-domain or out-domain). In-domain refers to using the same dataset for pre-training (contrastive learning), fine-tuning, and knowledge distillation, followed by testing the trained model on the same dataset. Out-domain refers to using a different dataset for the pre-training step. For instance, in the PCam-outdomain setting the NCT100K dataset was used for pre-training.

3.3. Evaluation Metrics

To evaluate the performance of our classification models, we used three primary metrics: Accuracy, F1 Score, and Area Under the ROC Curve (AUC).

- **Accuracy:** Measures the ratio of correctly classified instances to the total number of instances in the dataset, offering a basic measure of overall model performance. However, accuracy may be less informative when there is a class imbalance.
- **F1 Score:** This balances precision and recall, making it especially useful for evaluating model performance in cases of class imbalance. In this work, we employed the *weighted-average* F1 score, which calculates the F1 score for each class individually and then averages these scores according to the proportion of each class in the dataset. This approach ensures that classes with more instances contribute proportionally to the final score, providing a more balanced evaluation across all classes.
- **Area Under the ROC Curve (AUC):** Measures the area under the ROC curve, which plots the true positive rate against the false positive rate across different threshold values. The AUC score indicates how well the model can distinguish between classes, with values closer to 1.0 representing stronger performance in correctly classifying positive and negative instances. A score of 0.5 implies random performance, while a higher AUC reflects better discriminative ability.

These metrics together provide a comprehensive view of the model’s accuracy and its robustness in handling class imbalances.

3.4. UA SimCLRv2

We next introduced uncertainty awareness [25] to the SimCLRv2 framework. Our uncertainty aware SimCLRv2 is termed UA-SimCLRv2. The primary objective of UA-SimCLRv2 is to enhance the interpretability of the model’s predictions in the context of digital pathology. This is achieved by incorporating the theory of uncertainty estimation, which serves as the basis for uncertainty awareness in UA SimCLRv2.

In [25], the uncertainty estimation is approached from Dempster–Shafer theory of evidence (DST) perspective [8] assigning belief masses to subsets of a frame of discernment, which denotes the set of exclusive possible states. Subjective logic formalizes DST’s notion of belief assignments over a frame of discernment as a Dirichlet distribution. Term evidence is a measure of the amount of support collected from data in favor of a sample to be classified into a certain class. Through model training evidence

e_k ($k = 1, 2, \dots, K$) are collected and belief masses b_k ($k = 1, 2, \dots, K$) are assigned to each class based on the evidence collected and the remaining are marked as uncertainty u . For K mutually exclusive classes,

$$u + \sum_{k=1}^K b_k = 1. \quad (1)$$

Here $u \geq 0$ and $b_k \geq 0$, and they are calculated by,

$$b_k = \frac{e_k}{S} \quad \text{and,} \quad u = \frac{K}{S}, \quad \text{where} \quad S = \sum_{i=1}^K e_i + 1. \quad (2)$$

Observe that when there is no evidence, the belief for each class is zero and the uncertainty is one. A belief mass assignment, i.e., subjective opinion, corresponds to a Dirichlet distribution with parameters $\alpha_k = e_k + 1$. A Dirichlet distribution parameterized over evidence represents the density of each such probability assignment; hence it models second-order probabilities and uncertainty [19]. It is characterized by K parameters $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_K]$ and is given as

$$D(p|\alpha) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^K p_i^{\alpha_i-1}, & \text{if } p \in S_K \\ 0, & \text{otherwise,} \end{cases}$$

where $S_K = \{p | \sum_{i=1}^K p_i = 1 \text{ and } 0 \leq p_1, \dots, p_K \leq 1\}$ and $B(\alpha)$ is the K -dimensional multinomial beta function [21].

