MAKING THE MOST OF LIMITED DATA WITH SELF-SUPERVISED LEARNING FOR BREAST CANCER SCREENING

Christopher Clark[1] , Scott Kinder[1] , Syed Rakin Ahmed[2] , Giacomo Nebbia[1] , Yoga Advaith Veturi[1] , Praveer Singh[1] , and Jayashree Kalpathy-Cramer[1]
[1]University of Colorado School of Medicine, USA
[2]Harvard University, USA

Purpose

- The collection of high-quality, labeled images for medical classification remains a difficult and costly endeavor, making standard *Supervised Learning* (SL) medical classification tasks difficult. With *Self-Supervised Learning* (SSL), features can be learned from unlabeled data and fine-tuned on limited labeled samples.
- In this work, we compare SL methods with SSL methods at fractionations of 1, 50, and 100% of the labeled data.
- Further, we compare pretext SSL fine-tuning methods on our unlabeled data with using off-the-shelf ImageNet pretext pre-trained weights only.

Methods

- We are using two SSL training methods, *knowledge DIstillation* with NO labels (DINOv1) [1] *and Masked AutoEncode*r (MAE) [2], a contrastive and generative method, respectively.
- The DINOv1 method has as a backbone a CNN, ResNet50, and a Vision Transformer (ViT).
- The MAE uses a ViT, as well.

Experiments

- **Dataset:** Images are black and white, three-channel digital mammography images We have 83,039 unlabeled images for the SSL pretext fine-tuning and 72,606 images for the downstream classification training task.
- **Metrics:** Linearly-weighted κ
- **Classes:** Four classes indicating severity
- **Fractionations:** 100% SSL unlabeled data with 1, 10, and 100% SL labeled data in the downstream classification task

Results

Comparison of Models at 1%, 50%, and 100% Labeled Data Fractionation with Linearly Weighted κ score (Lower, Upper)

Model	1%	50%	100%
SL ResNet50	0.32 (0.19, 0.46)	0.54 (0.54, 0.55)	0.56 (0.55, 0.57)
SL ViT	0.45 (0.43, 0.47)	0.58 (0.57, 0.58)	0.59 (0.59. 0.60)
SSL ViTMAE	0.47 (0.46, 0.49)	0.58 (0.58, 0.59)	0.59 (0.59, 0.60)
SSL ViTMAE Domain	0.53 (0.52, 0.54)	0.59 (0.58, 0.59)	0.60 (0.60, 0.60)
SSL DINO ResNet50	0.54 (0.53, 0.55)	0.57 (0.56, 0.58)	0.59 (0.58, 0.59)
SSL DINO ResNet50 Domain	0.55 (0.54, 0.56)	0.58 (0.58, 0.59)	0.59 (0.58, 0.60)
SSL DINO ViT	0.48 (0.46, 0.49)	0.56 (0.55, 0.57)	0.58 (0.58, 0.59)
SSL DINO ViT Domain	0.55 (0.54, 0.55)	0.58 (0.57, 0.58)	0.59 (0.59, 0.59)

Conclusion

- With lower amounts of labeled data (1%), the best performing model is the SSL DINO ResNet50 and DINO ViT with in-domain pretext fine-tuning at 0.55.
- Even at higher labeled data regimes (100%), all the models except the SL ResNet50 reach around 0.59/0.60, showing that, though the benefit of SSL diminishes with more labeled data, these models are still as competitive as SL approaches.
- Comparing the in-domain vs ImageNet-only pretext fine-tuning, we see that at 1 and 50% labeled data, the in-domain versions of ResNet50 and ViT, both MAE and DINO SSL pretraining pipelines, outperform the ImageNet-only, and at 100%, they are very similar.

References

1. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
2. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint *arXiv:2010.11929, 2020*.

CONFORMAL PREDICTION AND MONTE CARLO INFERENCE FOR ADDRESSING UNCERTAINTY IN CERVICAL CANCER SCREENING

Christopher Clark[1], Scott Kinder[1], Didem Egemen[2], Brian Befano[3], KananDesai[2], Syed Rakin Ahmed[4], Praveer Singh[1], Ana Cecilia Rodriguez[2], Jose Jeronimo[2], Silvia De Sanjose[5], Nicolas Wentzensen[2], Mark Schiffman[2], and Jayashree Kalpathy-Cramer[1]

[1]University of Colorado School of Medicine, USA, [2]National Cancer Institute, USA, [3]Information Management Services, USA, [4]Harvard University, USA, [5]Cancer Epidemiology Research Programme, Spain

Purpose

•**Background:** Uncertainty is important in the medical domain, where misdiagnoses have important consequences. This project focuses on an in-development deep learning cervical cancer screening tool for use in low and middle-income countries.

•**Goal**: Compare two methods of uncertainty quantification

1. **Coefficient of Variation of Monte-Carlo Inference** (CoV) [1]
2. **Conformal Uncertainty Quantification** (CUQ) to an ongoing cervical cancer screening project PAVE to determine the relationship between uncertainty and performance [2]

Methods

•**CoV**: Run 50 rounds of inferences per sample with dropout left on, giving us 50 unique prediction vectors

•μ (σ): Average (standard deviation) of each prediction's *expected value*, then calculate the *Coefficient of Variation*

$$CoV = \sigma/\mu$$

•**CUQ**: Extract the length of the conformal prediction set $\hat{C}(x)$ for sample x with label y so that $\mathbb{P}[y \in \hat{C}(x)] \geq 1-\alpha$, with α the error rate (to be manually selected)

•Least Ambiguous Set-Valued Classifier (LAC) [2]

