Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Contrastive Deep Encoding Enables Uncertainty-aware Machine-learning-assisted Histopathology

Dushan Wadduwage (wadduwage@fas.harvard.edu)
Harvard University https://orcid.org/0000-0002-0689-9454
Nirhoshan Sivaroopan
University of Moratuwa
Chamuditha Jayanga
University of Moratuwa
Chalani Ekanayake
University of Moratuwa
Hasindri Watawana
University of Moratuwa
Jathurshan Pradeepkumar
Harvard University
Mithunjha Anandakumar
Harvard University
Ranga Rodrigo
University of Moratuwa
Chamira Edussooriya
University of Moratuwa https://orcid.org/0000-0001-6715-6198

Article

Keywords: Whole Slide Image (WSI), Uncertainty Awareness (UA), knowledge distillation, self supervised learning.

Posted Date: October 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3317700/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

DEEP ENCODING FOR MACHINE LEARNING ASSISTED HISTOPATHOLOGY

Nirhoshan Sivaroopan[1,†], Chamuditha Jayanga[1,†], Chalani Ekanayake[1,†], Hasindri Watawana[1,†], Jathurshan Pradeepkumar[2], Mithunjha Anandakumar[2], Ranga Rodrigo[1], Chamira U. S. Edussooriya[1], & Dushan N. Wadduwage[2,*]

[1] Dept. of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka
[2] Center for Advanced Imaging, Harvard University, Cambridge, MA, 02138
[*] wadduwage@fas.harvard.edu

ABSTRACT

Deep neural network models can learn clinically relevant features from millions of histopathology images. However generating high-quality annotations to train such models for each hospital, each cancer type, and each diagnostic task is prohibitively laborious. On the other hand, terabytes of training data –while lacking reliable annotations– are readily available in the public domain in some cases. In this work, we explore how these large datasets can be consciously utilized to pre-train deep networks to encode informative representations. We then fine-tune our pre-trained models on a fraction of annotated training data to perform specific downstream tasks. We show that our approach can reach the current state-of-the-art (SOTA) for patch-level classification with only 1-10% randomly selected annotations compared to other SOTA approaches. Moreover, we propose an uncertainty-aware loss function, to quantify the model confidence during inference. Quantified uncertainty helps experts select the best instances to label for further training. Our uncertainty-aware labeling reaches the SOTA with significantly fewer annotations compared to random labeling. Last, we demonstrate how our pre-trained encoders can surpass current SOTA for whole-slide image classification with weak supervision. Our work lays the foundation for data and task-agnostic pre-trained deep networks with quantified uncertainty.

Index Terms— Whole Slide Image (WSI), Uncertainty Awareness (UA), knowledge distillation, self-supervised learning.

1. INTRODUCTION

Computer-assisted diagnostics using deep learning is an emerging field. Especially for cancer, deep models have demonstrated their potential in the clinic by effectively streamlining labor-intensive and error-prone procedures, thereby complementing pathologists in the diagnostic process. For instance, in the detection of breast cancer metastases, a prevalent form of cancer, the groundbreaking work by [1] has made significant strides in reducing human error rates by integrating deep learning models with human pathologists' assessments. Similarly, during the last decade, deep learning has demonstrated its capability as a promising tool in digital pathology, with numerous studies highlighting its diverse clinical applications. These include clinical diagnosis [2–5], prognosis/survival prediction [6–9], treatment response forecasting [10, 11], and the identification of regions of interest (RoIs) that exhibit substantial diagnostic value [12–14]. Thus, machine-learning-driven digital pathology can potentially enhance multiple facets of the clinical process including accuracy, analysis speed, and reproducibility [15–17].

Despite notable achievements of deep learning in histopathology, several challenges persist [18]. First, unlike conventional images where the target object occupies a substantial portion of the image, cancer-positive histopathology images typically feature small regions of positive activations over a large area of normal tissue. One should therefore image many biopsy specimens to collect a sufficiently large training dataset. Some cancer types are rare and access to such samples is limited to many. Furthermore, patient data cannot be readily shared due to privacy concerns. This lack of access to training data discourages open-source communities from developing better models. Second, to train a fully supervised model, images should be finely annotated by an expert pathologist identifying visually intricate patterns critical for accurate diagnosis. Such careful annotations are time-consuming. Return on expert time invested in annotating is also not guaranteed in terms of the final performance of the model. This uncertainty discourages expert pathologists from spending time annotating large datasets to a standard of accuracy needed to train a reliable deep model fully supervised. Third, most deep-learning models for computational pathology are not interpretable. The end user

[†] These authors contributed equally to the work.

is not made aware of the uncertainty in model predictions. This lack of transparency makes it challenging to integrate these models into the existing clinical decision process. In this work we attempt to tackle all three of these challenges by: using publicly available large datasets; self-supervised pre-training followed by supervised fine-tuning with only a few annotations; and explicitly quantifying the level of uncertainty.

To this end, based on the seminal SimCLRv2 self-supervised framework [19], we introduce an uncertainty-aware contrastive learning method for digital pathology. Our work results in three significant advancements over the current state-of-the-art (SOTA). First, we overcome annotation limitations by training an accurate model with minimal labeled training samples. Second, we establish our model as a versatile framework capable of adapting to various clinical tasks. Third, we quantify the uncertainty in our model predictions, empowering pathologists to make more informed decisions in the context of model confidence. With these contributions, our approach outperformed current SOTA in both patch and slide-level predictions on multiple benchmark datasets while uniquely quantifying risks associated with uncertain model predictions.

2. RELATED WORK

2.1. Deep learning for digital pathology

Some seminal work that laid foundation for machine-learning-assisted histopathology include: [20] that applied logistic regression to investigate the correlation between histology of the surrounding connective tissue of tumor cells and prognosis in breast cancer; [21] that studied survival prediction for lung cancers using regularized machine learning and automatic feature extraction; and [22] that visually examined individual node responses in the last hidden layer of a convolutional neural network, to uncover biologically insightful interpretable characteristics in histopathology images.

2.2. Self-supervised representation learning

Self-supervised learning (SSL) [23–30], has proven successful in computer vision to pre-train large models with unlabelled data. SimCLR [31], a contrastive learning framework, is an example of this approach to learn visual representations of images. SimCLRv2 [19] (i.e., the second version) enhanced SimCLR in multiple ways: utilizing larger backbone networks during pre-training, increasing the capacity of the projection head, incorporating a memory mechanism from MoCo [32], and leveraging knowledge distillation [33] using unlabeled examples. Masked auto-encoding (MAE) [34] is another SSL approach that reconstructs signals from partial observations. It consists of an encoder that maps a partially masked image to a latent representation and a decoder that reconstructs the original image from it. By masking 75% of the input patches of the image, MAE enabled scalable training of large models using a reconstruction loss on masked patches. The efficacy of MAE has been demonstrated in various downstream tasks, including image classification.

Both SimCLR and MAE allow learning of rich representations from large amounts of unlabeled data and have also been effectively used in digital pathology. [35] adapted SimCLR for digital pathology. [36] used MAE in digital pathology and introduced a self-distillation loss on top of the masked patch reconstruction loss from original MAE. Nevertheless, more advanced SimCLRv2 has not been adopted for digital pathology. Moreover, none of these SSL models have been investigated in relation to the uncertainty of their predictions.

2.3. Uncertainty Quantification

Estimating uncertainty of deep learning models is an active area of reaserch in machine learning. A straightforward appraoch is to use the model to output a distribution of predictions for each input rather than a single prediction. The variability of the predicted distribution can then be used to quanitfy uncertainity. Monte Carlo dropout method is one such approach, where each input is forward passed multiple times through the model (during both training and inference), while randomly dropping out different network nodes. Here, the standard deviation of the generated distribution of outputs is an estimator for uncertainty [37, 38]. Deep ensembles is another method to quantify uncertainty. It generates a distribution of predictions by forwarding each input through a set of deep learning models that are trained with different initializations and hyperparameters [39, 40]. Entropy of the model output is then used to estimate uncertainty. Test-time-augmentation is another uncertainty quantification method. It applies random transformations to each image and obtains predictions for each. Prediction variance among the perturbed versions of the same image is used as the uncertainty estimation.

Despite the critical clinical importance, most deep learning frameworks for histopathology do not quantify the confidence of model predictions. A notable exception is [41] that quantified histologic ambiguity of images using an uncertainty score. The uncertainty here is estimated during inference by sending each image tile through 30 forward passes in a dropout-enabled network, resulting in a distribution of predictions. The standard deviation of predictions for a single image patch represents patch-level uncertainty.

In our work we augmented SSL frameworks to explicitly include uncertainity estimation. We adopted a Bayesian approach first proposed by [42] using the the-

ory of evidence . [42] compared this Bayesian approach with others to show its utility for out-domain training. As we focus on task-agnostic out-domain learning from public datasets, we leverage this method in our work.