Model training follows the classical neural network architecture with a softmax layer replaced with ReLU activation layer to ascertain non-negative output, which is taken as the evidence vector for the predicted Dirichlet distribution. For network parameters θ , let $f(x_i|\theta)$ be the evidence vector predicted by the network for the classification. Corresponding Dirichlet distribution’s parameters $\alpha_i = f(x_i|\theta) + 1$ are calculated and their means ($\frac{\alpha_i}{S}$) are considered as the class probabilities. Let y_i be one hot vector encoding the ground-truth class label of a sample x_i . Treating $D(p_i|\alpha_i)$ as a prior on the sum of squares loss $\|y_i - p_i\|_2^2$, we obtain the loss function

$$L_i(\theta) = \int \|y_i - p_i\|_2^2 \frac{1}{B(\alpha_i)} \prod_{i=1}^K p_{ij}^{\alpha_{ij}-1} dp_i. \quad (3)$$

By decomposing the first and second moments, minimization of both the prediction error and the variance of the Dirichlet experiment for each sample is achieved by the above loss function. Further some evidence collected might strengthen the belief for multiple classes. To avoid situations where evidence with more ambiguity assigns more belief to incorrect class, Kullback-Leibler (KL) divergence

term is appended to the loss function. Following is the total loss used for UA fine-tuning.

$$L(\theta) = \sum_{i=1}^N L_i(\theta) + \lambda_t \sum_{i=1}^N KL[D(p_i || \tilde{\alpha}_i) || D(p_i || <1, \dots, 1 >)] \quad (4)$$

where $\lambda_t = \min(1, t/10) \in [0, 1]$ is the annealing coefficient, t is the index of the current training epoch, $D(p_i || <1, \dots, 1 >)$ is the uniform Dirichlet distribution, and $\tilde{\alpha}_i = y_i + (1 - y_i) * \alpha_i$ is the Dirichlet parameters after removal of the non-misleading evidence from predicted parameters α_i for sample i . The KL divergence term in the loss can be calculated as

$$\begin{aligned} & KL[D(p_i || \tilde{\alpha}_i) || D(p_i || 1)] \\ &= \log \left(\frac{\Gamma(\sum_{k=1}^K \tilde{\alpha}_{ik})}{\Gamma(K) \prod_{k=1}^K \Gamma(\tilde{\alpha}_{ik})} \right) \\ &+ \sum_{k=1}^K (\tilde{\alpha}_{ik} - 1) \left[\psi(\tilde{\alpha}_{ik}) - \psi \left(\sum_{j=1}^K \tilde{\alpha}_{ij} \right) \right] \end{aligned}$$

where 1 represents the parameter vector of K ones, $\Gamma(\cdot)$ is the gamma function, and $\psi(\cdot)$ is the digamma function. By gradually increasing the effect of the KL divergence in the loss through the annealing coefficient, the neural network is allowed to explore the parameter space and avoid premature convergence to the uniform distribution for the misclassified samples, which may be correctly classified in future epochs.

3.5. Uncertainty-aware Active learning (UA-AL)

Last, we leveraged uncertainty scores in UA-SimCLRv2, as the querying strategy for AL. We demonstrated our AL for subtyping of NCT100k dataset. Starting with a pre-trained model, we first labeled 1% of images randomly and fine-tuned the model. Subsequently, we iteratively queried the top 1% uncertain images and added them to the training set with expert-annotated labels. This process continued until 10% of the labels were used in training. As in Fig. 2.A, We compared UA-AL with both regular SimCLRv2 and UA-SimCLRv2 models with random sampling for labeling at each iteration, assessing model performance on the test dataset. As discussed in section 2.3, because the related work showed comparable or slightly lower accuracy than the random sampling strategy, we conducted the comparison directly against the random sampling method.

4. Results

In Section 4.1, we evaluate various SSL frameworks for patch-level classification to identify the most effective one.

We then incorporate uncertainty awareness into the chosen SSL framework. Then, we visualize and analyze the suitability of the selected uncertainty estimation method as a querying strategy in AL. Finally, in Section 4.2, we discuss the effectiveness of uncertainty awareness in AL and observe that we achieve SOTA results in patch-level classification using only 2% of in-domain labels.

