Are We Ready for Out-of-Distribution Detection in Digital Pathology?

Ji-Hun Oh[1], Kianoush Falahkheirkhah[1,2], and Rohit Bhargava[1]

[1] University of Illinois Urbana-Champaign, Urbana, IL, USA
[2] CZ Biohub Chicago, LLC, Chicago, IL, USA
{jihunoh2, kf4, rxb}@illinois.edu

Abstract. The detection of semantic and covariate out-of-distribution (OOD) examples is a critical yet overlooked challenge in digital pathology (DP). Recently, substantial insight and methods on OOD detection were presented by the ML community, but how do they fare in DP applications? To this end, we establish a benchmark study, our highlights being: 1) the adoption of proper evaluation protocols, 2) the comparison of diverse detectors in both a single and multi-model setting, and 3) the exploration into advanced ML settings like transfer learning (ImageNet *vs.* DP pre-training) and choice of architecture (CNNs *vs.* transformers). Through our comprehensive experiments, we contribute new insights and guidelines, paving the way for future research and discussion. We continuously update our code at https://github.com/jihunoh2/OODD4DP.

Keywords: Digital pathology · Out-of-distribution detection · Misclassified detection · Robustness · Transfer learning

1 Introduction

The fickleness and fragility of deep neural networks (DNNs) makes them prone to overconfident but erroneous predictions, particularly under distribution shifts. In high-stake domains like digital pathology (DP), a subsequent misdiagnosis can be catastrophic, thus far hindering real-world DNN deployments. To facilitate trustworthy AI practices in DP, it is pivotal for DNNs to communicate "*I don't know*" when unsure of its own prediction, allowing clinicians to intervene.

Background. Known as an *out-of-distribution (OOD) detection* problem, let us denote $f : \mathcal{X} \to \mathcal{Y}$ the classifier model, the focus of our article. Letting $P(\mathcal{X}_{ID}, \mathcal{Y}_{ID}) \subseteq \mathcal{X} \times \mathcal{Y}$ the joint in-distribution (ID) defined by the training set, a sample is OOD if $(x, y) \notin P(\mathcal{X}_{ID}, \mathcal{Y}_{ID})$. In literature, OOD is categorized into either *semantic* or *covariate*; semantic OOD arise from label-altering shifts $y \notin P(Y|X)$, whereas covariate OOD preserves the ID labels $y \in P(Y|X)$ but are modified by the image space $x \notin P(X)$. Note, $y \notin P(Y|X)$ generally entails $x \notin P(X)$, and thus, semantic OOD is shifted by the image space too. Since the model cannot inherently handle semantic OOD, we seek to flag them using some scoring function expressing OOD extent; examples include (*but is not restricted*

to) the softmax uncertainty. However, in the case of covariate OOD, unconditional detection is undesirable as it conflicts with *generalization*,[1] limiting its open-world application. To avoid such a dilemma, the emerging consensus in the ML community is to detect just its failures (*i.e.*, misclassified instances) [6, 13]. From here onwards, we refer to semantic and misclassified covariate OOD detection as S-OODD and MC-OODD, respectively.

In the field of DP, S-OODD is required foremost because models are conceived in a closed-set environment. For instance, many breast datasets do not include borderline atypical lesion or rarer carcinoma subtypes due to its rarity and/or annotating difficulties. Likewise, the need for MC-OODD is inevitable because DNNs are prone to overfit or memorize, even finding shortcuts by picking up spurious correlations (*e.g.*, site-specific attributes coming from the staining and scanning procedures) in lieu of pathological generalization.

Related work. Few work have studied detection tasks in DP [2, 19, 29]. Ref. [29] inspected MC-OODD (across different organ & hospital) and S-OODD (head/neck SCC) over the Camelyon[2] dataset. Ref. [2] benchmarked various OOD detectors in PatchCamelyon, designating external datasets as OOD. Recently, [19] investigated the detection of diffuse large B-cell lymphoma *w.r.t.* Camelyon, and prostate images containing colorectal sections *w.r.t.* colon-free prostate biopsies. We however recognize several deficiencies in these work.

i) Misleading practice. Not all detection objectives therein conform to the above-mentioned consensus. For instance, let us consider the example of detecting colorectal invasion in prostate biopsies [19], which arises due to their anatomical proximity. Such a shift from co-occurrence does not necessarily induce a label change and is closer to a covariate OOD, thus, calling for MC-OODD. Note, it is possible for certain cases to potentially alter the patient's treatment/prognosis and become semantic OOD; in these scenarios, their detection makes sense, but such a distinction is not made in the above work. In addition, the goal of MC-OODD is to quantify how separable misclassifications are from its correct counterpart, *regardless of their ratios*. However, using AUROC or AUPR like in [29] yields a systemic bias as these metrics are sensitive to the model's accuracy [23, 24], preventing a fair study when comparing factors affecting OOD detection performance but also simultaneously the model's base performance *e.g.*, choice of DNN architecture.