3. RESULTS

3.1. SimCLRv2 and uncertainty-aware(UA-) SimCLRv2

We experimented with two state-of-the-art self supervised pre-training approaches on two publicly available patch-level datasets. We tested masked auto-encoding (MAE) [34] using a transformer backbone and contrastive learning using SimCLR frameworks (v1 [31] and v2 [43]). Out of these SimCLR-v2 performed best and hence was selected for this work. The training procedure of SimCLRv2 involves three stages (see Fig. 1.A1-3). First, a ResNet backbone undergoes self-supervised pre-training, where it learns representations through a stack of projection layers (see Fig. 1.A1). This process utilizes a contrastive loss function to compare the embeddings generated by the projected head. Second, a classification head is appended to the pre-trained encoder, and the model is fine-tuned fully-supervised using a small fraction of annotated patches (see Fig. 1.A2). Last, the learned classifier, i.e. the teacher model, transfers its knowledge to an architecturally identical student model through a process known as knowledge distillation (see Fig. 1.A3). Here the teacher's pseudo labels on the entire large unlabelled dataset are used as ground-truths for the student model. Note that the models use annotations only in the fine-tuning stage (see Fig. 1.A2).

While the SimCLRv2 model with a ResNet-50 backbone demonstrates superior performance in various clinical tasks (see our results in the next section), it lacks interpretability. Medical artifical intelligence (AI) applications, however, require transparent and explainable results due to the potential risks associated with predictions. To address this issue, we propose a modified version of SimCLRv2 called Uncertainty-aware SimCLRv2 (UA-SimCLRv2) shown in Fig. 1.B. UA-SimCLRv2 follows a similar pre-training process as SimCLRv2; however, it undergoes fine-tuning and distillation using MSE loss in the Bayesian view. Notably, UA-SimCLRv2 not only provides class labels but also provides an associated uncertainty score for each patch prediction. We discuss UA-SimCLRv2 in detail in the methods section 5.

In the next section, we present SimCLRv2 and UA-SimCLRv2 results for patch-level classification tasks in the digital pathology domain.

3.2. Patch classification results using SimCLRv2 and UA-SimCLRv2

Histopathology studies analyze data at multiple levels of hierarchy including patch-level, slide-level, and patient-level. We first employ SimCLRv2 and UA-SimCLRv2 as patch-level classifiers (later we adapt patch-level trained models to slide-level). In our study, we utilized two datasets: PatchCamelyon (PCam) and NCT-CRC-HE-100k (NCT100k). The experimental setup for evaluating model performances on these datasets included two variables: (1) the percentage of annotations used for fine-tuning from the training set, and (2) whether pre-training was performed on in-distribution data or out-distribution data. For example, in PCam experiments, out-distribution-pre-training means that pre-training was done on the NCT-CRC-HE-100k dataset (and vice versa in NCT-CRC-HE-100k experiments).

Binary classification results on PCAM. The PCam dataset [44] comprised a collection of 327,680 patches extracted from histopathology scans of lymph node sections in the CAMELYON16 dataset [45]. Each patch had a corresponding binary label indicating the presence of metastatic tissue. According to [44], a positive patch contained at least one pixel of tumor tissue in its center 32x32 pixel region. Patches were 96x96 pixels in size, but for consistency across experiments, we resized all patches to 224x224 pixels (see the methods section for more details). We divided the dataset into training (75%), validation (12.5%), and test (12.5%) splits, ensuring a balanced distribution of positive and negative examples with no overlap between the splits.

Fig. 2 and Table 1 show the accuracy, F1 score, and AUC score (Area Under the ROC Curve) 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 current state-of-the-art (SOTA) approach, i.e., MAE [34]. Notably, out-distribution-pre-trained UA-SimCLRv2 performed best with 2.89% increase in accuracy, 2.74% in F1 score and 0.77% in AUC score compared to the current SOTA.

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 current SOTA (see Fig. 2 and Table 1). Last, we fine-tuned our models on 1% training labels. Interestingly the SimCLRv2 and UA-SimCLRv2 models still performed comparable to the current SOTA (see the 1% setting on Fig. 2 and Table 1). However, at the 1% setting UA-SimCLRv2 consistently underperformed compared to SimCLRv2, perhaps due to the limited evidence available for uncertainty awareness (see method section 5 for more details).

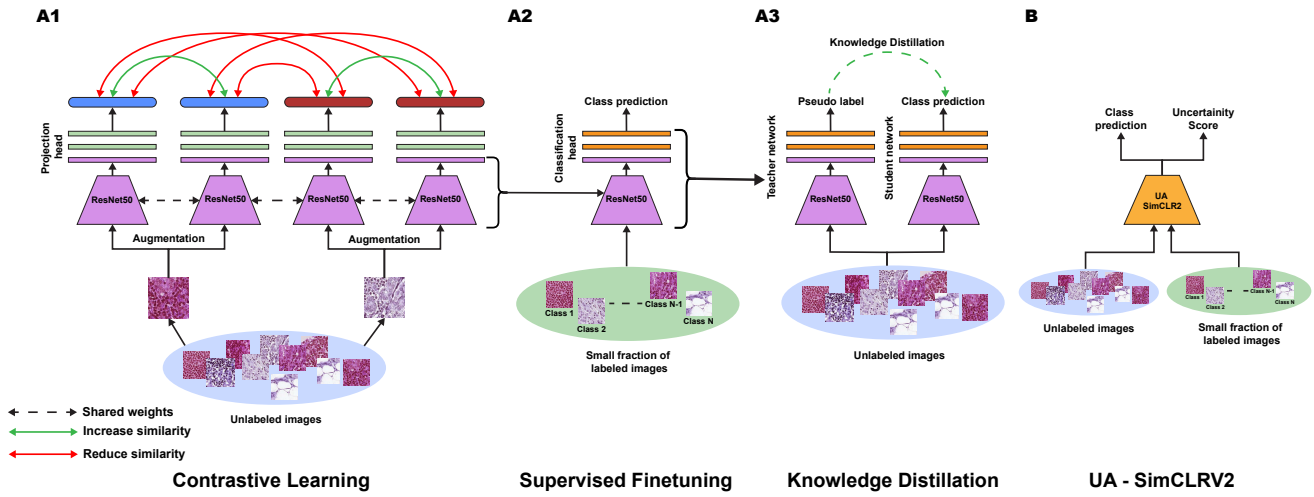


Fig. 1: The SimCLRv2 framework. (A1) The pre-training step. Contrastive Learning is used to pre-train a deep neural encoder using a large set of unlabelled images. (A2) The supervised fine-tuning step. A classification head is added to the pre-trained encoder. The model is then fine-tuned using a small fraction of labeled images. (A3) The knowledge distillation step. The model from 'A2' is used as a teacher network to generate pseudo labels for all unlabeled training images. Then the pseudo labels are used to train a student network (with the same architecture). (B) The proposed Uncertainty-aware(UA) SimCLRv2 model with an additional output for the uncertainty score.

Multi class classification on NCT100k. The NCT100k dataset [51] comprised 100,000 non-overlapping image patches extracted from hematoxylin and eosin (H&E) stained histological images of human colorectal cancer and normal tissues. Each patch was 224x224 pixels in size, and annotated with one of nine classes based on the type of tissue: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM). According to NCT100k dataset's guidelines, we trained our models with NCT100k and tested with CRC-VAL-HE-7K (CRC7k) dataset [51], which consisted of samples from the same data distribution.

Fig. 3 and Table 3 show 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 current SOTA. Interestingly, out-distribution-pre-trained SimCLRv2 was the best-performing model and surpassed the current SOTA by a small margin. At the 10% setting, our models still performed comparable to the 100% baseline and current SOTA. But at the 1% setting, we observed a clear degradation of performance by a few percentage points.

We further investigated the model behaviors at 100% and 1% settings using t-distributed stochastic neighbor embedding (T-SNE) on the learned feature representations. Fig. 3.C&E show the T-SNE maps for SimCLRv2 at 100% & 1% settings respectively. Fig. 3.D1&F1 show the same for UA-SimCLRv2. Compared to the 100% setting, the 1% setting of UA-SimCLRv2 showed more overlapping clusters. For instance, the NORM class heavily overlapped with the TUM class (see Fig. 3.F1). Fig. 3.D2&F2 show the same T-SNE plots from UA-SimCLRv2 in D1&F1 but color-coded with the associated uncertainty of the predictions. Interestingly, the overlapping regions showed high uncertainty. In Fig. 3.D3&F3 we further color-coded only the incorrect predictions. We observed that in the 100% setting most incorrect predictions were associated with higher uncertainty (Fig. 3.D3). But in the 1% setting some incorrect predictions were associated with lower uncertainty (Fig. 3.F3). We also plotted the histograms of uncertainty values of correct and incorrect predictions (see Fig. 3.G&H). For both 100% and 1% settings, correct predictions showed a left-skewed distribution; incorrect predictions showed a right-skewed distribution. Thus uncertainty in predictions allows us to identify the data classes that lead to inaccurate predictions. This insight enabled us to develop a sample selection procedure we call *uncertainty-aware training* (UA-training) to fine-tune UA-SimCLRv2. Uncertainty-aware training deviates from the random selection of 1% annotations employed while keeping the fine-tuning process unchanged.