4.1. Patch level Classification

4.1.1 Binary Class Classification

Table 3 show the accuracy, F1 score, and AUC score for the PCam binary classification. To establish a baseline, we first fine-tuned our models with all training labels (i.e., the 100% setting). Here, our models outperformed the SOTA approach, i.e., MAE [15]. In-domain pre-trained SimCLRv2, performed best with 2.16% increase in accuracy compared to the SOTA and UA-SimCLRv2 performed even better. Next, we fine-tuned our models on 10% training labels. 10%-fine-tuned models performed slightly worse than the 100% baseline. Nevertheless, the 10%-fine-tuned SimCLRv2 and UA-SimCLRv2 still performed on par with or better than the SOTA. Then, we fine-tuned our models on 1% training labels. Interestingly the SimCLRv2 and UA-SimCLRv2 models still performed comparable to the SOTA (see the 1% setting on Table 3). However, at the 1% setting UA-SimCLRv2 consistently underperformed compared to SimCLRv2, perhaps due to the limited evidence available for uncertainty awareness. Rows in bold highlight the best results within their respective sections.

4.1.2 Multi-Class Classification

Table 4 shows multi-class classification results for the NCT100k dataset. Similar to the binary case, we experimented at 100%, 10% and 1% fine-tuning settings. First, at the 100% setting our SimCLRv2 and UA-SimCLRv2 performed on par with the SOTA. Interestingly, out-domain pre-trained SimCLRv2 was the best-performing model and surpassed the SOTA by a small margin. But at the 1% setting, we observed a degradation of performance by a few percentage points. We would further explore the impact of in-domain and out-domain setting in future work.

Tables 3 and 4 demonstrate that SimCLRv2 stands out as the superior SSL framework when compared to current SOTA models. Furthermore, the incorporation of uncertainty awareness into SimCLRv2 not only maintains its high accuracy but also enhances the interpretability of its predictions.

4.1.3 Visualizing Uncertainty estimation

Our t-SNE analysis in Fig. 3 initially showed that compared to SimCLRv2, UA-SimCLRv2’s T-SNE maps demonstrate

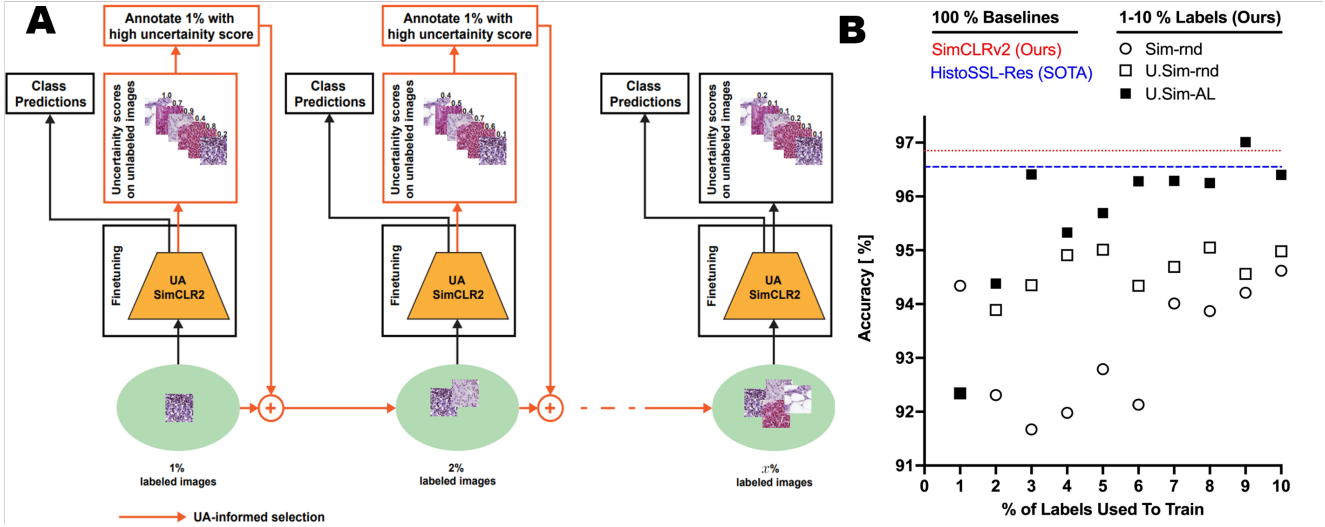


Figure 2. (A) UA-AL framework (B) Outdomain training of UA-AL outperforms random label selection, achieving comparable results to the SOTA with just 3% of labels, and surpassing the SOTA with 9% of labels.