ii) Limited OOD detectors. Many works [19, 29] adopt multi-model uncertainty quantification (UQ) to score OOD-ness, *e.g.*, ensembles or approximate Bayes. However, these uncertainty measures like Shannon Entropy and Mutual Information can be falsely low far away from the ID data [10, 30]. While [2] explored diverse detectors beyond those UQ-based, all are from 2020 or prior. More recent SOTA methods such as ViM [31] have yet to be explored in DP.

iii) Easy or non-public datasets. The datasets used are simple such as the binary classification task in Camelyon. In addition, some "OODs" therein are *far*

[1] This is because covariate OODs share the ID labels, resulting in robust models capable of generalizing across such out-of-domains.

[2] https://camelyon17.grand-challenge.org/Home/

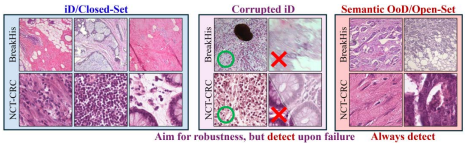

Fig. 1: **Our experiments.** We aim to i) generalize over corrupted (covariate-shifted) ID or detect its misclassification, and ii) always detect semantic OOD.

(or *very-far*) *w.r.t.* ID, making them easy to discern, *e.g.*, prostate *vs.* colon *vs.* breast lymph nodes [19], and H&E *vs.* other dyes [2]. Oftentimes, the work is also not reproducible because internal datasets were used [19, 29].

iv) Limited depth. They overlook crucial factors known to influence robustness and OOD behavior like pre-training [9] and DNN architecture [24].

Contributions. Acknowledging these shortcomings, we herein present an enhanced benchmark study. Concretely, we contribute in the following ways.

i) Proper protocols. Adopting public datasets, we simulate an Open Set Recognition (OSR) setting by excluding a small fraction of the class during training and afterwards perform S-OODD *w.r.t.* the held-out classes, permitting S-OODD evaluation in an indisputable way. As for MC-OODD, we apply common DP corruptions in [36] to the ID test set and report the Prediction Rejection Ratio (PRR) [23], a metric agnostic to the model accuracy. Our experiments (illustrated in Fig. 1) are free from any of the above-mentioned bias/malpractice or shortcomings, providing thus the most objective and reproducible benchmark to date on OOD detection in the context of DP.

ii) Wider scope. In addition to UQ scores, we include a variety of recent frequentist (*i.e.*, single-model) detection methods from top-tier ML conferences. We also investigate the impact of transfer learning (TL), specifically when adopting pre-trained DNN weights conceived from natural and DP images. Last but not least, we comparatively explore different architectures, namely fully-convnets (CNNs) *vs.* transformers, the latter having seen a surge in popularity lately.

iii) Novel insights. Through our extensive experiments, we answer questions like: Is there a detector best suited for S-OODD and/or MC-OODD? Does ensemble-based UQ (widely regarded the gold-standard) really perform best? Should we always pre-train over DP data? Are transformers better than CNNs like recent studies suggest [1]? Whereas these are questions are popular subjects in the ML OOD detection literature, they have yet to be recognized and studied in the broader DP community as OOD detection has just recently started to gain traction. Our findings serve as a guide for practitioners and open up research questions and discussions for the future.

2 Benchmark Setup

Our overall pipeline involves training a classifier using an ID train set, and afterwards evaluating S-OODD and MC-OOD by respectively performing the binary

classification of *semantic OOD vs. ID test set* and *correctly classified covariate OOD vs. misclassified covariate OOD* using some OOD detection method.

Datasets. We consider two public H&E image datasets: **BreakHis** [26], consisting of eight breast carcinoma subtypes, and **NCT-CRC** [15], including nine colorectal tissue types. We simulate OSR splits in Tab. 1. Overall, the semantic OOD includes both *near (hard)* and *far (easy)* cases. For instance, the fine-grained MUS in NCT-CRC is hard to detect *w.r.t.* STR (and vice versa) due to their shared fibrous structure. In contrast, we anticipate ADI, BACK (NCT-CRC) or M-LC (BreakHis) to be much easier given their drastically different appearances. For BreakHis, we perform a 4:1 train-test split among the ID samples, while using the CRC-VAL as the test set for NCT-CRC. To generate covariate OOD, we apply the DP corruptions from [36] with a severity level of 3, which includes several types of digitization, blur, color, and artifact.

Training details. All classifiers are trained using the focal loss. We crop images at 224×224 and apply random flips and transpose. We allocate a small hold-out subset for validation purposes, evaluating its accuracy at every 25 iterations. Using a mini-batch size of 32, we initialize the Adam learning rate to 1e-4 and downstep by 0.3 upon reaching the [2^{nd}, 4^{th}, 5^{th}, 6^{th}] times the validation set's performance fail to improve, terminating the training afterwards.