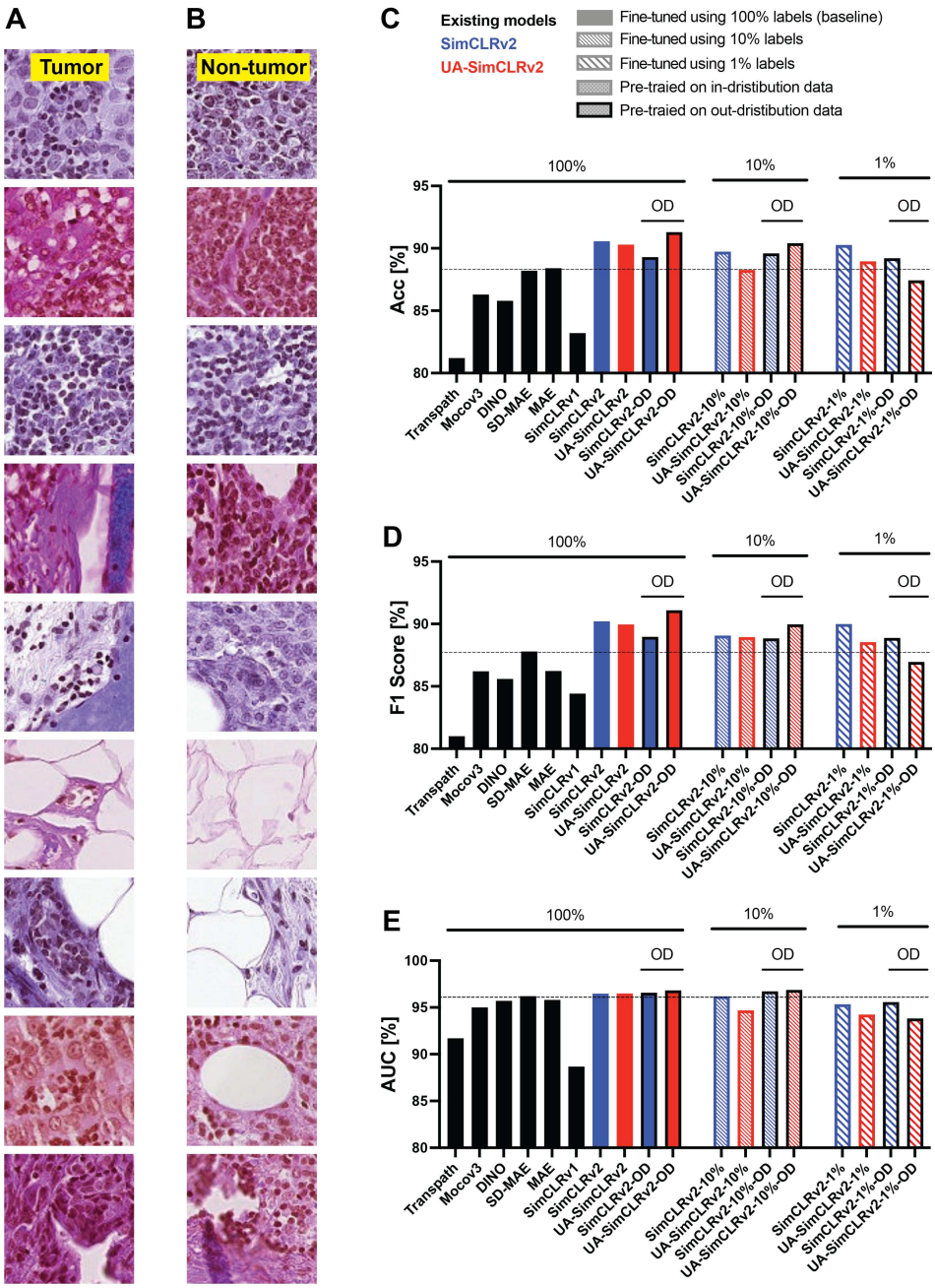


Fig. 2: Binary classification results for PCam dataset for an ensemble of models. (A) Representative images from the tumor class. (B) Representative images from the non-tumor class. (C) Classification accuracy for competing models (shown in black) vs. the proposed models: SimCLRv2 and UA-SimCLRv2 (i.e. Uncertainty Aware SimCLRv2). The proposed models were first trained on 100% of annotations to generate the baselines. Models were pretrained using in-distribution data (i.e. training data from NCT100k) as well as out-of-distribution data (i.e. training data from PCam). Then the same experiments were repeated with 10% of annotations, and 1% of annotations. (D) F1-scores for the same experiments in 'C'. (E) Area under the curve (AUC) for the same classification experiments in 'C'.

Table 1: Binary classification results for PCam dataset for an ensemble of models under 100%, 10%, and 1% labels settings. For selected cases, MAE, Simclrv1, and Simclrv2 models were modified with uncertainty-aware loss; the corresponding results are shown in the "Uncertainty-aware Model" columns. Results marked by $*$ are quoted from [36]. The rest are from our experiments. Results marked by † are from our selected Simclrv2 approach. The original references for the model architectures are shown next to each model.

			Regular Model			Uncertainty-aware Model		
Labels	**Training**	**Model**	Acc (%)	F1 (%)	AUC (%)	Acc (%)	F1 (%)	AUC (%)
100%	Indomain	TransPath* [46]	81.20	81.00	91.70	-	-	-
	Indomain	Mocov3* [47]	86.30	86.20	95.00	-	-	-
	Indomain	DINO* [48]	85.80	85.60	95.70	-	-	-
	Indomain	SD-MAE* [36]	88.20	87.80	96.20	-	-	-
	Indomain	MAE [34]	88.41	86.23	95.81	-	-	-
	Indomain	Simclrv1 [31]	83.21	84.40	88.67	-	-	-
	Indomain	**Simclrv2**† [43]	90.57	90.20	96.47	90.29	89.95	96.49
	Outdomain	**Simclrv2**† [43]	89.30	88.97	96.58	**91.30**	**91.09**	**96.83**
10%	Indomain	**Simclrv2**† [43]	89.73	89.07	96.19	88.27	88.94	94.69
	Outdomain	**Simclrv2**† [43]	89.60	88.84	96.73	**90.41**	**89.97**	**96.87**
1%	Indomain	MAE [34]	86.10	94.45	95.81	85.81	86.10	94.45
	Indomain	Simclrv1 [31]	88.67	81.52	83.45	87.77	88.67	81.52
	Indomain	**Simclrv2**† [43]	**90.27**	**89.99**	**95.34**	88.96	88.54	94.24
	Outdomain	**Simclrv2**† [43]	89.21	88.88	95.57	87.43	86.96	92.33

Table 2: Multi-class classification results for NCT100k dataset for an ensemble of models under 100%, 10% and 1% labels settings. For selected cases, MAE, Simclrv1, and Simclrv2 models were modified with uncertainty-aware loss; the corresponding results are shown in the "Uncertainty-aware Model" columns. Results marked by $*$ are quoted from [36]; Results marked by $**$ are quoted from [49]. The rest are from our experiments. Results marked by † are from our selected Simclrv2 approach. The original references for the model architectures are shown next to each model.

			Regular Model		Uncertainty-aware Model	
Labels	**Training**	**Model**	Acc (%)	F1 (%)	Acc (%)	F1 (%)
100%	Indomain	TransPath* [46]	92.80	89.90	-	-
	Indomain	Mocov3* [47]	94.40	92.60	-	-
	Indomain	DINO* [48]	94.40	91.60	-	-
	Indomain -	BYOL** [50]	93.93	-	-	-
	Indomain	HistoSSL-Res** [49]	96.55	-	-	-
	Indomain	HistoSSL-ViT** [49]	96.18	-	-	-
	Indomain	SD-MAE* [36]	95.30	93.50	-	-
	Indomain	MAE [34]	94.70	94.20	-	-
	Indomain	Simclrv1 [31]	92.10	92.20	-	-
	Indomain	**Simclrv2**† [43]	96.28	96.25	96.44	96.39
	Outdomain	**Simclrv2**† [43]	**96.85**	**96.82**	95.88	95.82
10%	Indomain	**Simclrv2**† [43]	**96.28**	**96.25**	95.82	95.73
	Outdomain	**Simclrv2**† [43]	94.62	94.56	94.98	94.87
1%	Indomain	MAE [34]	93.40	92.68	-	-
	Indomain	**Simclrv2**† [43]	94.27	94.12	91.70	91.65
	Outdomain	**Simclrv2**† [43]	**94.34**	**94.23**	92.34	92.85