Table 3. Binary classification results for PCam dataset for a variety of models. Results marked by * are quoted from [23].

Labels	Model	Regular Model			Uncertainty-aware Model			
		Acc (%)	F1 (%)	AUC (%)	Acc (%)	F1 (%)	AUC (%)	
100%	Indomain	TransPath* [31]	81.20	81.00	91.70	-	-	-
	Indomain	Mocov3* [6]	86.30	86.20	95.00	-	-	-
	Indomain	DINO* [1]	85.80	85.60	95.70	-	-	-
	Indomain	SD-MAE* [23]	88.20	87.80	96.20	-	-	-
	Indomain	MAE [15]	88.41	86.23	95.81	-	-	-
	Indomain	SimCLRv1 [3]	83.21	84.40	88.67	-	-	-
	Indomain	SimCLRv2 [4]	90.57	90.20	96.47	90.29	89.95	96.49
	Outdomain	SimCLRv2 [4]	89.30	88.97	96.58	91.30	91.09	96.83
10%	Indomain	MAE [15]	86.10	84.45	94.81	-	-	-
	Indomain	SimCLRv1 [3]	88.67	81.52	83.45	-	-	-
	Indomain	SimCLRv2 [4]	89.73	89.07	96.19	88.27	88.94	94.69
	Outdomain	SimCLRv2 [4]	89.60	88.84	96.73	90.41	89.97	96.87
1%	Indomain	MAE [15]	85.81	86.10	94.45	-	-	-
	Indomain	SimCLRv1 [3]	87.77	88.67	81.52	-	-	-
	Indomain	SimCLRv2 [4]	90.27	89.99	95.34	88.96	88.54	94.24
	Outdomain	SimCLRv2 [4]	89.21	88.88	95.57	87.43	86.96	92.33

improved interpretability, characterized by better cluster border refinement and organization of points based on prediction uncertainty (refer Fig. 3.B2). Furthermore, it was observed that more interpretable predictions (i.e., incorrect predictions with higher uncertainty scores) were attainable when a larger number of labels were available. This highlights the effectiveness of uncertainty estimation in discerning prediction reliability.

We also plotted the histograms of uncertainty values of correct and incorrect predictions (see Fig. 4). Note that incorrect prediction histograms correspond to ‘B3’ and ‘D3’

of Fig. 3. In both the 100% and 1% settings, correct predictions exhibited a left-skewed distribution, while incorrect predictions displayed a right-skewed distribution. This observation indicates that the majority of incorrect predictions correlate with high uncertainty, whereas correct predictions tend to exhibit low uncertainty. This alignment underscores the effectiveness of the uncertainty estimation method in enhancing the interpretability of model predictions. The insight from Fig. 3 & Fig. 4 enabled us to develop a querying strategy for UA-AL.

Table 4. Multi-class classification results for NCT100k dataset for a variety of models. Results marked by * are quoted from [23]; Results marked by ** are quoted from [18].