ML configs. It is well-established that TL by initializing from pre-trained weights of another task improves robustness in downstream applications [9]. The de-facto protocol in computer vision is to fine-tune from ImageNet-1K (IN1K) weights. However, it is common belief that TL is more beneficial when the pre- and fine-tuning sources are similar in domain to capitalize feature reuse. Multiple recent works have thus proposed general-purpose DP models for TL, developed via non-supervised learning over massive amounts of unlabeled data [12, 14, 25, 35]. We delve into both TL options across two DNN backbones: **(C1) ResNet-50:** We compare training from scratch (*i.e.*, no TL; de-novo) and TL from IN1K and three DP-specific models in [14], each embodying a distinct self-supervised learning (SSL) method trained with 19M H&E patches extracted from The Cancer Genome Atlas (TCGA): MoCo v2 [4], SwAV [3], and BT [33]. **(C2) ViT-B/16** [5]: We compare de-novo, IN1K, and three foundational large vision-language models (LVLMs): CLIP [25], BiomedCLIP [35], and QuiltNet [12], in order of decreasing domain gap *w.r.t.* our downstream datasets.

Table 1: **OSR configs.** I and O denotes ID (closed-set) and semantic OOD (open-set), respectively. The four/three-fold OSR split is arbitrary, aimed to reduce bias in our analysis. †B, Benign; A, Adenosis; F, Fibroadenoma; PT, Phyllodes Tumor; TA, Tubular Adenoma; M, Malignant; DC, Ductal Carcinoma (C); LC, Lobular C; MC, Mucinous C; PC, Papillary C. ‡ADI, adipose; BACK, background; DEB, debris; LYM, lymphocytes; MUC, mucus; MUS, muscle; NORM, normal mucosa; STR, C-associated stroma; TUM, adenocarcinoma epithelium.

	BreakHis† [26]									NCT-CRC‡ [15]								
OSR	B-A	B-F	B-TA	B-PT	M-DC	M-LC	M-MC	M-PC		ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM
no. 1	O	I	I	I	O	I	I	I		O	I	I	O	I	I	O	I	I
no. 2	I	O	I	I	I	O	I	I		I	O	I	I	O	I	I	O	I
no. 3	I	I	O	I	I	I	O	I		I	I	O	I	I	O	I	I	O
no. 4	I	I	I	O	I	I	I	O										

Table 2: **Overview of frequentist methods.** Feat. refers to the classifier's penultimate activation map. We follow the original paper's guidelines to set hyperparameters where c and d is the class number and the feat's dimension.

Detector	Space	ID data free?	Working principle	Hyperparameter
MSP (ICLR'17 [8])	Prob.	Y	Max softmax prob.	n/a
Maha (NeurIPS'18 [18])	Feat.	N	Min feat. dist. to train set's classwise centroids	n/a
R+E (NeurIPS'21 [27])	Feat./Logit	N	Feat. truncation followed by energy function	Truncate by ID's 98%.
GrN (NeurIPS'21 [11])	Prob.	Y	Backpropagated gradients	n/a
MLS (ICML'21 [7])	Logit	Y	Max logit	n/a
KLM (ICML'21 [7])	Prob.	N	Min KL div. *w.r.t.* the train set's class-wise probs.	n/a
KNN (ICML'22 [28])	Feat.	N	k-th nearest neighbor dist. *w.r.t.* train set's feats.	$k \approx 2.5 \times c$
ViM (CVPR'22 [31])	Feat./Logit	N	Norm of feat's residual projection, plus energy	Subspace dim. $\approx d/2$
GEN (CVPR'23 [20])	Prob.	Y	Generalized entropy of prob.	$\gamma = 0.1$; no suppression

CLIP is trained with 300M natural image-caption pairs, BiomedCLIP incorporates 15M pairs from PubMed but does not include H&E, whereas QuiltNet utilizes 1M pairs of H&E-text pairs scraped from MedTwitter and YouTube. We take the IN1K and CLIP pre-trained weights from the public timm-repository.[3]

DNN architecture is also known to influence OOD behavior. We further inspect **(C3) different architectures**, notably CNNs (ResNet, ConvNeXt [22]) *vs.* transformers (ViT, Swin [21]). For a fair comparison, it is crucial to employ the same TL recipe to minimize its influence, but is not possible with existing pre-trained checkpoints which are limited to just a single architecture backbone. Hence, we pre-train with the 32 class pan-cancer TCGA dataset using the patch-level annotations in [16], adopting a training recipe similar to the above. Note, unlike (C1) & (C2), this pre-training is *fully-supervised*, developed in a much *light-weight* manner with ∼272K patches, trainable in a single commercial GPU in less than a few hours, presenting hence a unique perspective on DP-specific TL. We dub our checkpoints as SIAYN (short for *Supervision Is All You Need*) and make them public along with our code.