•Adaptive Prediction Set (APS) – *results not shown*

Experiments

•**Dataset:** 17,013 cervical images from 9,462 women from 5 studies across Costa Rica, the US, the Netherlands

•**classes**: Normal, Gray-Zone, Precancer+

•**Model**: DenseNet121

•**Task 1**: Relationship between conformal set length and type of prediction (correct, incorrect, single-, and double-class misclassification)

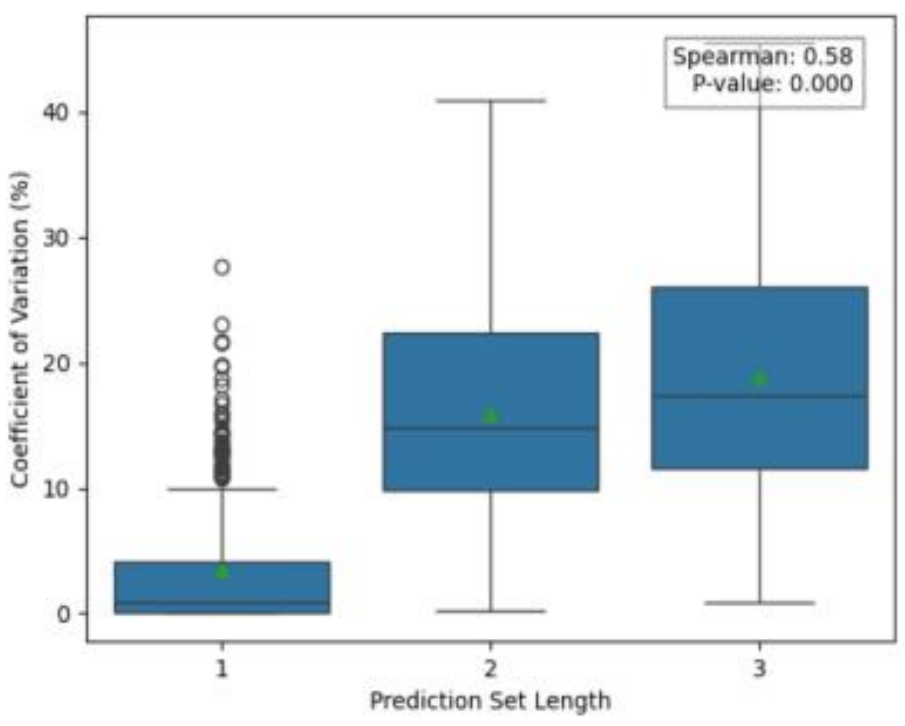
•**Task 2**: Correlation between CoV and CUQ

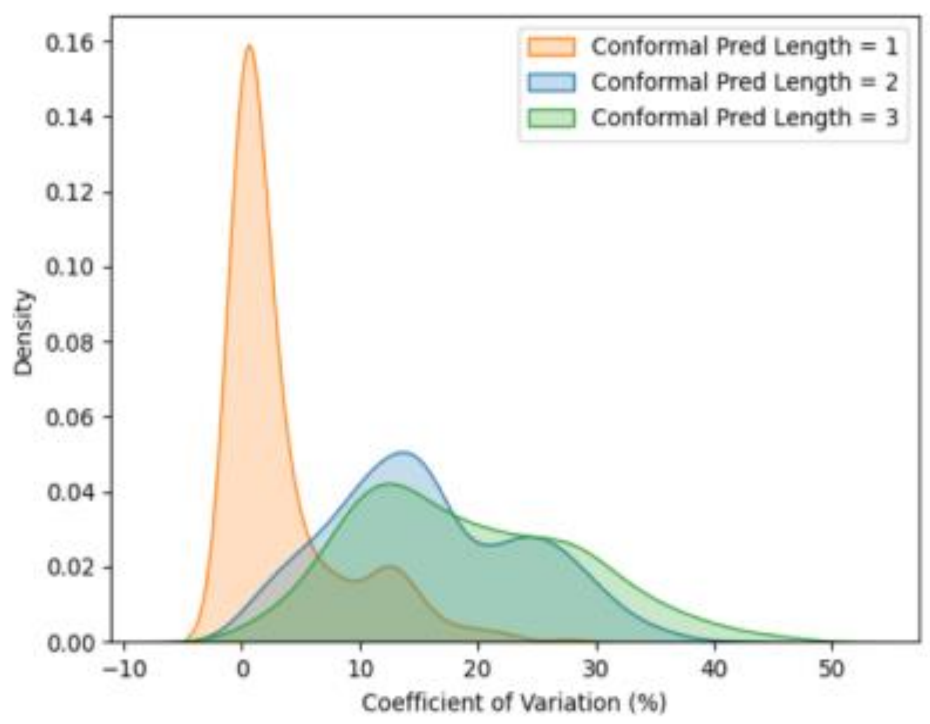
Results: Task 1

Prediction	Set Length	P-value < .05	
Correct	1.78	(ref.)	
Incorrect	2.38	*	
Single-class	2.43	*	(ref.)
Double-class	1.93	×	*

Ground Truth	Set Length Correct	Set Length Incorrect	P-value < .05
Normal	1.58	2.54	*
Gray-Zone	2.39	1.90	*
Precancer +	1.88	2.31	*

Results: Task 2





Conclusion

Task 1:

•When the model is incorrect, it is more uncertain. We see this comparing average prediction set length of 1.78 and 2.38.

•Gray-zone class associated with more uncertainty when the model is correct, 2.38 vs 1.90

Task 2:

•Statistically significant correlation between CoV and conformal prediction set lengths with $\rho = 0.58$

References

1. Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. PMLR, 2016.
2. Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv preprint arXiv:2107.07511 (2021).