Labels	Model	Regular Model		Uncertainty-aware Model		
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	
100%	Indomain	TransPath* [31]	92.80	89.90	-	-
	Indomain	Mocov3* [6]	94.40	92.60	-	-
	Indomain	DINO* [1]	94.40	91.60	-	-
	Indomain	BYOL** [13]	93.93	-	-	-
	Indomain	HistoSSL-Res** [18]	96.55	-	-	-
	Indomain	HistoSSL-ViT** [18]	96.18	-	-	-
	Indomain	SD-MAE* [23]	95.30	93.50	-	-
	Indomain	MAE [15]	94.70	94.20	-	-
	Indomain	SimCLRv1 [3]	92.10	92.20	-	-
	Indomain	SimCLRv2 [4]	96.28	96.25	96.44	96.39
Outdomain	SimCLRv2 [4]	96.85	96.82	95.88	95.82	
10%	Indomain	SimCLRv2 [4]	96.28	96.25	95.82	95.73
	Outdomain	SimCLRv2 [4]	94.62	94.56	94.98	94.87
1%	Indomain	SimCLRv2 [4]	94.27	94.12	91.70	91.65
	Outdomain	SimCLRv2 [4]	94.34	94.23	92.34	92.85

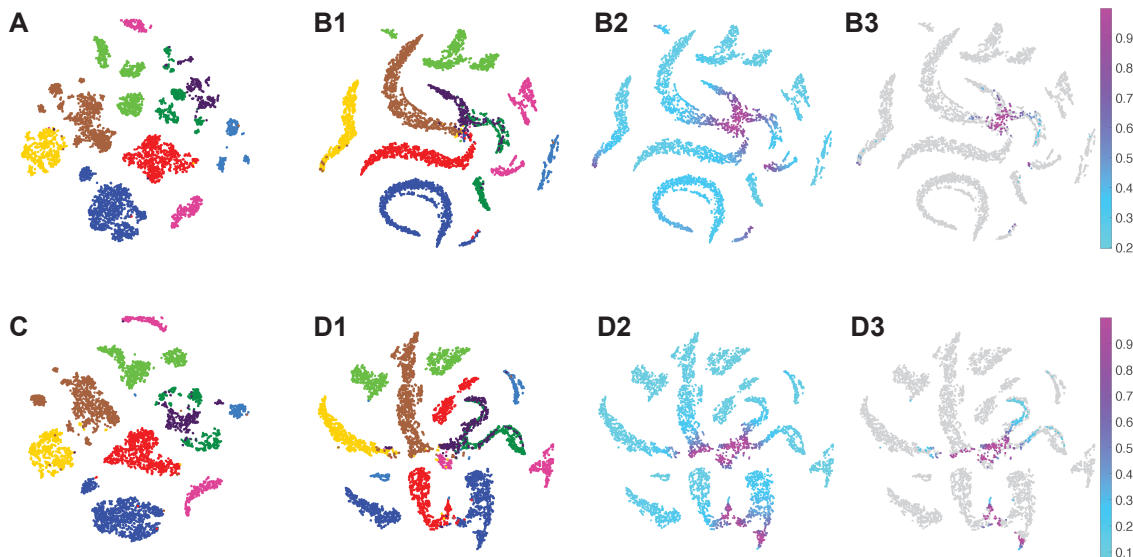


Figure 3. T-SNE plot (A) for SimCLRv2 trained in distribution with 100% annotations (B1) for UA-SimCLRv2 trained in distribution with 100% annotations (B2) color coded with the uncertainty values (Note that mixed cluster regions show high uncertainty) (B3) where only the Incorrect predictions are color coded. Note that most incorrect predictions show high uncertainty. C, D1, D2, D3 Corresponding versions of ‘A, B1, B2, B3’ with 1% of annotations. Note that in ‘D3’ there are more incorrect predictions with low uncertainty values than in ‘B3’.

