Selective Prediction for Semantic Segmentation under Distribution Shift

Bruno L. C. Borges, Bruno M. Pacheco & Danilo Silva
Federal University of Santa Catarina (UFSC), Florianópolis, Brazil
`danilo.silva@ufsc.br`

Abstract

Semantic segmentation plays a crucial role in various computer vision applications, yet its efficacy is often hindered by the lack of high-quality labeled data. To address this challenge, a common strategy is to leverage models trained on data from different populations, such as publicly available datasets. This approach, however, leads to the distribution shift problem, presenting a reduced performance on the population of interest. In scenarios where model errors can have significant consequences, selective prediction methods offer a means to mitigate risks and reduce reliance on expert supervision. This paper investigates selective prediction for semantic segmentation in low-resource settings, thus focusing on post-hoc confidence estimators applied to pre-trained models operating under distribution shift. We propose a novel image-level confidence measure tailored for semantic segmentation and demonstrate its effectiveness through experiments on three medical imaging tasks. Our findings show that post-hoc confidence estimators offer a cost-effective approach to reducing the impacts of distribution shift.

1 Introduction

Deep learning has demonstrated remarkable success in various computer vision tasks, offering a powerful aid to labor-intensive processes. However, the scarcity of labeled data hinders the use of state-of-the-art models in many applications. This is particularly true in semantic segmentation of medical images, as the high dimensionality of input and output coupled with the high specialization needed for expert annotation usually implies in high costs for labeled data.

To circumvent the data limitation, a common approach is to leverage models trained on datasets drawn from different populations, giving rise to the problem of distribution shift. For example, a hospital in a developing country that does not have enough magnetic resonance images with high-quality annotations to train a state-of-the-art tumor segmentation model can resort to publicly available datasets on that application. However, the images used for training come from a different distribution than those expected to be fed to the model during clinical practice, as the acquisition hardware and the subjects' population may differ. Generally, the greater the degree of this distribution shift between training and test (or deployment) data, the poorer the generalization performance (Koh et al., 2021; Malinin et al., 2022).

In scenarios where model mistakes are critical, selective prediction methods are valuable tools to reduce the costs with expert supervision, as proposed in the seminal works of Geifman & El-Yaniv (2017) and Hendrycks & Gimpel (2016). At the core of selective prediction is the *confidence estimator*, a scalar function used to determine whether the model should abstain from a prediction and instead an expert should be invoked. A good confidence estimator enables the trade-off between expected performance and abstention rate of the deep learning model.

In this paper, we investigate the problem of selective prediction for semantic segmentation in low-resource settings. Given a lack of high-quality annotated data, we assume that a pre-trained model will be used and focus on post-hoc confidence estimators. In other words, a deep learning model trained on a different data distribution is used for inference, and the confidence estimators have as input solely the model's output. To evaluate the resulting selective prediction methods, we perform our experiments in three medical imaging segmentation tasks with state-of-the-art pre-trained models. Furthermore, we propose a novel image-level confidence measure based on the soft Dice

loss (Sudre et al., 2017; Milletari et al., 2016), and show that it outperforms all other confidence measures we tested.

The main contributions of this paper are summarized as follows:

- This is the first paper to evaluate *image-level* post-hoc confidence measures for selective prediction in semantic segmentation, extending the existing methods for classification;
- We propose a novel image-level confidence estimation method designed specifically for image segmentation, which takes into account usual segmentation evaluation metrics;
- We evaluate all methods in 3 medical imaging tasks (polyp segmentation, optic cup segmentation and multiple sclerosis lesion segmentation) under distribution shift, and show that the proposed method significantly outperforms other confidence estimators.

1.1 RELATED WORK

Given the increased labeling costs (in comparison to classification), selective prediction methods for semantic segmentation hold great promise. Many works (Nair et al., 2020; Jungo & Reyes, 2019; DeVries & Taylor, 2018; Lambert et al., 2022) have investigated the natural extension of classification confidence measures to pixel-wise (or voxel-wise) confidence estimation on semantic segmentation tasks. Such confidence maps (or uncertainty maps) can be useful for explainability, but often the expert intervention is performed on the whole image, rather than at pixel-level or region-level, which results in a need for image-level confidence estimates. Although the aggregation of pixel-level measures seems a natural choice, in this paper we show that they are unsuitable for semantic segmentation and instead proposed a novel measure designed specifically for this task.