Detectors. In recent years, a plethora of OOD detection methods were proposed. We focus on **post-hoc/frequentist** (*i.e.*, single-model inference-time) methods which are competitive and (almost) cost-free, summarized in Tab. 2: Max Softmax Probability (MSP) [8], Mahalanobis distance (Maha) [18], ReAct followed by Energy (R+E) [27], GradNorm (GrN) [11], Max-Logit Score (MLS) [7], KL-Matching (KLM) [7], k^{th} Nearest Neighbor (KNN) [28], Virtual-Logit Matching (ViM) [31], and Generalized Entropy (GEN) [20]. They are chosen on the basis of popularity, competitiveness, and diversity in principle, in which information are leveraged from the feature, logit, and/or softmax probability spaces to handcraft a granular OOD scoring function beyond conventional uncertainty measures. Note, some methods are more prohibitive from needing access to ID training samples, which may be confidential, or embedding that can be inaccessible in black-box models; both are pertinent in the context of DP.

As hinted, **UQ** orthogonally provides a means to detect OOD as well. Hence, we inspect Deep Ensembles (DE) [17], the UQ gold-standard, using a heterogeneous setup comprised of four members and evaluate two uncertainty measures:

[3] Timm names: `resnet50.a1_in1k`, `vit_base_patch16_224.augreg2_in21k_ft_in1k`, `vit_base_patch16_clip_224.openai`.

Total Uncertainty (TU), the Shannon entropy of the DE-averaged probability, and Epistemic Uncertainty (EU), the mutual information across the members' probabilities. Unlike the post-hoc schemes, DE incurs overhead from training multiple models and performing multiple forward-passes during inference.

Metrics. For S-OODD, we report the AUROC (↑; arrow indicates which direction is better). For MC-OODD, we report the PPR (↑) [23], where closer to 100% indicates correlation of low confidence to mispredictions; conversely, zero/negative values mean no/anti-correlation and is thus undesirable. We refer to [23] for full calculation details. We also report the class-wise macro-averaged accuracy (Acc., ↑) of the ID test and covariate OOD sets to present a full picture on robustness (*e.g.*, a model with trivial performance can display near zero accuracy but high PRR). Note, when using the DE as the detector, we report the accuracy over the DE-averaged predictions.

Table 3: **(C1)** Results of various TL configs in ResNet-50.

Pre-train (TL)	Single-Model Acc%	MSP [8]	Maha [18]	R+E [27]	GrN [11]	MLS [7]	KLM [7]	KNN [28]	ViM [31]	GEN [20]	DE [17] (×4) Acc%	TU	EU
(S1) BreakHis: ID Acc% & S-OODD AUROC%													
None	73.31	57.85	61.58	57.39	54.11	56.89	58.4	70.15	62.33	56.47	74.04	59.24	61.83
IN1K	43.82	64.86	48.94	62.14	64.63	65.62	53.58	64.14	48.60	65.68	47.75	67.19	49.63
MoCo v2 [4,14]	95.42	72.22	82.53	74.40	73.15	73.69	70.72	74.62	81.94	75.31	95.88	77.30	77.60
SwAV [3,14]	**96.61**	**79.40**	**83.44**	**82.01**	**78.16**	**81.99**	**76.74**	**75.35**	**86.37**	**82.95**	**97.02**	**83.86**	77.09
BT [14,33]	94.89	72.81	80.79	76.25	66.83	76.22	73.09	74.23	81.05	75.98	95.73	76.89	**78.76**
(S2) NCT-CRC: ID Acc% & S-OODD AUROC%													
None	78.92	56.64	59.14	59.59	55.29	56.86	66.33	69.01	64.92	56.96	79.96	57.84	62.38
IN1K	93.83	**84.72**	76.45	84.99	70.77	**84.86**	**83.86**	87.25	84.38	83.16	94.62	**87.47**	76.31
MoCo v2 [4,14]	95.19	78.37	82.61	78.88	**79.87**	78.57	76.59	79.88	85.20	79.27	95.69	83.07	81.72
SwAV [3,14]	95.78	80.86	87.52	86.39	72.96	82.43	82.08	86.98	88.59	82.80	96.28	84.44	78.62
BT [14,33]	**96.18**	81.82	**87.79**	**86.64**	77.06	84.80	82.86	**92.08**	**91.44**	**85.21**	**96.81**	87.30	**84.19**
(S3) BreakHis: Covariate OOD Acc% & MC-OODD PRR%													
None	45.64	18.84	1.33	11.67	11.29	12.54	9.22	9.41	3.04	11.37	46.11	17.29	12.80
IN1K	28.91	21.53	-8.61	12.55	19.51	21.59	4.02	16.17	-7.36	20.72	30.47	22.43	-9.55
MoCo v2 [4,14]	**58.41**	21.12	33.34	16.23	17.88	18.26	21.78	22.40	34.44	18.81	**59.91**	25.39	30.16
SwAV [3,14]	57.30	**34.05**	33.41	**31.41**	**34.19**	**33.04**	**34.81**	33.66	**36.56**	**34.30**	57.95	**37.41**	19.38
BT [14,33]	58.23	27.79	**34.66**	27.62	28.27	27.18	27.25	**37.14**	35.55	27.68	59.43	34.11	**35.47**
(S4) NCT-CRC: Covariate OOD Acc% & MC-OODD PRR%													
None	53.06	15.85	11.34	7.25	5.24	7.44	16.00	25.00	13.32	7.46	53.30	15.88	17.12
IN1K	73.47	50.40	29.11	45.10	36.47	48.76	47.58	46.66	36.51	45.18	75.07	50.37	25.58
MoCo v2 [4,14]	80.19	49.29	42.60	46.87	**49.32**	48.11	45.95	48.84	48.11	48.18	82.32	53.22	49.26
SwAV [3,14]	**81.75**	**61.68**	**51.62**	**59.34**	48.45	**60.18**	**61.80**	**60.72**	53.20	**58.35**	83.31	**62.62**	50.42
BT [14,33]	81.63	58.01	49.36	52.80	40.30	53.95	55.97	55.47	**54.46**	53.83	**83.56**	62.00	**55.19**