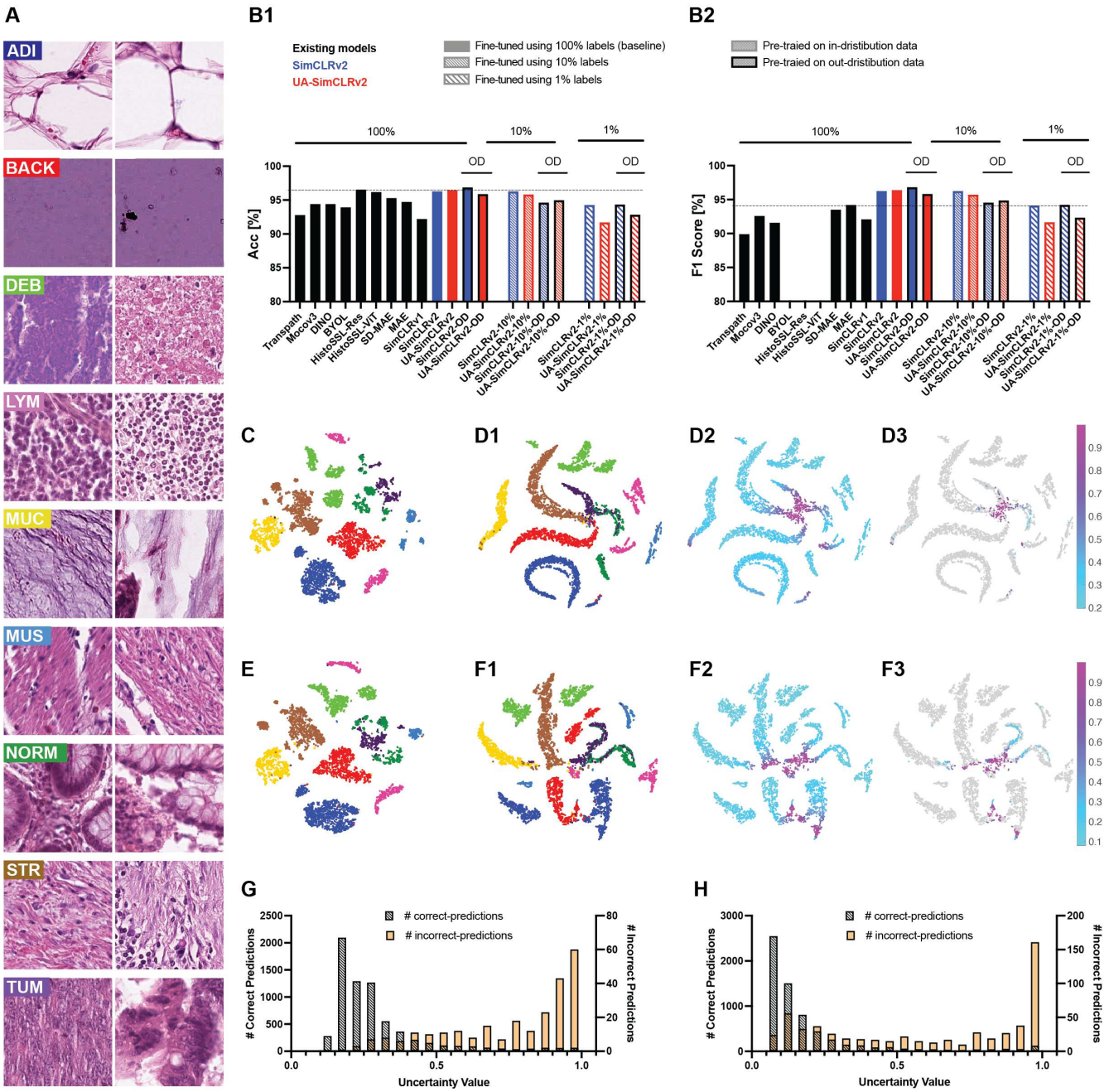


Fig. 3: Multi-class classification results for NCT100k dataset for an ensemble of models. (A) Representative images from nine classes in the dataset. (B1) Classification accuracy for competing models (shown in black) vs. the proposed models: SimCLRv2 and UA-SimCLRv2 (i.e. Uncertainty Aware SimCLRv2). The proposed models were first trained on 100% of annotations to generate the baselines. Models were pre-trained using in-distribution data (i.e. training data from NCT100k) as well as out-domain data (i.e. training data from PCam). Then the same experiments were repeated with 10% of annotations, and 1% of annotations. (B2) F1-scores for the same experiments in 'A'. (C) T-SNE plot for SimCLRv2 trained in distribution with 100% annotations. (D1) T-SNE plot for UA-SimCLRv2 trained in distribution with 100% annotations. Note that there are four clusters that were hard for the model to separate. (D2) T-SNE plot color coded with the uncertainty values. Note that mixed cluster regions show high uncertainty. (D3) T-SNE plot where only the Incorrect predictions are color coded. Note that most incorrect predictions show high uncertainty. (E, F1, F2, F3) Corresponding versions of 'C, D1, D2, D3' with 1% of annotations. Note that in 'F3' there are more incorrect predictions with low uncertainty values than in 'D3'. (G) Histogram of uncertainty values with correct vs. incorrect predictions for 100% annotations (H) Same as in 'G' with 1% annotations. Note that incorrect prediction histograms correspond to 'D3' and 'F3'.

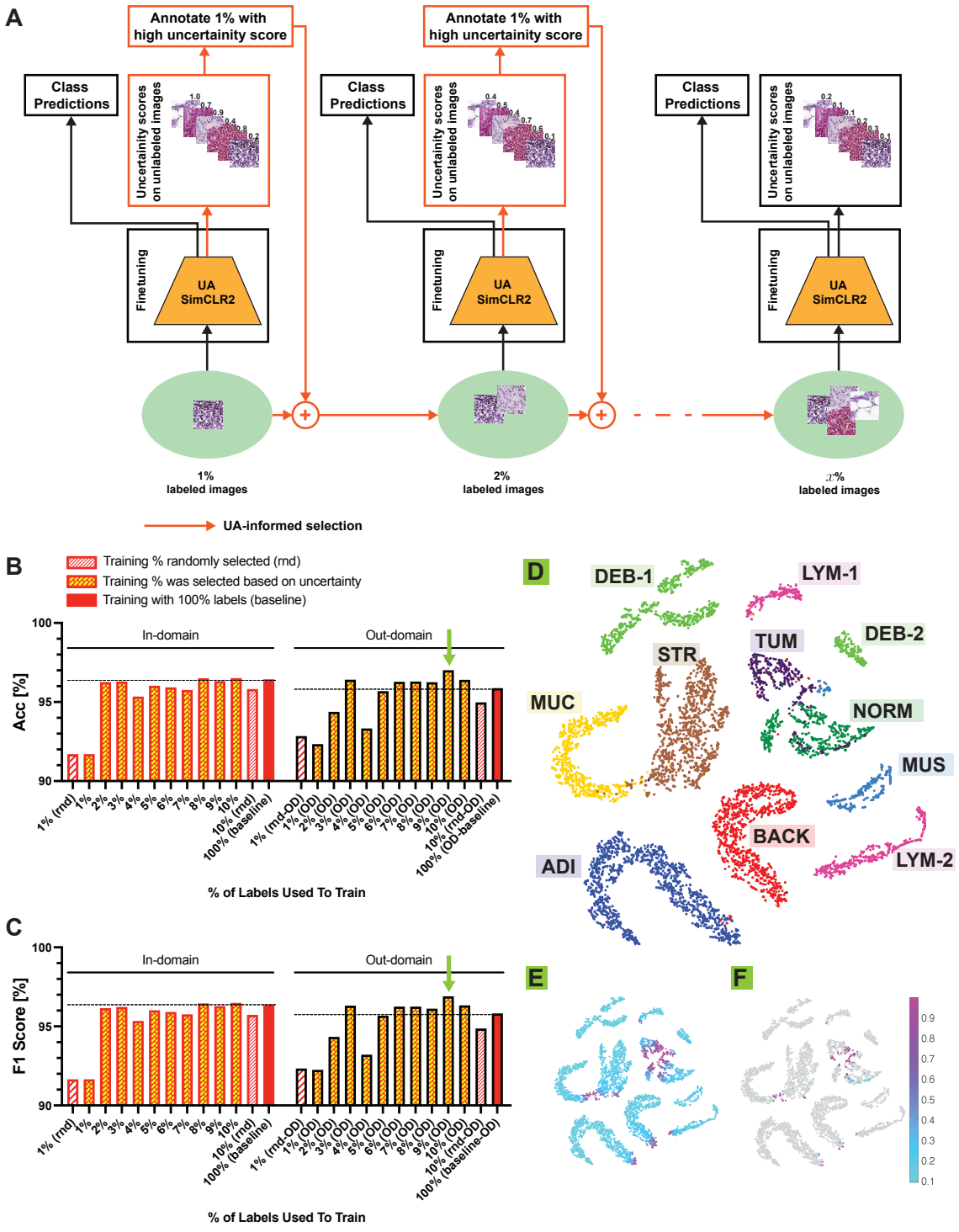


Fig. 4: (A) Uncertainty-aware fine-tuning of the UA-SimCLRv2 model for NCT100k. UA-SimCLRv2 was pretrained using in-distribution data (i.e. training data from NCT100k) as well as out-domain data (i.e. training data from PCam). Then the pretrained models were finetuned using a randomly selected 1% of annotations. Next, the uncertainty values were calculated using the 1%-finetuned model for all remaining training data and another 1% of labels with high uncertainty score were annotated to have 2% of annotations for further fine-tuning. This procedure was repeated until the 10% of annotations were used for fine-tuning. (B) Classification accuracy for Uncertainty-aware fine-tuning vs. randomly selected fine-tuning. (C) F1 score for Uncertainty-aware fine-tuning vs. randomly selected fine-tuning. (D) T-SNE map of the learned features from the final distilled model for the best performing case (i.e. out-domain pretrained model with Uncertainty-aware fine-tuning up to 9% annotations – also marked using the green arrow in 'B'). (E) The same T-SNE map in 'D' color coded using the uncertainty score. (F) The same T-SNE map in 'E' with only wrong classifications color coded.

Table 3: Results from uncertainty-informed training of UA-SimCLRv2 on the NCT100K dataset.

	In-domain		Out-domain	
Labels	Acc (%)	F1 (%)	Acc (%)	F1 (%)
1%	91.70	91.65	92.34	92.25
2%	96.26	96.15	94.38	94.34
3%	96.29	96.23	96.41	96.31
4%	95.35	95.35	93.33	93.21
5%	96.03	96.02	95.69	95.68
6%	95.93	95.91	96.28	96.25
7%	95.76	95.76	96.29	96.25
8%	96.50	96.45	96.25	96.12
9%	96.32	96.28	97.01	96.90
10%	96.51	96.49	96.40	96.33

3.3. Uncertainty-aware Fine-tuning

The proposed uncertainty-aware training approach is presented in Fig. 4.A. We started fine-tuning using 1% randomly selected labeled patches (from the training set of the target task). The resulting model was then inferred on the remaining training set, and uncertainty scores were computed. We then included annotations of the top 1% patches with the highest uncertainty scores. The annotated patches, totaling 2% of the training set, were then used for the subsequent fine-tuning step. This iterative process emulates a scenario where only highly uncertain patches are selectively labeled by an expert pathologist. The process was repeated until 10% of the training set was annotated for fine-tuning. At each step, the fine-tuned UA SimCLRv2 model was saved and evaluated on the CRC7k dataset (i.e. the test set for multi-class classification) to generate test performance metrics.