Within the uncertainty estimation literature, a popular approach to compute confidence is the use of ensembles, which combine multiple predictions made for the same input; this includes conventional ensembles (Lakshminarayanan et al., 2017; Wen et al., 2020; Kushibar et al., 2022) as well as techniques based on Bayesian reasoning (Gal & Ghahramani, 2016; Malinin et al., 2022; Lambert et al., 2022). However, ensembles are not well suited to low-resource settings, as they usually increase training, development and/or storage costs. At the very least, ensembles either require multiple network passes for inference, increasing operational costs, or require custom models with custom training regimes, preventing the use of state-of-the-art, publicly-available pre-trained models. To fit the low-resource paradigm, our scope is limited to confidence estimation methods applicable to pre-trained single models.

We do not consider techniques to increase robustness to distribution shift, as they are entirely orthogonal to our approach; as long as there is some remaining performance gap due to distribution shift, selective prediction can always be used as a last resort to reduce this gap.

2 SELECTIVE PREDICTION FOR SEMANTIC SEGMENTATION

2.1 SELECTIVE PREDICTION

Let P be an unknown distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space, and let $h : \mathcal{X} \to \mathcal{Y}$ be a (predictive) model. The *risk* of h is $R(h) = E_P[\ell(h(x), y)]$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function. A *selective model* (Geifman & El-Yaniv, 2017; 2019) is a pair (h, g), where $g : \mathcal{X} \to \mathbb{R}$ is a *confidence estimator* (also known as *confidence score function* or *confidence-rate function*), which quantifies the model's confidence on its prediction for a given input. For some fixed threshold t, given an input x, the selective model makes a prediction $h(x)$ if $g(x) \geq t$, otherwise the prediction is rejected. A selective model's *coverage* $\phi(h, g) = P[g(x) \geq t]$ is the probability mass of the selected samples in $\mathcal{X}$, while its *selective risk* $R(h, g) = E_P[\ell(h(x), y) \mid g(x) \geq t]$ is its risk restricted to the selected samples. In particular, a model's risk equals its selective risk at *full coverage* (i.e., for t such that $\phi(h, g) = 1$). These quantities can be evaluated empirically given a given a test dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ drawn i.i.d. from P, yielding the *empirical coverage* $\hat{\phi}(h, g) = (1/N) \sum_{i=1}^N \mathbb{1}[g(x^{(i)}) \geq t]$ and the *empirical selective risk*

$$\hat{R}(h, g) = \frac{\sum_{i=1}^N \ell(h(x^{(i)}), y^{(i)}) \mathbb{1}[g(x^{(i)}) \geq t]}{\sum_{i=1}^N \mathbb{1}[g(x^{(i)}) \geq t]}. \tag{1}$$

Note that, by varying t, it is generally possible to trade off coverage for selective risk, i.e., a lower selective risk can usually (but not necessarily always) be achieved if more samples are rejected. This tradeoff is captured by the *risk-coverage (RC) curve* (Geifman & El-Yaniv, 2017), a plot of $\hat{R}(h, g)$ as a function of $\hat{\phi}(h, g)$. While the RC curve provides a full picture of the performance of a selective classifier, it is convenient to have a scalar metric that summarizes this curve. A commonly used metric is the *area under the RC curve* (AURC) (Ding et al., 2020; Geifman et al., 2019), denoted by $\text{AURC}(h, g)$. However, the AURC is not always easy to interpret. For a more tangible metric, we propose here a simple generalization of the *selective accuracy constraint* (Galil et al., 2023) from classification to generic prediction: the *coverage at selective risk* (CSR), defined as the maximum coverage allowed for a model to achieve a specified risk.

For semantic segmentation, inputs are images and labels are segmentation masks, so, e.g., $\mathcal{X} = [0, 1]^{n_C \times n_W \times n_H}$ and $\mathcal{Y} = \{0, 1, \dots, K-1\}^{n_W \times n_H}$ if images are $n_W \times n_H$ pixels with n_C channels and segmentation masks are divided into K classes. Here we consider only binary segmentation, so $K = 2$. To simplify notation, we assume *flattened* segmentation masks, so that $\mathcal{Y} = \{0, 1\}^n$, where $n = n_W n_H$ is the total number of output pixels. In the binary case, a commonly used quality metric is the Dice Similarity Coefficient (DSC), defined as $D(\hat{\mathbf{y}}, \mathbf{y}) = 2 \sum_{j=1}^n \hat{y}_j y_j / \sum_{j=1}^n (\hat{y}_j + y_j)$, where $\hat{\mathbf{y}}, \mathbf{y} \in \mathcal{Y}$. We convert it to a loss by taking $\ell(\hat{\mathbf{y}}, \mathbf{y}) = 1 - D(\hat{\mathbf{y}}, \mathbf{y})$.