Table 4: **(C2)** Results of various TL configs in ViT-B/16.

Pre-train (TL)	Single-Model Acc%	MSP [8]	Maha [18]	R+E [27]	GrN [11]	MLS [7]	KLM [7]	KNN [28]	ViM [31]	GEN [20]	DE [17] (×4) Acc%	TU	EU
(S1) BreakHis: ID Acc% & S-OODD AUROC%													
None	43.28	53.26	58.92	54.81	53.89	54.01	55.33	63.37	60.49	54.70	47.75	54.47	57.87
IN1K	95.06	68.58	77.43	73.85	67.27	73.70	68.26	72.22	78.98	73.57	**96.60**	75.06	76.04
CLIP [25]	56.93	55.96	63.97	59.35	56.33	58.57	54.80	60.90	65.09	58.63	69.45	58.88	54.18
BiomedCLIP [35]	**95.49**	**73.71**	**77.74**	**78.25**	**74.42**	**78.51**	**74.16**	**76.18**	**80.81**	**78.51**	**96.60**	**79.17**	**80.68**
QuiltNet [12]	60.16	56.03	64.17	60.25	59.40	59.36	54.02	62.08	66.13	60.07	67.51	61.19	60.12
(S2) NCT-CRC: ID Acc% & S-OODD AUROC%													
None	76.85	63.63	67.55	62.24	64.09	63.20	66.95	68.35	69.98	63.50	78.00	66.36	65.83
IN1K	**97.21**	**86.03**	**84.20**	**87.86**	82.30	**87.46**	**86.62**	86.21	**87.24**	**87.98**	**97.84**	**90.73**	**87.38**
CLIP [25]	93.91	71.26	78.77	72.76	72.93	73.02	74.48	81.95	79.84	73.03	96.33	75.61	73.2
BiomedCLIP [35]	96.49	83.95	74.23	86.78	**82.94**	86.52	84.99	**86.48**	83.36	86.89	97.04	87.92	85.76
QuiltNet [12]	95.02	74.38	81.73	72.05	73.43	73.09	75.75	83.18	82.29	73.94	96.44	78.79	77.80
(S3) BreakHis: Covariate OOD Acc% & MC-OODD PRR%													
None	37.06	20.50	6.05	14.46	15.70	17.04	2.52	11.93	10.27	14.84	39.65	20.18	9.73
IN1K	67.37	41.77	18.16	31.31	32.80	34.48	34.96	22.51	21.49	35.74	70.47	45.05	33.34
CLIP [25]	41.19	19.37	4.92	16.56	15.87	18.27	2.43	7.31	7.74	17.23	48.46	20.65	8.88
BiomedCLIP [35]	**70.07**	**48.61**	**26.45**	**39.81**	**42.83**	**43.14**	**45.60**	**34.71**	**32.95**	**44.09**	**73.14**	**52.83**	**42.77**
QuiltNet [12]	44.06	20.69	5.26	17.50	17.64	19.98	1.30	11.42	9.74	18.39	48.89	20.81	9.96
(S4) NCT-CRC: Covariate OOD Acc% & MC-OODD PRR%													
None	63.87	40.89	8.10	16.19	33.64	23.30	29.48	24.13	12.23	22.08	65.04	37.35	25.27
IN1K	**90.13**	**75.09**	**46.39**	**65.11**	**67.61**	**69.29**	**72.90**	**57.45**	**51.61**	**68.66**	**91.70**	**74.92**	**62.14**
CLIP [25]	75.15	45.88	33.14	36.35	40.58	40.48	41.99	42.18	35.99	40.05	78.29	49.63	36.70
BiomedCLIP [35]	87.03	68.05	27.74	58.42	62.38	62.00	65.39	56.36	43.15	62.14	88.76	68.03	58.38
QuiltNet [12]	77.17	49.40	43.37	39.91	42.37	44.75	46.18	48.04	44.20	43.68	79.99	53.54	46.49