Fig. 4 and Table 3 demonstrate the accuracy and F1 score for uncertainty-aware training. In the in-domain training settings, the accuracy and F1-scores immediately reached just below the 100%-setting baseline with only 2% labels; but the performance saturated afterwards. Nevertheless, in both in- and out-domain pre-training settings, UA-training outperformed fine tuning with random selections (compare 2-10% bars with their corresponding 10% rnd bars in Fig. 4 [A-B]). The best-performing case was achieved in the out-domain pre-training setting (i.e. UA SimCLRv2 pre-trained on PCam) at 9% of the labels. Interestingly, this model outperformed both the current SOTA, i.e. HistoSSL-Res [49], and the 100% baseline model.

These results establish the proposed UA-SimCLRv2 as the SOTA patch classifier on NCT100k benchmarks. We next present how the SimCLRv2 and UA-SimCLRv2 models trained on patch-level data can be used as encoders in slide-level classification.

3.4. Whole Slide Image (WSI) classification using Multiple Instance Learning (MIL)

In addition to establishing SimCLRv2 as the SOTA patch classifier, we demonstrate its versatility in adapting to whole slide image (WSI) classification at the slide-level. WSI classification is typically performed using multiple instance learning (MIL) under weakly-supervised settings. Here only slide-level annotations are available. We adapted SimCLRv2 and UA-SimCLRv2 for WSI classification on the CAMELYON16 dataset [45]. CAMELYON16 consisted of 400 H&E WSIs of lymph nodes, with metastatic regions labeled. The dataset consists of two sets: a training set consisting 270 WSIs and a testing set consisting 130 WSIs.

We adapted DTFD-MIL [52], the current SOTA WSI classifier on CAMELYON16, as our WSI classifier (Fig. 1[D]). DTFD-MIL leverages a ResNet-50 backbone to extract features from patches obtained from WSIs. Features extracted from all the patches (from a particular slide) are then combined using an attention mechanism to predict the slide-level label. In our adaptation, we replaced the ResNet-50 feature extractor (which is pre-trained on ImageNet [58] in a fully supervised setting) with a ResNet-50 that is pre-trained, fine-tuned, and distilled within our proposed SimCLRv2 and UA-SimCLRv2 frameworks using either the CAMELYON16, NCT100k, or PCam datasets.

Fig. 5 shows variation of test accuracy, F1 score and AUC scores comparing DTFD-MIL with and without our encoder. The exact values are reported in Table 4. In all 100%, 10%, and 1% settings, some version of the SimCLRv2 and UA-SimCLRv2 models outperformed the current SOTA by a few percentage points in all three performance metrics. This result shows the value of introducing a few patch-level annotations to train an encoder for slide-level classification. We further investigated the effect of patch-level contrastive pre-training alone on the slide-level encoder. In this setting, no patch-level annotations were used to pre-train the encoder. To this end, we pre-trained the SimCLRv2 encoder using patches obtained by splitting the WSIs in CAMELYON16. This model too outperformed the current SOTA by a few percentage points, in all three performance metrics (see Camelyon16 case in Fig. 5 and)Table 4).

These results highlight the capability of our proposed approach to achieve accurate and interpretable machine learning with minimal supervision.

4. DISCUSSION AND CONCLUSION

In summary, based on the seminal SimCLRv2 self-supervised framework [19], we introduce an uncertainty-aware contrastive learning method for digital pathology.

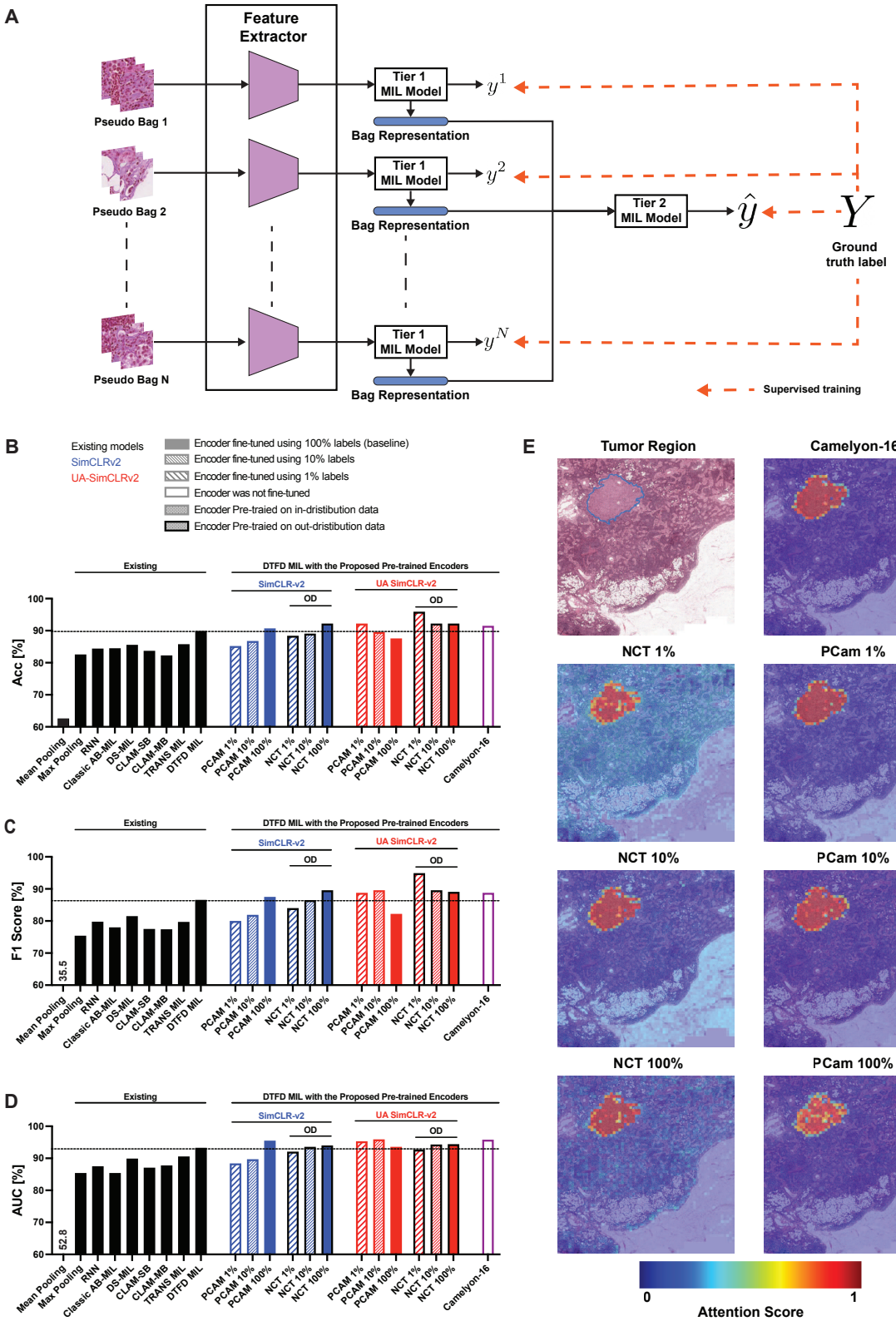


Fig. 5: Using the proposed models trained on patch level data as encoders for multiple instance learning (MIL). (A) DTFD [52] architecture has been used for MIL evaluation. (B) MIL classification accuracy for existing methods and DTFD using our proposed encoders with multiple combinations of models, % of annotated patches, datasets. Shown in solid black are the competing methods with DTFD performing the best. We then replaced the encoder of DTFD with our encoders and re-trained the rest of the MIL framework and the results are shown in colored bars. (C) F1 score for the same experiments in 'B'. (D) Area under the curve (AUC) for the same experiments in 'B'. Note that, in the spirit of only using slide-level ground truths for training, the camelion-16 pretrained model didn't use any annotated patch level labels. (E) MIL attention score maps for a representative test slide for different models.