2.2 CONFIDENCE ESTIMATORS

We restrict attention to segmentation models that can be decomposed as $h(x) = \mathbb{1}[f(x) > \gamma]$, where $f : \mathcal{X} \to [0, 1]^n$ is a neural network, $\gamma \in [0, 1]$ is a decision threshold (e.g., $\gamma = 0.5$), and $\mathbb{1}[\cdot]$ denotes the indicator function extended to vectors in element-wise fashion. In other words, if we denote $\hat{\mathbf{p}} = f(x)$, then $\hat{p}_j$ can be interpreted as an estimate of the posterior probability $P[y_j = 1|x]$ that the jth pixel of the ground truth mask is equal to 1.

We focus on confidence estimators that can be computed directly from the model outputs $\hat{\mathbf{p}} = f(x)$ and $\hat{\mathbf{y}} = h(x)$. To the best of our knowledge, image-level confidence estimators applicable to a single model had not been proposed before, so first we consider a few natural baselines.

The first two are based on widely-used pixel-level confidence estimators, namely, the *maximum softmax probability* (MSP) (Hendrycks & Gimpel, 2016; Geifman & El-Yaniv, 2017; Ding et al., 2020) $g_j(x) = \max\{\hat{p}_j, 1 - \hat{p}_j\}$ and the *negative entropy*[1] (Belghazi & Lopez-Paz, 2021) $g_j(x) = \hat{p}_j \log \hat{p}_j + (1 - \hat{p}_j) \log(1 - \hat{p}_j)$. We adapt them to image-level confidence estimators by taking their mean over all pixels, $g(x) = (1/n) \sum_{j=1}^n g_j(x)$. The third baseline exploits a peculiarity of the Dice score. As shown in (Carass et al., 2020; Raina et al., 2023; Malinin et al., 2022), the Dice score is biased towards providing higher scores to images with higher prevalence of lesion. With this in mind, we propose as a baseline confidence estimator the *predicted lesion load*, computed as the proportion of the predicted foreground class of an image, namely, $g(x) = \hat{\rho}(x) \triangleq (1/n) \sum_{j=1}^n \hat{y}_j$.

Finally, we propose a confidence measure based on the Dice coefficient. Our motivation is to have a confidence estimator somehow connected to the evaluation metric. We define the *soft Dice confidence* (SDC) as[2] $g(x) = 2 \sum_{j=1}^n \hat{p}_j \hat{y}_j / \sum_{j=1}^n (\hat{p}_j + \hat{y}_j)$. Note that the SDC increases as the confidence of a gixen pixel increases within the same class; in particular, it is highest (equal to 1) when $\hat{\mathbf{p}} = \hat{\mathbf{y}}$. On the other hand, when the model is least confident at the pixel level, $\hat{\mathbf{p}} \approx \gamma$, it can be shown that $g(x) \approx 2\gamma/(1 + \gamma/\hat{\rho}(x))$, which is an increasing function of the predicted lesion load $\hat{\rho}(x)$. Thus, in this case, it at least retains the properties of the lesion load baseline.

3 EXPERIMENTS AND RESULTS

3.1 SEMANTIC SEGMENTATION TASKS

In the following we describe the segmentation tasks and respective models considered, as well as the datasets used for in-distribution (ID) and out-of-distribution (OOD) evaluation.

---

[1] Note that any uncertainty estimator can be used as a confidence estimator by taking its negative.

[2] The SDC is inspired by and closely related to the *soft Dice loss* (SDL) (Sudre et al., 2017); specifically, it equals 1 minus SDL when the hard predictions $\hat{\mathbf{y}}$ are used in place of the ground truth $\mathbf{y}$.