4.2. Uncertainty-aware Label Selection

Table 5 showcases the accuracy and F1 score outcomes from training the UA-SimCLRv2 model, utilizing

an uncertainty-aware label selection approach. This is set against the backdrop of the SimCLRv2 model’s performance, as well as the UA-SimCLRv2 model when trained with a random image selection for labeling. In scenarios

Table 5. Results of UA-AL and random sampling of labels on the NCT100K dataset. rnd - Random Labeling, Sim - SimCLRv2, U.S - UA-SimCLRv2

Lab.	Indomain						Outdomain					
	Sim-rnd		U.Sim-rnd		U.Sim-AL		Sim-rnd		U.Sim-rnd		U.Sim-AL	
	Acc-%	F1-%	Acc-%	F1-%	Acc-%	F1-%	Acc-%	F1-%	Acc-%	F1-%	Acc-%	F1-%
1%	94.27	94.12	91.70	91.65	91.70	91.65	94.34	94.22	92.34	92.25	92.34	92.25
2%	93.57	93.46	94.59	94.13	96.26	96.15	92.31	92.23	93.89	93.58	94.38	94.34
3%	92.01	91.87	93.23	92.95	96.29	96.23	91.67	91.65	94.35	94.32	96.41	96.31
4%	91.22	91.20	95.30	95.18	95.35	95.35	91.98	91.92	94.91	94.90	95.33	95.21
5%	91.69	91.69	94.56	94.49	96.03	96.02	92.79	92.45	95.01	94.93	95.69	95.68
6%	91.68	91.65	94.06	93.94	95.93	95.91	92.13	92.11	94.34	94.21	96.28	96.25
7%	92.48	92.42	95.12	94.91	95.76	95.76	94.01	93.97	94.69	94.56	96.29	96.25
8%	92.08	92.05	95.01	94.89	96.50	96.45	93.87	93.46	95.05	94.98	96.25	96.12
9%	94.62	94.54	96.32	96.28	96.51	96.42	94.21	94.12	94.56	94.32	97.01	96.90
10%	96.28	96.25	95.82	95.73	96.51	96.49	94.62	94.56	94.98	94.87	96.40	96.33

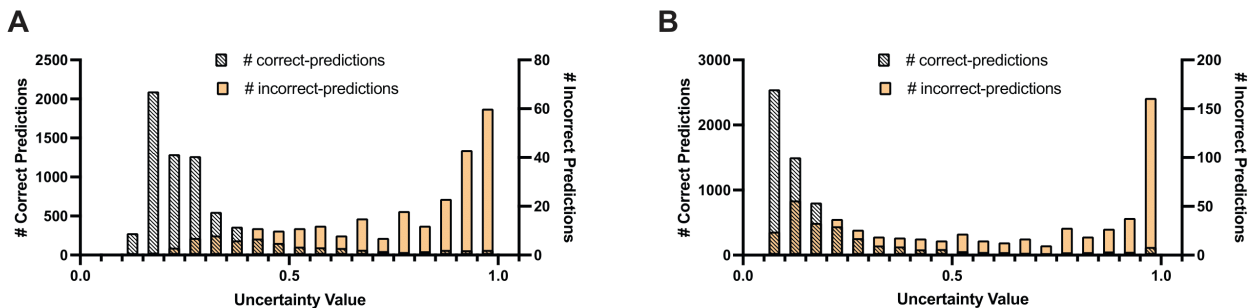


Figure 4. Histograms for (A) 100% annotations (B) 1% annotations demonstrating the tight coupling between model predictions accuracy and uncertainty awareness.

where training is conducted for indomain, we observed that the accuracy and F1 scores rapidly approached the baseline set by the 100% label setting with just 2% of labels. However, beyond this point, performance gains began to plateau. Notably, the UA-training method consistently outperformed models fine-tuned with randomly selected labels. The peak performance was recorded in an out-domain pre-training context, where it achieved superior results (refer Fig. 2.B) with only 9% of labels using uncertainty-aware labeling. This model not only surpassed the current state-of-the-art, HistoSSL-Res [18], but also outperformed the baseline model trained with 100% labels (as detailed in Table. 4).

The implications of these findings are threefold: firstly, UA-SimCLRv2 emerges as the foremost patch classifier on the NCT100k benchmarks. Secondly, even with randomly selected labels, UA-SimCLRv2 outperforms SimCLRv2 as label quantity increases. Lastly, employing uncertainty-aware label selection consistently leads to higher accuracy compared to random selection methods.