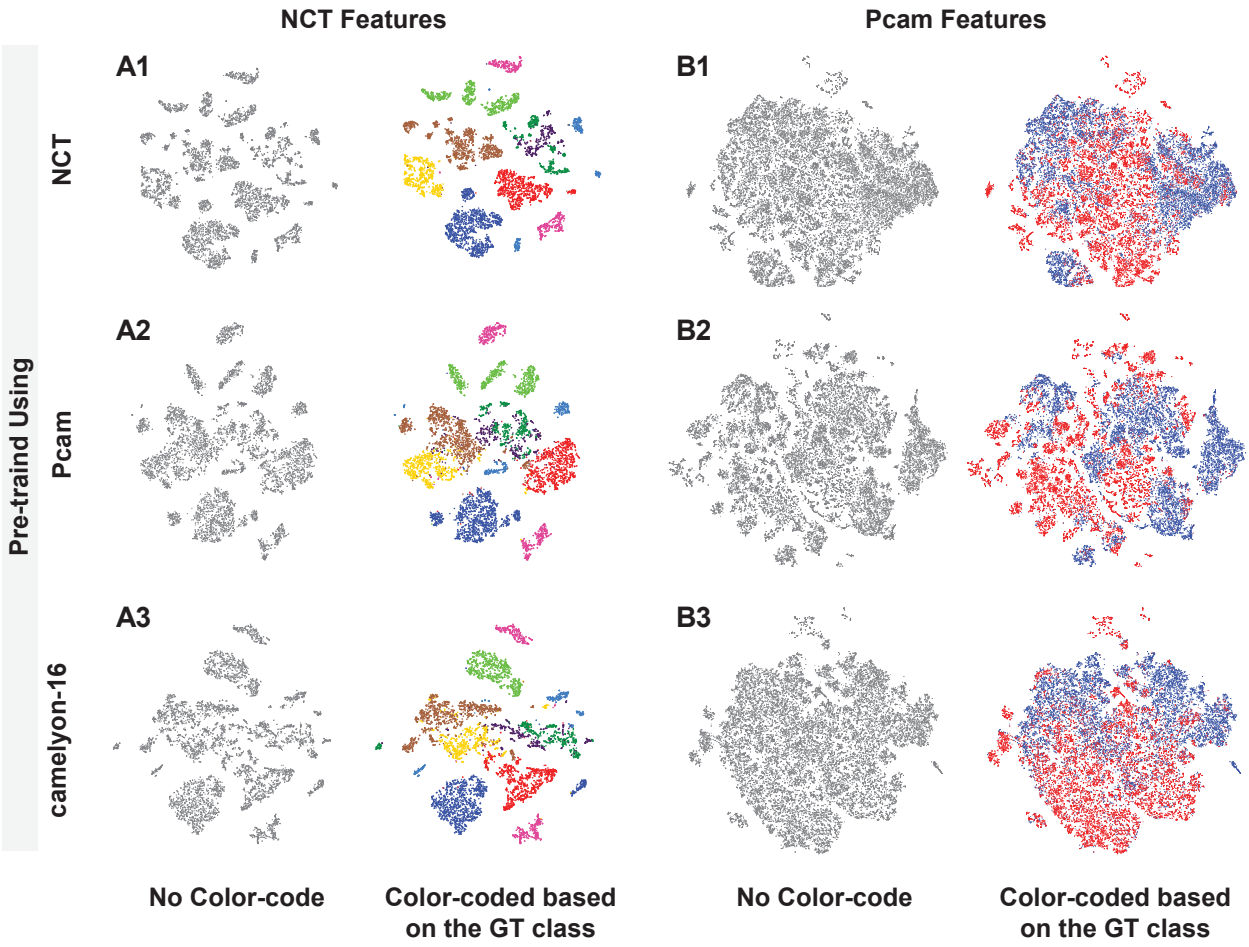


Fig. 6: Interpreting the pre-training process through T-SNE maps. The first two columns show T-SNE maps for the NCT100k test dataset's features after pre-training SimCLRv2 using different datasets. Note that the clusters are partially formed for in-domain as well as out-domain pre-training. The remaining two columns show T-SNE maps for the PCam test dataset's features from the same pre-trained SimCLRv2 models. Both non-color-coded and color-coded T-SNE map pairs are provided (based on ground truth labels mentioned in Fig. 4D). (A1) NCT 100K dataset features extracted from SimCLRv2 pre-trained on the NCT100K dataset itself. (A2) NCT 100K dataset features extracted from SimCLRv2 pre-trained on the Pcam dataset. (A3) NCT 100K dataset features extracted from SimCLRv2 pre-trained on the Camelyon-16 dataset. (B1) Pcam dataset features extracted from SimCLRv2 pre-trained on the NCT100K dataset. (B2) Pcam dataset features extracted from SimCLRv2 pre-trained on the Pcam dataset itself. (B3) Pcam dataset features extracted from SimCLRv2 pre-trained on the Camelyon-16 dataset.

Table 4: Whole slide classification results for CAMELYON16 dataset. The column "ImgNet weights" shows the models that used ResNet50 encoders trained on ImageNet in fully supervised settings. The column "Finetuned" shows the percentage of data from the same pre-trained dataset used to finetune our encoders. Results marked by $*$ are quoted from [52]

Method	ImgNet weights	Pretrained	Finetuned	Acc(%)	F1(%)	AUC(%)
Mean Pooling*	✓	-	-	62.6	35.5	52.8
Max Pooling*	✓	-	-	82.6	75.4	85.4
RNN-MIL* [53]	✓	-	-	84.4	79.8	87.5
Classic AB-MIL* [54]	✓	-	-	84.5	78.0	85.4
DS-MIL* [55]	✓	-	-	85.6	81.5	89.9
CLAM-SB* [56]	✓	-	-	83.7	77.5	87.1
CLAM-MB* [56]	✓	-	-	82.3	77.4	87.8
Trans-MIL* [57]	✓	-	-	85.8	79.7	90.6
DTFD-MIL* [52]	✓	-	-	89.9	86.6	**93.3**
DTFD-MIL + Our encoder	✗	PCAM	100%	90.7	87.5	**95.5**
DTFD-MIL + Our encoder	✗	NCT-CRC	100%	92.2	89.6	94.0
DTFD-MIL + Our encoder (UA)	✗	PCAM	100%	87.6	82.2	93.6
DTFD-MIL + Our encoder (UA)	✗	NCT-CRC	100%	92.2	89.1	94.4
DTFD-MIL + Our encoder	✗	PCAM	10%	86.8	81.9	89.7
DTFD-MIL + Our encoder	✗	NCT-CRC	10%	89.1	86.5	93.6
DTFD-MIL + Our encoder (UA)	✗	PCAM	10%	89.6	89.6	**95.9**
DTFD-MIL + Our encoder (UA)	✗	NCT-CRC	10%	92.2	89.6	94.3
DTFD-MIL + Our encoder	✗	PCAM	1%	85.2	80.0	88.4
DTFD-MIL + Our encoder	✗	NCT-CRC	1%	88.4	84.0	92.1
DTFD-MIL + Our encoder (UA)	✗	PCAM	1%	92.2	88.8	**95.3**
DTFD-MIL + Our encoder (UA)	✗	NCT-CRC	1%	95.9	94.9	92.7
DTFD-MIL + Our encoder	✗	Camelyon-16	-	91.5	88.8	**95.8**

Through a series of experiments, we showcase the performance of our framework across various histopathology tasks, even when confronted with limited annotated data. Furthermore, we address a critical limitation in histopathology models by incorporating uncertainty awareness, which has significantly enhanced the interpretability of our novel model. This newfound interpretability may empower clinicians to make well-informed decisions, facilitating the seamless integration of AI into the manual clinical decision-making process.

Our findings indicate that vision models with CNN backbones outperform transformer-based models in histopathology visual learning. Specifically, SimCLRv2, equipped with a simple ResNet-50 backbone, surpasses current SOTA models. It has the potential to achieve even better performance with a deeper ResNet backbone. The superiority of SimCLRv2 may be due to contrastive learning. As depicted in Fig. 6, the feature clusters formed by the SSL pre-trained model reveal its ability to differentiate distinct classes and their corresponding features, even before any annotations are introduced. The learning procedure of SimCLRv2, which effectively utilizes a large unlabelled dataset in two steps of the pipeline, confers the advantage of learning a highly effective encoder for patch-level data.

When a sufficient number of annotations are available for UA training, UA SimCLRv2 tends to outperform SimCLRv2. The transition from the feature space clusters of the SimCLRv2 model to those of UA SimCLRv2 results in more well-defined and tightly shaped clusters, indicating an improved ability to classify accurately (see Fig. 3). Moreover, the alignment of high uncertainty with incorrect predictions demonstrates the model's capability to identify challenging cases and exhibit lower confidence in predicting them.

Our results also suggest that when enough labels are available for fine-tuning, pre-training on the out-distribution data results in higher performance. The superiority of this out-domain trained model can be attributed to the advantage gained from a large number of classes available for contrastive pre-training. With an increased diversity of classes, the encoder can effectively compare and cluster features, leading to a better understanding of distinctive tissue characteristics and the establishment of clear boundaries within the feature space. Last, the seamless integration of our models into the WSI classification demonstrates the versatility and adaptability of our approach.

However, in scenarios with fewer annotations, such as our 1% setting, UA-SimCLRv2 exhibits significant

underperformance, often providing incorrect predictions with high confidence, particularly when multiple data classes are present. This poses a substantial risk in the context of medical AI, as the model's confidence does not necessarily reflect the correctness of its decisions. Thus we recommend using approaches such as uncertainty-aware fine-tuning (proposed in the section 3.3) to mitigate such risks when fine-tuning with few annotations. We also highly recommend rigorous testing with independent test datasets that aren't used during any stage of the training process including the hyper-parameter estimation.

In conclusion, our extensive patch- and slide-level experiments with contrastive learning and uncertainty quantification set new benchmarks in digital pathology classification. Our results consistently suggest the efficient use of large datasets -despite having no annotations- helps build better models for digital pathology. We believe that our work sets the foundation upon which multiple clinical tasks can use large digital pathology datasets efficiently, accurately, and interpretably.

ACKNOWLEDGEMENTS

This work was supported by the Center for Advanced Imaging and the John Harvard Distinguished Science Fellowship Program within the FAS Division of Science of Harvard University.