Polyp segmentation. Polyp segmentation in colonoscopy images aids identification and evaluation of these clumps, which is key in colorectal cancer prevention (Jha et al., 2019). In our study, we build the dataset for polyp segmentation as reported in Dong et al. (2023). More specifically, 900 images from KvasirSEG (Jha et al., 2019) and 548 images from ClinicDB (Bernal et al., 2015) build the training set, while the remaining 162 images from both are used for ID evaluation. The OOD dataset is composed with 196 images from ETIS (Silva et al., 2014), 380 images from ColonDB (Tajbakhsh et al., 2016), and 60 images from EndoScene (Vázquez et al., 2017). As segmentation models, we use the Polyp-PVT, a deep learning model based on the pyramid vision transformer (PVT), and a baseline U-Net. Both models were made available (both structure and weights) by Dong et al. (2023).

Optic cup segmentation. Accurately segmenting the cup and disc in color fundus photography (CFP) plays a key role in the early diagnosis of glaucoma and the prevention of irreversible vision loss (Orlando et al., 2020). To instantiate optic cup segmentation task, we use three publicly available datasets: REFUGE (Orlando et al., 2020), with 1600 images; ORIGA (Zhuo Zhang et al., 2010), with 650 images; and G1020 (Bajwa et al., 2020), with 1020 images. 1200 images from REFUGE compose the training set, while the remaining 400 images are used as the ID evaluation set, and the entirety of ORIGA and G1020 compose the OOD evaluation set. All images (and corresponding masks) were preprocessed by the MNet DeepCDR model (Fu et al., 2018) to crop a region of interest containing the optic disc and cup. To maintain consistency with the other tasks, images corresponding to labels with $y = 0$ (no optic cup appearing in the ground truth mask) were systematically excluded as they trivially indicate a preprocessing failure. Therefore, the resulting revised evaluation datasets comprise 394 images from REFUGE, 648 from ORIGA, and 782 from G1020. For segmentation, we use the Segtran model, developed and made available by Li et al. (2021). The authors aim to streak a balance between increasing context and improving localization accuracy by combining a convolutional backbone with a series of squeeze-and-expansion transformer layers.

Multiple Sclerosis White Matter Lesion (MSWML) segmentation. The segmentation of white matter lesions caused by multiple sclerosis helps physicians follow the progress of the disease through the evolution of the lesions (Thompson et al., 2018). We use the datasets and baseline segmentation model provided in the Shifts 2.0 challenge (Malinin et al., 2022). We consider the ISBI (Carass et al., 2017) and MSSEG-1 (Commowick et al., 2018) as ID data, while the PubMRI (Lesjak et al., 2018) is considered OOD. The ID evaluation set contains 33 images from each source, while the OOD set contains 25 images. As our baseline segmentation model, we use a (single-model) 3D U-Net made available by the competition organizers, as few competitors were able to overcome the baseline performance in terms of ID and OOD Dice score.

3.2 Results

For each task and model, we evaluate the confidence estimators presented in Sec. 2.2 with respect to their selective prediction performance on OOD data, simulating the low-resource setting described in Sec. 1. As a bound, we also evaluate a hypothetically optimal confidence estimator (denoted "ideal") that perfectly orders samples in decreasing order of their losses. Following the literature on selective prediction, we use the AURC metric, with selective risk taken as 1 minus the average Dice score over the selected samples (see Sec. 2.1). The results are shown in Table 1 where, for comparison, the performance on ID data is also shown. The full RC curve on OOD data for three examples (one for each task) is shown in Fig. 1. Even though the performances differ significantly between tasks (and even between models, as in the case of the polyp segmentation), the proposed confidence estimator SDC visibly achieves the best overall performance, providing the smallest risk in almost all coverage levels for all tasks and models.

Finally, to give a more tangible illustration of the potential of selective prediction, in Table 2 we measure the maximum coverage on OOD data that produces a selective risk equal to that obtained at full coverage on ID data (assuming that the ID performance of the models would be considered satisfactory). In other words, for each confidence estimator, we estimate how much of the OOD data can be processed solely by the deep learning model (instead of resorting to an expert) so that no performance drop is observed due to the distribution shift. It can be seen that the SDC is not only the best option by a wide margin, but also the only that provided consistent results. However, we can also see that this metric is highly dependent on the task and model.