Table 5: **(C3)** Results of diff. architectures using our SIAYN TL. We chose size variants of ConvNeXt, ViT, Swin to fit in 24GB of GPU.

Pre-train (TL)	Single-Model Acc%	MSP [8]	Maha [18]	R+E [27]	GrN [11]	MLS [7]	KLM [7]	KNN [28]	ViM [31]	GEN [20]	DE [17] (×4) Acc%	TU	EU
(S1) BreakHis: ID Acc% & S-OODD AUROC%													
ResNet-50	93.98	76.25	79.15	79.11	74.41	79.04	73.43	74.47	83.33	79.29	94.43	80.22	73.70
ConvNeXt-S [22]	**96.47**	**78.38**	83.64	**81.71**	**78.31**	**82.02**	**78.43**	74.63	85.13	**82.60**	97.40	**82.75**	**82.84**
ViT-B/16 [5]	96.22	77.20	80.50	80.59	71.71	80.62	74.86	72.66	82.13	80.80	96.94	81.55	80.62
Swin-T [21]	96.37	74.75	**85.28**	79.35	71.48	79.27	74.27	**75.40**	**87.05**	79.71	**97.41**	80.07	81.10
(S2) NCT-CRC: ID Acc% & S-OODD AUROC%													
ResNet-50	95.37	82.51	78.08	84.49	72.92	83.78	82.20	84.03	82.68	82.47	95.95	85.04	79.45
ConvNeXt-S [22]	96.04	**86.20**	**92.76**	**87.65**	**86.03**	**87.52**	85.72	**92.74**	**92.84**	**88.38**	97.06	**89.42**	85.94
ViT-B/16 [5]	**96.79**	85.64	92.08	86.73	80.07	86.58	**86.69**	90.79	92.69	86.96	**97.62**	88.77	**88.09**
Swin-T [21]	95.82	82.07	91.45	84.96	77.60	84.57	81.91	89.48	91.95	84.89	96.77	86.56	86.20
(S3) BreakHis: Covariate OOD Acc% & MC-OODD PRR%													
ResNet-50	53.06	26.66	17.44	27.03	17.91	26.09	28.90	21.83	19.70	22.74	54.34	27.78	17.41
ConvNeXt-S [22]	62.93	**45.78**	**26.82**	**36.68**	**42.33**	**41.40**	**42.74**	**31.63**	**29.86**	**41.04**	65.63	47.82	**37.10**
ViT-B/16 [5]	**67.37**	45.71	9.80	36.35	37.92	40.19	42.64	19.27	17.27	40.62	**69.61**	**47.98**	35.95
Swin-T [21]	62.07	39.72	23.49	32.02	33.92	35.32	36.68	27.38	26.82	35.79	64.38	42.76	35.28
(S4) NCT-CRC: Covariate OOD Acc% & MC-OODD PRR%													
ResNet-50	76.52	49.29	34.53	43.90	34.08	48.27	47.88	41.83	37.49	42.14	78.81	48.99	34.51
ConvNeXt-S [22]	**84.88**	**63.53**	**37.85**	**52.21**	**58.36**	**57.85**	**62.03**	**51.01**	**40.92**	**56.64**	**88.47**	**62.19**	52.07
ViT-B/16 [5]	84.38	61.44	23.11	50.55	55.56	54.76	59.21	43.61	30.65	54.52	87.77	59.45	**52.22**
Swin-T [21]	79.30	54.44	35.40	47.21	48.57	50.44	51.49	46.31	37.76	50.81	83.09	56.70	50.85

3 Results and Discussion

We report the results of experimental configs. (C1), (C2), (C3) in Tabs. 3-5, respectively. For clarity, in each table, we organize our results into four row *Sections*: (S1) and (S2) lists the *ID Acc.* and *each detector's S-OODD AUROC* over BreakHis and NCT-CRC, respectively. Note, we display ID Acc. separately for DE. We report the averaged metrics over the entire OSR splits, each with 10 trials using different random seeds. Similarly, (S3) and (S4) respectively shows the *covariate OOD Acc.* and *each detector's MC-OODD PRR* over BreakHis and NCT-CRC. The best metric across the TL (Tabs. 3-4) or architecture (Tab. 5) configs are in **bold** (*column-wise direction*). For each TL/architecture config, we further highlight the top-3 (1^{st}, 2^{nd}, 3^{rd}) performant detectors (*row-wise*).