5. METHODOLOGY

5.1. UA SimCLRv2

To address the critical need for interpretability, we introduce UA-SimCLRv2. The primary objective of UA-SimCLRv2 is to enhance the interpretability of the model's predictions in the context of histopathology analysis. This is achieved by incorporating the theory of uncertainty estimation, which serves as the basis for uncertainty awareness in UA SimCLRv2.

In [42], the uncertainty estimation is approached from Dempster–Shafer theory of evidence (DST) perspective [59] assigining belief masses to subsets of a frame of discernment, which denotes the set of exclusive possible states. Subjective logic (SL) formalizes DST's notion of belief assignments over a frame of discernment as a Dirichlet distribution. Term evidence as 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 are collected and belief masses b_k ($k = 1, 2...K$) are assigned to each class based on the evidences collected and the remaining are marked as uncertainty u. For K mutually exclusive classes,

$$u + \sum_{k=1}^{K} b_k = 1 \tag{1}$$

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

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

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 [60]. It is characterized by K parameters $\alpha = [\alpha_1, \alpha_2...., \alpha_K]$ and is given as follows.

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

where $S_K = \{p\| \sum_{i=1}^{K} p_i = 1 \quad \text{and} \quad 0 \leq p_1, ...p_k \leq 1\}$ and $B(\alpha)$ is the K-dimensional multinomial beta function [61].

Model training follows the classical neural network architecture with 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 its mean $(\frac{\alpha_i}{S})$ is considered as the class probabilities. Let y_i be one hot vector encoding the ground-truth class label of sample x_i. Treating $D(p_i\|\alpha_i)$ as a prior on the sum of squares loss $\|y_i - p_i\|_2^2$, obtain the loss function

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

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 evidences collected might stregthen the belief for multiple classes. To avoid situations where evidence with more ambiguity assign 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.

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

where $\lambda_t = \min(1.0, t/10) \in [0, 1]$ is the annealing coefficient, t is the index of the current training epoch, $D(p_i\| < 1, ..., 1 >)$ is the uniform Dirichlet distribution, and $\alpha_i = y_i + (\tilde{1} - y_i) * \alpha_i$ 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

$$KL[D(p_i\|\tilde{\alpha_i})||D(p_i\|1)] = \log(\frac{\Gamma(\sum_{k=1}^{K}\tilde{\alpha_{ik}})}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha_{ik}})}) + \sum_{k=1}^{K}(\tilde{\alpha_{ik}} - 1)[\psi(\tilde{\alpha_{ik}}) - \psi(\sum_{j=1}^{K}\tilde{\alpha_{ij}})]$$

where 1 represents the parameter vector of K ones, Γ is the gamma function, and ψ 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.

5.2. DTFD-MIL framework for MIL

DTFD-MIL [52] uses a pseudo bag concept to virtually increase the number of bags and uses a doublt-tier approach for WSI classification. One WSI bag is randomly divided into multiple pseudo bags with relatively less number of patches. A pseudo bag is given the label of the parent bag. DTFD-MIL is applied on top of the pseudo bags to predict a bag label. It uses the commonly used attention based MIL approach in each tier. First, patch features are extracted from each pseudo bag using a ResNet backbone. These features are forwarded to the attention based tier-1 model which computes attention scores and instance probabilities for each patch. Tier-1 model aggregates patch features into an embedding that represents the pseudo bag. A feature distillation is performed on top of patch embeddings, using tier-1 instance probabilities, to extract a distilled feature vector. Distilled feature vectors from all pseudo bags are forwarded to the tier-2 attention based model, which aggregates them using attention to learn the final bag embedding for the parent bag. Bag labels from all of the tier-1 models and tier-2 model is compared with ground truth parent bag label to drive the cross entropy loss for training.

REFERENCES

[1] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[2] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework," *Sensors*, vol. 21, no. 3, p. 748, 2021.

[3] Y. Jiang, M. Yang, S. Wang, X. Li, and Y. Sun, "Emerging role of deep learning-based artificial intelligence in tumor pathology," *Cancer communications*, vol. 40, no. 4, pp. 154–166, 2020.

[4] J. Wang, C.-J. Wu, M.-L. Bao, J. Zhang, X.-N. Wang, and Y.-D. Zhang, "Machine learning-based analysis of mr radiomics can help to improve the diagnostic performance of pi-rads v2 in clinically relevant prostate cancer," *European radiology*, vol. 27, pp. 4082–4090, 2017.

[5] Z. Lu, S. Xu, W. Shao, Y. Wu, J. Zhang, Z. Han, Q. Feng, and K. Huang, "Deep-learning–based characterization of tumor-infiltrating lymphocytes in breast cancers from histopathology images and multiomics data," *JCO clinical cancer informatics*, vol. 4, pp. 480–490, 2020.

[6] R. Li, A. Shinde, A. Liu, S. Glaser, Y. Lyou, B. Yuh, J. Wong, and A. Amini, "Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival," *JCO Clinical Cancer Informatics*, vol. 4, pp. 637–646, 2020.

[7] E. Wulczyn, D. F. Steiner, Z. Xu, A. Sadhwani, H. Wang, I. Flament-Auvigne, C. H. Mermel, P.-H. C. Chen, Y. Liu, and M. C. Stumpe, "Deep learning-based survival prediction for multiple cancer types using histopathology images," *PloS one*, vol. 15, no. 6, p. e0233678, 2020.

[8] X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016, pp. 544–547.

[9] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, p. e1002730, 2019.

[10] M. Byra, K. Dobruch-Sobczak, Z. Klimonda, H. Piotrzkowska-Wroblewska, and J. Litniewski, "Early prediction of response to neoadjuvant chemotherapy in breast cancer sonography using siamese convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 797–805, 2020.

[11] L. Lu, L. Dercle, B. Zhao, and L. H. Schwartz, "Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging," *Nature communications*, vol. 12, no. 1, p. 6654, 2021.

[12] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.

[13] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[14] Z. Liu, L. Tong, L. Chen, Z. Jiang, F. Zhou, Q. Zhang, X. Zhang, Y. Jin, and H. Zhou, "Deep learning based brain tumor segmentation: a survey," *Complex & intelligent systems*, vol. 9, no. 1, pp. 1001–1026, 2023.

[15] H. H. Rashidi, N. K. Tran, E. V. Betts, L. P. Howell, and R. Green, "Artificial intelligence and machine learning in pathology: the present landscape of supervised methods," *Academic pathology*, vol. 6, p. 2374289519873088, 2019.

[16] T. Sakamoto, T. Furukawa, K. Lami, H. H. N. Pham, W. Uegami, K. Kuroda, M. Kawai, H. Sakanashi, L. A. D. Cooper, A. Bychkov *et al.*, "A narrative review of digital pathology and artificial intelligence: focusing on lung cancer," *Translational Lung Cancer Research*, vol. 9, no. 5, p. 2255, 2020.

[17] J. H. Harrison Jr, J. R. Gilbertson, M. G. Hanna, N. H. Olson, J. N. Seheult, J. M. Sorace, and M. N. Stram, "Introduction to artificial intelligence and machine learning for pathology," *Archives of pathology & laboratory medicine*, vol. 145, no. 10, pp. 1228–1254, 2021.

[18] H. R. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: challenges and opportunities," *Journal of pathology informatics*, vol. 9, no. 1, p. 38, 2018.

[19] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[20] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. Van De Vijver, R. B. West, M. Van De Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science translational medicine*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011.

[21] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature communications*, vol. 7, no. 1, p. 12474, 2016.

[22] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, E. I. Chang *et al.*, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–17, 2017.

[23] P. Bachman, R. D. Hjelm, and W. Buchwalter, *Learning Representations by Maximizing Mutual Information across Views*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[24] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706–6716, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208617491

[25] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[26] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9912–9924. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf

[27] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 271–21 284. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf

[28] X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 15 750–15 758. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html

[29] S. Gidaris, A. Bursuc, G. Puy, N. Komodakis, M. Cord, and P. Pérez, "Learning representations by predicting bags of visual words," in *CVPR*, 2021.

[30] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 12 310–12 320. [Online]. Available: http://proceedings.mlr.press/v139/zbontar21a.html

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[32] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[33] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[35] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, p. 100198, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827021000992

[36] Y. Luo, Z. Chen, and X. Gao, "Self-distillation augmented masked autoencoders for histopathological image classification," *arXiv preprint arXiv:2203.16983*, 2022.

[37] S. Wang and C. Manning, "Fast dropout training," in *international conference on machine learning*. PMLR, 2013, pp. 118–126.

[38] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[39] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[40] F. Wenzel, J. Snoek, D. Tran, and R. Jenatton, "Hyperparameter ensembles for robustness and uncertainty quantification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6514–6527, 2020.

[41] J. M. Dolezal, A. Srisuwananukorn, D. Karpeyev, S. Ramesh, S. Kochanny, B. Cody, A. S. Mansfield, S. Rakshit, R. Bansal, M. C. Bois *et al.*, "Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology," *Nature communications*, vol. 13, no. 1, p. 6572, 2022.

[42] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[43] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 22 243–22 255. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf

[44] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant CNNs for digital pathology," Jun. 2018.

[45] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, , and the CAMELYON16 Consortium, "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 12 2017. [Online]. Available: https://doi.org/10.1001/jama.2017.14585

[46] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, and X. 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*. Springer, 2021, pp. 186–195.

[47] X. Chen*, S. Xie*, and K. He, "An empirical study of training self-supervised vision transformers," *arXiv preprint arXiv:2104.02057*, 2021.