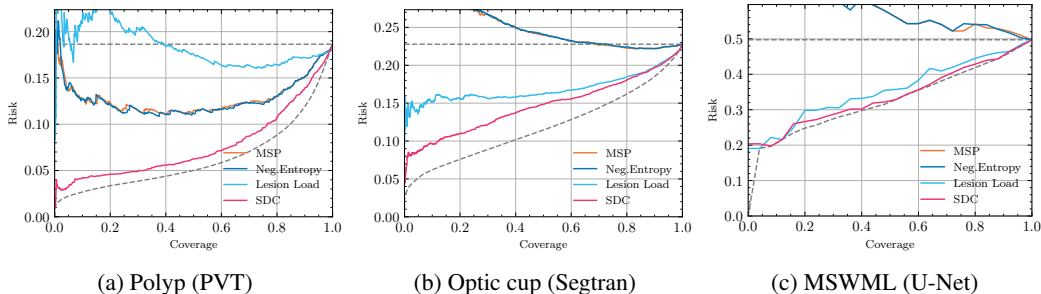(a) Polyp (PVT) (b) Optic cup (Segtran) (c) MSWML (U-Net)

Figure 1: Risk-coverage curves for all confidence estimators across all tasks (one model for each task) on the respective OOD evaluation sets. The dashed lines represent the random confidence estimator (horizontal line) and the ideal confidence estimator.

Table 1: AURC of confidence estimators across all task-model pairs (lower is better). *Risk* indicates the original risk without selective prediction, which equals the AURC of the random estimator.

	Polyp (PVT)		Polyp (U-Net)		Optic cup (Segtran)		MSWML (U-Net)	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Risk	0.072	0.186	0.173	0.507	0.074	0.228	0.377	0.497
MSP	0.068	0.129	0.099	0.370	0.078	0.270	0.414	0.600
Negative Entropy	0.068	0.128	0.091	0.319	0.078	0.268	0.418	0.595
Lesion Load	0.063	0.185	0.170	0.333	0.064	0.169	0.302	0.355
SDC	0.036	0.076	0.059	0.217	0.058	0.146	0.282	0.338
Ideal	0.028	0.062	0.049	0.195	0.047	0.119	0.259	0.326

Table 2: Maximum coverage of the OOD dataset using the designated confidence estimator that reproduces the ID risk at 100 % coverage (higher is better). Blank values (-) indicate that the corresponding ID risk could not be achieved for any coverage levels. *Ideal* is the maximum coverage derived from using the ideal confidence estimator.

	Polyp (PVT)	Polyp (U-Net)	Optic cup (Segtran)	MSWML (U-Net)
MSP	-	-	-	-
Negative Entropy	-	6.0 %	-	-
Lesion Load	-	-	-	56.0 %
SDC	60.5 %	49.5 %	0.8 %	64.0 %
Ideal	70.9 %	56.8 %	18.3 %	64.0 %

Our full results on OOD data broken-down by each data source are presented in Appendix A.

4 CONCLUSION

In this paper, we evaluated confidence estimators for semantic segmentation with respect to their selective prediction capacity in the context of distribution shift due to the lack of in-domain, high-quality data. The proposed confidence estimator, soft Dice confidence, outperformed all other confidence estimators evaluated, as shown through the RC curves of our results (see Fig. 1 and Table 1). Furthermore, our results suggest that selective prediction can be a viable option for tackling the distribution shift problem when other options have been exhausted. With a suitable confidence estimator, a selective model provides a sensible way of achieving ID performance during clinical practice, while saving on expert costs. The evaluation through the maximum coverage to achieve ID performance in OOD data (see Table 2) demonstrates such goal, as this ratio can be directly translated into expert hours that could be saved by employing a selective model. However, despite a significant performance in the tasks of polyp and MSWML segmentation, the poor performance on optic cup segmentation indicates that there is still a wide avenue for future research.

REFERENCES

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection. 2020. doi: 10.48550/ARXIV.2006.09158. URL https://arxiv.org/abs/2006.09158. Publisher: arXiv Version Number: 1.

Mohamed Ishmael Belghazi and David Lopez-Paz. What classifiers know what they don't?, July 2021. URL http://arxiv.org/abs/2107.06217.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, July 2015. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2015.02.007. URL https://www.sciencedirect.com/science/article/pii/S0895611115000567.

Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L. Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H. Sudre, Manuel Jorge Cardoso, Niamh Cawley, Olga Ciccarelli, Claudia A. M. Wheeler-Kingshott, Sébastien Ourselin, Laurence Catanese, Hrishikesh Deshpande, Pierre Maurel, Olivier Commowick, Christian Barillot, Xavier Tomas-Fernandez, Simon K. Warfield, Suthirth Vaidya, Abhijith Chunduru, Ramanathan Muthuganapathy, Ganapathy Krishnamurthi, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Leonardo O. Iheme, Devrim Unay, Saurabh Jain, Diana M. Sima, Dirk Smeets, Mohsen Ghafoorian, Bram Platel, Ariel Birenbaum, Hayit Greenspan, Pierre-Louis Bazin, Peter A. Calabresi, Ciprian M. Crainiceanu, Lotta M. Ellingsen, Daniel S. Reich, Jerry L. Prince, and Dzung L. Pham. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148:77–102, March 2017. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2016.12.064. URL https://www.sciencedirect.com/science/article/pii/S1053811916307819.

Aaron Carass, Snehashis Roy, Adrian Gherman, Jacob C. Reinhold, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Mohsen Ghafoorian, Bram Platel, Ariel Birenbaum, Hayit Greenspan, Dzung L. Pham, Ciprian M. Crainiceanu, Peter A. Calabresi, Jerry L. Prince, William R. Gray Roncal, Russell T. Shinohara, and Ipek Oguz. Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Scientific Reports*, 10(1):8242, May 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-64803-w. URL https://www.nature.com/articles/s41598-020-64803-w. Number: 1 Publisher: Nature Publishing Group.

Olivier Commowick, Audrey Istace, Michaël Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Améli, Jean-Christophe Ferré, Anne Kerbrat, Thomas Tourdias, Frédéric Cervenansky, Tristan Glatard, Jérémy Beaumont, Senan Doyle, Florence Forbes, Jesse Knight, April Khademi, Amirreza Mahbod, Chunliang Wang, Richard McKinley, Franca Wagner, John Muschelli, Elizabeth Sweeney, Eloy Roura, Xavier Lladó, Michel M. Santos, Wellington P. Santos, Abel G. Silva-Filho, Xavier Tomas-Fernandez, Hélène Urien, Isabelle Bloch, Sergi Valverde, Mariano Cabezas, Francisco Javier Vera-Olmos, Norberto Malpica, Charles Guttmann, Sandra Vukusic, Gilles Edan, Michel Dojat, Martin Styner, Simon K. Warfield, François Cotton, and Christian Barillot. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Scientific Reports*, 8(1):13650, September 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-31911-7. URL https://www.nature.com/articles/s41598-018-31911-7. Number: 1 Publisher: Nature Publishing Group.

Terrance DeVries and Graham W. Taylor. Leveraging Uncertainty Estimates for Predicting Segmentation Quality. 2018. doi: 10.48550/ARXIV.1807.00502. URL https://arxiv.org/abs/1807.00502. Publisher: arXiv Version Number: 1.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the Evaluation of Uncertainty Estimation and Its Application to Explore Model Complexity-Uncertainty Trade-Off. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 22–31, June 2020. doi: 10.1109/CVPRW50498.2020.00010. URL https://ieeexplore.ieee.org/document/9150782. ISSN: 2160-7516.

Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers, April 2023. URL `http://arxiv.org/abs/2108.06932`. arXiv:2108.06932 [cs, eess].

Huazhu Fu, Jun Cheng, Yanwu Xu, Changqing Zhang, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Disc-aware Ensemble Network for Glaucoma Screening from Fundus Image. *IEEE Transactions on Medical Imaging*, 37(11):2493–2501, November 2018. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2018.2837012. URL `http://arxiv.org/abs/1805.07549`. arXiv:1805.07549 [cs].

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, June 2016. PMLR. URL `https://proceedings.mlr.press/v48/gal16.html`.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What Can We Learn From The Selective Prediction And Uncertainty Estimation Performance Of 523 Imagenet Classifiers, February 2023. URL `http://arxiv.org/abs/2302.11874`. arXiv:2302.11874 [cs].

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html`.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151–2159. PMLR, May 2019. URL `https://proceedings.mlr.press/v97/geifman19a.html`. ISSN: 2640-3498.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers, April 2019. URL `http://arxiv.org/abs/1805.08206`. arXiv:1805.08206 [cs, stat].