Does DP domains pose more challenges? Not really – At the very least, the AUROC and PRR of our experiments do not deviate from the statistics therein natural imagery benchmarks [24,34]. This is promising as it suggests that *we can tackle this problem in a general framework catered to natural images*. However, another popular line of research is Outlier Exposure (OE), in which an auxiliary OOD set is exploited during training to encourage low confidence. Unlike natural imagery where a plethora of effective OE candidate sets are publicly available, collecting or synthesizing such a dataset in DP is challenging, even unfeasible. We leave this to future work.

Accuracy can be deceptive. While accuracy generally translates to S-OODD and MC-OODD performance, solely relying on it to rank TL or architecture configs. when all accuracies are within ±1% of each other is precarious. For instance, ViT-B/16 in Tab. 5 displays the highest accuracy in (S2) & (S3) but seldom beats ConvNeXt-S in S-OODD or MC-OODD across most detectors.

No universal SOTA detector. The ranking among detectors in Tabs. 3-5 is volatile, sensitive to the dataset, TL, and architecture. In general though, **ViM is superior in S-OODD** followed by KNN, Maha, TU, and GEN. Conversely, **TU and MSP excels in MC-OODD** followed by KLM and ViM. As a rule

of thumb, we recommend these detectors as an initial choice. From these trends, we surmise that the information from the rich feature representation is vital for S-OODD, whereas MC-OODD benefits more when the detector operates near the final output space. *Thus, achieving all-round superior performance in both tasks via a single detector is elusive.* Note, the differences in practical constraint (*e.g.*, ViM, KNN, Maha require ID samples and the feature embedding, while TU adopts the costly DE framework, which is not suited for rapid or mobile applications) further underscores this "*no winner*" posture.

The contradiction of UQ. Uncertainty due to OOD-ness must *axiomatically* be encoded in EU, whereas TU is the combined sum of OOD and aleatoric sources (AU, *e.g.*, ambiguity near class boundaries). Surprisingly though, *EU detectors often underperform TU*, further supported by [10, 30, 32]. Is this a problem? That is, can't we just use TU regardless of such an axiomatic contradiction? While doing so imposed no practical setback in our experiments, we anticipate it to be a problem when fuzzy data are non-negligible or the disentanglement of AU-EU sources is crucial (*e.g.*, in active learning, we generally wish to avoid querying fuzzy data).

Is DE worth the extra cost? The combination of ViM+MSP enjoys comparable/superior performance to the DE. *Hence, a powerful frequentist detector(s) may be all you need.* This is promising news for efficient DP as novelties and errors can be sought in a seamless and timely fashion. We recognize though that ensembling always boosts accuracy, and thus, DE holds merit in this respect.

DP-based TL helps, albeit unpredictably. We see compelling improvement over de-novo. However, further statements cannot be made. For instance, there is no clear winner among the SSL methods (Tab. 3), *e.g.*, SwAV is superior in S-OODD over BreakHis but never in NCT-CRC. QuiltNet also surprisingly fails to outperform BiomedCLIP (Tab. 4), in spite of the exposure to H&E in the former but not the latter; in fact, Quiltnet's gain is relatively marginal in BreakHis. Moreover, SIAYN (Tab. 5) is unexpectedly effective, sometimes on par or surpassing their SSL and LVLM counterparts (*e.g.*, MoCo v2, BiomedCLIP), despite using a much smaller dataset. *Although we affirm the advantages of DP TL, these examples (some counterintuitive) demonstrate that much research is still needed to codify its exact effects.*

What about TL with natural images? Such TL still often helps, sometimes more than DP TL as demonstrated by the NCT-CRC experiments in Tab. 4 wherein IN1K outperformed BiomedCLIP. In large, however, it is less advantageous, even performing worse than de-novo in certain cases (BreakHis experiments in Tab. 3). Our verdict is, *TL with natural imagery on average helps, however, further research is likewise needed (especially on when it fails) and should be implemented with caution.* TL from DP is the safer option and we recommend IN1K only when it is unavailable. In this spirit, we extend our SIAYN checkpoints to over ten popular, off-the-shelf DNN architectures and release them with our code.