[48] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[49] X. Jin, T. Huang, K. Wen, M. Chi, and H. An, "Histossl: Self-supervised representation learning for classifying histopathology images," *Mathematics*, vol. 11, no. 1, p. 110, 2022.

[50] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[51] J. N. Kather, N. Halama, and A. Marx, "100,000 histological images of human colorectal cancer and healthy tissue," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1214456

[52] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 802–18 812.

[53] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[54] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[55] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 318–14 328.

[56] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[57] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5206848/

[59] A. P. Dempster *et al.*, "Upper and lower probabilities induced by a multivalued mapping." *Classic works of the Dempster-Shafer theory of belief functions*, vol. 219, no. 2, pp. 57–72, 2008.

[60] A. Jøsang, "Generalising bayes' theorem in subjective logic." in *MFI*, 2016, pp. 462–469.

[61] S. Kotz, N. Balakrishnan, and N. Johnson, "Continuous multivariate distributions–vol. 1, john wiley & sons, new york," 2000.

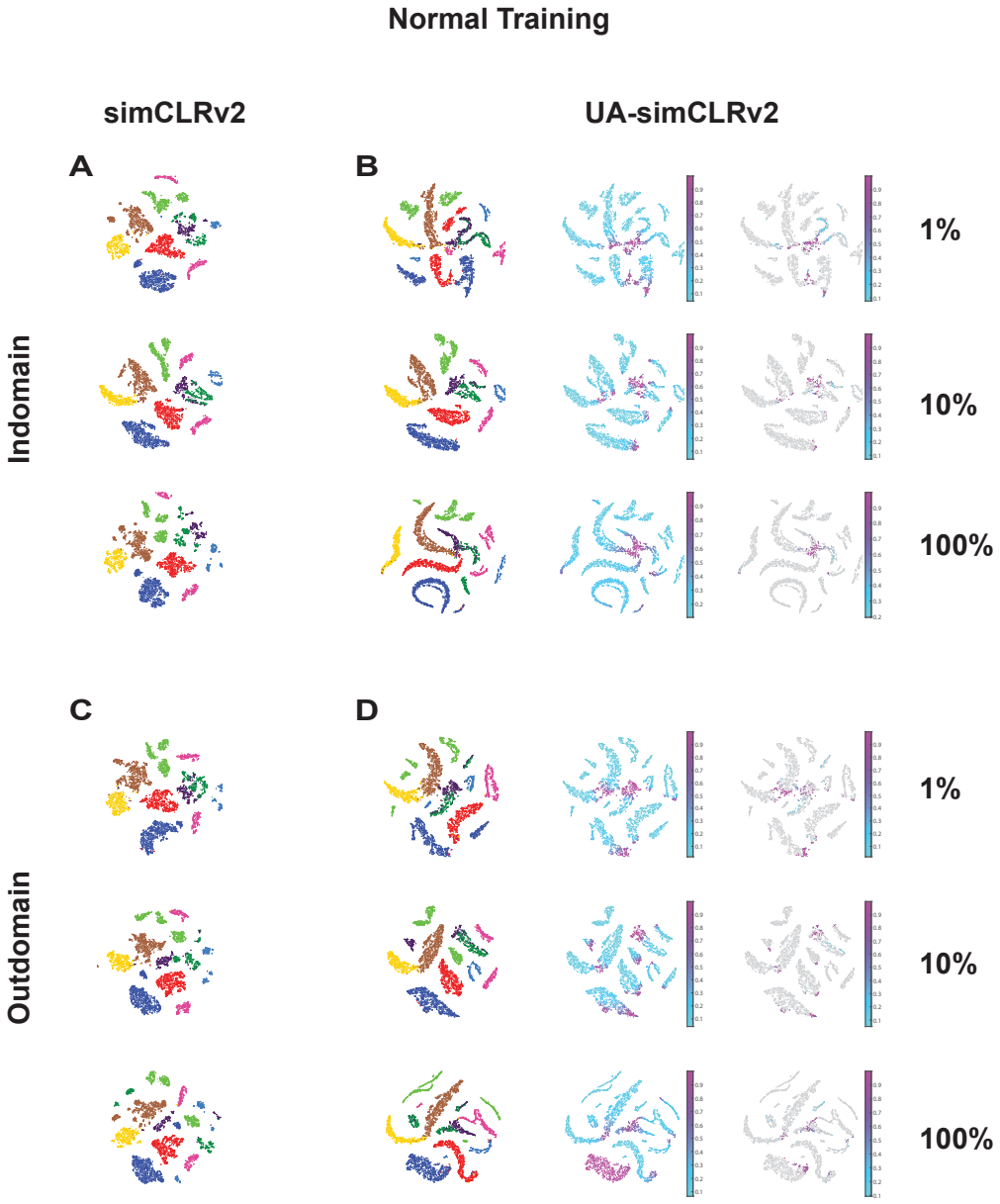


Fig. S1: Variation of NCT100K dataset feature space learnt by SimCLRv2 and UA-SimCLRv2 based on a percentage of annotations and the pre-training distribution. In the "Indomain" setting, the model undergoes pre-training and fine-tuning on the NCT100k. In the "Outdomain" setting, the model undergoes pre-training on the PCam dataset and fine-tuning on the NCT100K. The T-SNE maps on the first two columns are colour-coded based on NCT100K classes as shown in Fig.4D. The third column has T-SNE maps color-coded based on their uncertainty scores and the fourth column T-SNE maps has only the misclassified points highlighted. (A) and (B) showcase T-SNE maps for features extracted from SimCLRv2 and UA-SimCLRv2 models respectively in the "Indomain" setting. Similarly, (C) and (D) showcase T-SNE maps for features extracted from SimCLRv2 and UA-SimCLRv2 models respectively in the "Outdomain" setting.

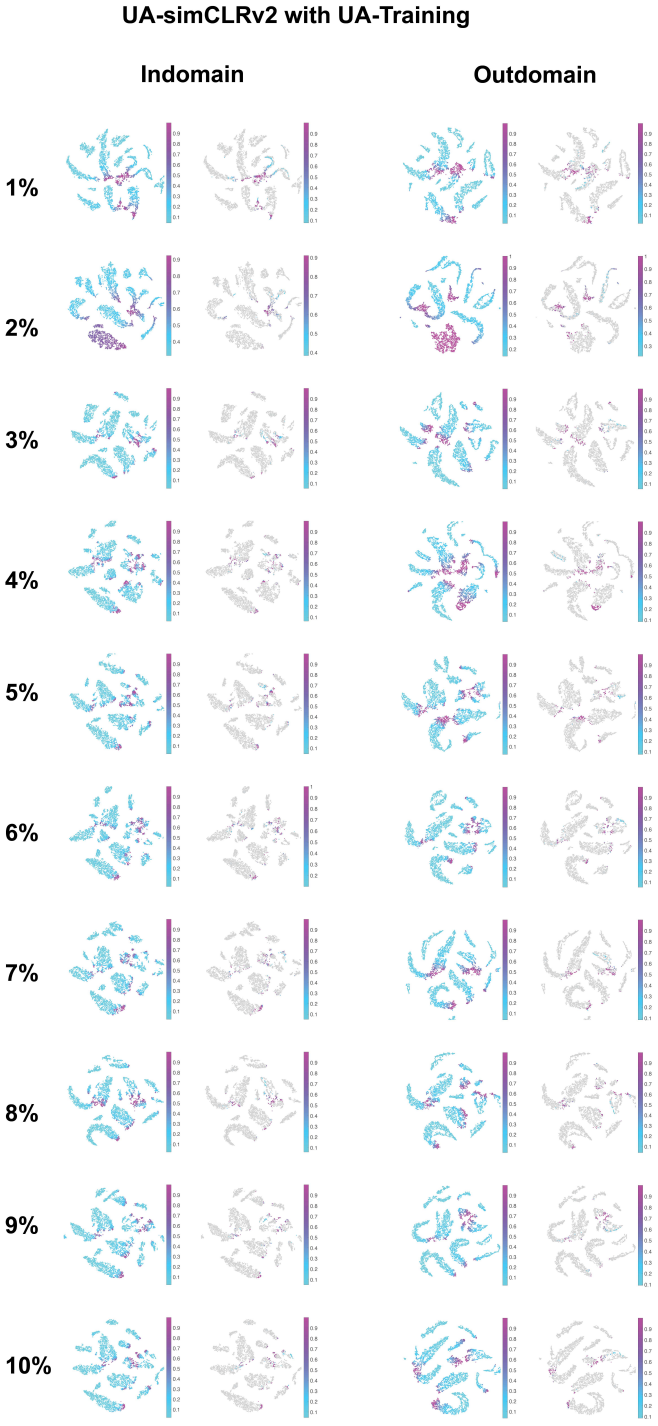


Fig. S2: Variation of NCT100K dataset feature space learnt by UA-SimCLRv2 in selective labeling approach based on a percentage of annotations and the pre-training distribution. In the "Indomain" setting, the UA-SimCLRv2 model undergoes pre-training, fine-tuning and UA-Training on the NCT100k dataset. In the "Outdomain" setting, the UA-SimCLRv2 model undergoes pre-training on the PCam dataset and fine-tuning and UA-Training on the NCT100K dataset. We present paired T-SNE maps for each scenario. The first map is color-coded by the uncertainty score whereas the second map in each pair highlights only the misclassified instances.