5. Conclusion

Our research represents a significant advancement in cancer subtyping for digital pathology, by integrating uncertainty awareness into SSL and AL frameworks. The UA-SimCLRv2 model offers superior interpretability in model predictions and performance, surpassing SOTA approaches with minimal labeled data. By strategically querying uncertain samples for annotation, our framework not only reduces annotation burdens but also enhances model precision and efficiency. These findings underscore the importance of incorporating uncertainty awareness into the learning process, particularly in critical domains like digital pathology.

With UA-SimCLRv2 established as the leading classifier on digital pathology benchmark datasets, our research sets a new standard in cancer subtyping in histopathology. This work can further be extended to whole slide image classification by using our fine-tuned encoder as the backbone to MIL approach. In essence, our work transforms digital pathology image analysis, introducing a new era of precision and efficiency led by the UA-AL with UA-SimCLRv2.

References

- [1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6, 7
- [2] Jacob Carse and Stephen McKenna. Active learning for patch-based digital pathology using convolutional neural networks to reduce annotation costs. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 20–27. Springer, 2019. 1, 2
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 6, 7
- [4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 1, 2, 6, 7
- [5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2
- [6] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 6, 7
- [7] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 1, 2
- [8] Arthur P Dempster et al. Upper and lower probabilities induced by a multivalued mapping. *Classic works of the Dempster-Shafer theory of belief functions*, 219(2):57–72, 2008. 4
- [9] James M Dolezal, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Brittany Cody, Aaron S Mansfield, Sagar Rakshit, Radhika Bansal, Melanie C Bois, et al. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature communications*, 13(1):6572, 2022. 2
- [10] Baolin Du, Qi Qi, Han Zheng, Yue Huang, and Xinghao Ding. Breast cancer histopathological image classification via deep active learning and confidence boosting. In *International Conference on Artificial Neural Networks*, pages 109–116. Springer, 2018. 2
- [11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2
- [12] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6928–6938, 2020. 2
- [13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 7
- [14] Wenlong Hang, Yecheng Huang, Shuang Liang, Baiying Lei, Kup-Sze Choi, and Jing Qin. Reliability-aware contrastive self-ensembling for semi-supervised medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 754–763. Springer, 2022. 1
- [15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 5, 6, 7
- [16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2
- [17] Xu Jin, Hong An, Jue Wang, Ke Wen, and Zheng Wu. Reducing the annotation cost of whole slide histology images using active learning. In *Proceedings of the 2021 3rd International Conference on Image Processing and Machine Vision*, pages 47–52, 2021. 1, 2
- [18] Xu Jin, Teng Huang, Ke Wen, Mengxian Chi, and Hong An. Histoss: Self-supervised representation learning for classifying histopathology images. *Mathematics*, 11(1):110, 2022. 7, 8
- [19] Audun Jøsang. Generalising bayes’ theorem in subjective logic. In *MFI*, pages 462–469, 2016. 4
- [20] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018. 3
- [21] S Kotz, N Balakrishnan, and NL Johnson. Continuous multivariate distributions—vol. 1, john wiley & sons, new york, 2000. 4
- [22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2
- [23] Yang Luo, Zhineng Chen, and Xieping Gao. Self-distillation augmented masked autoencoders for histopathological image classification. *arXiv preprint arXiv:2203.16983*, 2022. 1, 2, 6, 7
- [24] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 2
- [25] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 1, 2, 4
- [26] Ashish Seth, Sreyan Ghosh, S Umesh, and Dinesh Manocha. Fusdom: Combining in-domain and out-of-domain knowledge for continuous self-supervised learning. *arXiv preprint arXiv:2312.13026*, 2023. 1

- [27] Burr Settles. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011. [1](#)
- [28] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. [2](#)
- [29] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. [3](#)
- [30] Sida Wang and Christopher Manning. Fast dropout training. In *international conference on machine learning*, pages 118–126. PMLR, 2013. [2](#)
- [31] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021. [6](#), [7](#)
- [32] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems*, 33:6514–6527, 2020. [2](#)
- [33] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. [2](#)