ConvNeXt is robust, as are transformers. Recent studies to date have been dedicated to explaining the superior robustness properties of transformers

[1]. However, [24] showed that more advanced CNNs like ConvNeXt can behave just as robustly. Looking at Tab. 5, our findings agree with the latter study wherein ConvNeXt-S surpass ViT-B/16 and Swin-T in more number of metrics. Nonetheless, transformers perform well too and we conclude by saying *all three are SOTA*. However, not all CNNs are equal as we see a considerable gap between ResNet-50 and ConvNeXt-S. Hence, lumping them as one or analyzing just the ResNet family of architectures may give a false impression of CNNs. From this result, *we also urge to move away from ResNets, which remain a popular choice in DP, and onto more SOTA CNNs like ConvNeXt.*

4 Concluding Remark

We present an in-depth robustness study on S-OODD and MC-OODD in DP, with an emphasis on proper procedures and diverse settings. We reveal insights, some of which challenge the status quo taken for granted, *e.g.*, DE uncertainties may be redundant, TL from DP may not always behave as expected, CNNs can be just as robust as transformers. We hope our findings serve as a stepping stone to further exploration of pertinent themes in the realms of DP.

Acknowledgments. Research reported in this publication was supported by the National Institutes of Health under award numbers R01EB009745, R01CA260830, and R21CA263147. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by NSF grant 2243257, the National Science Foundation Science and Technology Center for Quantitative Cell Biology. Rohit Bhargava is a CZ Biohub Investigator.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Bai, Y., Mei, J., Yuille, A., Xie, C.: Are transformers more robust than cnns? In: Adv. Neural Inform. Process. Syst. pp. 26831–26843 (2021)
2. Cao, T., Huang, C.W., Hui, D.Y.T., Cohen, J.P.: A benchmark of medical out of distribution detection. J. Mach. Learn. Biomed. Imaging (2020)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 9912–9924 (2020)
4. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2020)
6. Guérin, J., Delmas, K., Ferreira, R., Guiochet, J.: Out-of-distribution detection is not all you need. In: AAAI Conf. Artif. Intell. vol. 37, pp. 14829–14837 (2023)

7. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: Int. Conf. Mach. Learn. pp. 8759–8773 (2022)
8. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Int. Conf. Learn. Represent. (2016)
9. Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty. In: Int. Conf. Mach. Learn. pp. 2712–2721 (2019)
10. Henning, C., D'Angelo, F., Grewe, B.F.: Are bayesian neural networks intrinsically good at out-of-distribution detection? In: Int. Conf. Mach. Learn. Worksh. (2021)
11. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: Adv. Neural Inform. Process. Syst. vol. 34, pp. 677–689 (2021)
12. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. In: Adv. Neural Inform. Process. Syst.
13. Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification. In: Int. Conf. Learn. Represent. (2022)
14. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3344–3354 (2023)
15. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med. **16**(1), e1002730 (2019)
16. Komura, D., Kawabe, A., Fukuta, K., Sano, K., Umezaki, T., Koda, H., Suzuki, R., Tominaga, K., Ochi, M., Konishi, H., et al.: Universal encoding of pan-cancer histology by deep texture representations. Cell Rep. **38**(9) (2022)
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Adv. Neural Inform. Process. Syst. pp. 6405–6416 (2017)
18. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Adv. Neural Inform. Process. Syst. pp. 7167–7177 (2018)
19. Linmans, J., Elfwing, S., van der Laak, J., Litjens, G.: Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Med. Image Anal. **83**, 102655 (2023)
20. Liu, X., Lochman, Y., Zach, C.: Gen: Pushing the limits of softmax-based out-of-distribution detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23946–23955 (2023)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10012–10022 (2021)
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11976–11986 (2022)
23. Malinin, A., Mlodozeniec, B., Gales, M.: Ensemble distribution distillation. In: Int. Conf. Learn. Represent. (2019)
24. Pinto, F., Torr, P.H., K. Dokania, P.: An impartial take to the cnn vs transformer robustness contest. In: Eur. Conf. Comput. Vis. pp. 466–480 (2022)

25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
26. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7), 1455–1462 (2015)
27. Sun, Y., Guo, C., Li, Y.: React: out-of-distribution detection with rectified activations. In: Adv. Neural Inform. Process. Syst. pp. 144–157 (2021)
28. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: Int. Conf. Mach. Learn. pp. 20827–20840 (2022)
29. Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J.D., Dahl, A.B.: Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: Med. Image Comput. Comput. Assist. Interv. pp. 824–833 (2020)
30. Ulmer, D., Cinà, G.: Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection. In: Uncertain. Artif. Intell. pp. 1766–1776 (2021)
31. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4921–4930 (2022)
32. Wimmer, L., Sale, Y., Hofman, P., Bischl, B., Hüllermeier, E.: Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In: Uncertain. Artif. Intell. pp. 2282–2292 (2023)
33. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Int. Conf. Mach. Learn. pp. 12310–12320 (2021)
34. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., et al.: Openood v1. 5: Enhanced benchmark for out-of-distribution detection. In: Adv. Neural Inform. Process. Syst. Worksh. (2023)
35. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv (2023)
36. Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., Yang, L.: Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In: Med. Image Comput. Comput. Assist. Interv. pp. 242–252 (2022)