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. 2016. doi: 10.48550/ARXIV.1610.02136. URL `https://arxiv.org/abs/1610.02136`. Publisher: arXiv Version Number: 3.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-SEG: A Segmented Polyp Dataset, November 2019. URL `http://arxiv.org/abs/1911.07069`. arXiv:1911.07069 [cs, eess].

Alain Jungo and Mauricio Reyes. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, volume 11765, pp. 48–56. Springer International Publishing, Cham, 2019. ISBN 978-3-030-32244-1 978-3-030-32245-8. doi: 10.1007/978-3-030-32245-8_6. URL `https://link.springer.com/10.1007/978-3-030-32245-8_6`. Series Title: Lecture Notes in Computer Science.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, July 2021. URL `https://proceedings.mlr.press/v139/koh21a.html`.

Kaisar Kushibar, Víctor Manuel Campello, Lidia Garrucho Moras, Akis Linardos, Petia Radeva, and Karim Lekadir. Layer Ensembles: A Single-Pass Uncertainty Estimation in Deep Learning for Segmentation, March 2022. URL `http://arxiv.org/abs/2203.08878`. arXiv:2203.08878 [cs, eess].

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka, and Michel Dojat. Fast Uncertainty Quantification for Deep Learning-based MR Brain Segmentation. pp. 1, January 2022. URL https://hal.science/hal-03498120.

Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. *Neuroinformatics*, 16(1):51–63, January 2018. ISSN 1539-2791, 1559-0089. doi: 10.1007/s12021-017-9348-7. URL http://link.springer.com/10.1007/s12021-017-9348-7.

Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Goh. Medical Image Segmentation Using Squeeze-and-Expansion Transformers, June 2021. URL http://arxiv.org/abs/2105.09511. arXiv:2105.09511 [cs, eess] version: 3.

Andrey Malinin, Neil Band, Ganshin, Alexander, German Chesnokov, Yarin Gal, Mark J. F. Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Roginskiy, Denis, Mariya Shmatova, Panos Tigas, and Boris Yangel. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks, February 2022. URL http://arxiv.org/abs/2107.07455. arXiv:2107.07455 [cs, stat].

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, Stanford, CA, USA, October 2016. IEEE. ISBN 978-1-5090-5407-7. doi: 10.1109/3DV.2016.79. URL http://ieeexplore.ieee.org/document/7785132/.

Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, January 2020. ISSN 13618415. doi: 10.1016/j.media.2019.101557. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841519300994.

José Ignacio Orlando, Huazhu Fu, João Barbossa Breda, Karel van Keer, Deepti R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murugesan, Valery Naranjo, Sai Samarth R. Phaye, Sharath M. Shankaranarayana, Apoorva Sikka, Jaemin Son, Anton van den Hengel, Shujun Wang, Junyan Wu, Zifeng Wu, Guanghui Xu, Yongli Xu, Pengshuai Yin, Fei Li, Xiulan Zhang, Yanwu Xu, Xiulan Zhang, and Hrvoje Bogunović. REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs. *Medical Image Analysis*, 59:101570, January 2020. ISSN 13618415. doi: 10.1016/j.media.2019.101570. URL http://arxiv.org/abs/1910.03667. arXiv:1910.03667 [cs].

Vatsal Raina, Nataliia Molchanova, Mara Graziani, Andrey Malinin, Henning Muller, Meritxell Bach Cuadra, and Mark Gales. Tackling Bias in the Dice Similarity Coefficient: Introducing nDSC for White Matter Lesion Segmentation, February 2023. URL http://arxiv.org/abs/2302.05432. arXiv:2302.05432 [cs, eess].

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, March 2014. ISSN 1861-6410, 1861-6429. doi: 10.1007/s11548-013-0926-3. URL http://link.springer.com/10.1007/s11548-013-0926-3.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi,

Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu (eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 10553, pp. 240–248. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67557-2 978-3-319-67558-9. doi: 10.1007/978-3-319-67558-9_28. URL `http://link.springer.com/10.1007/978-3-319-67558-9_28`. Series Title: Lecture Notes in Computer Science.

Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, February 2016. ISSN 1558-254X. doi: 10.1109/TMI.2015.2487997. URL `https://ieeexplore.ieee.org/document/7294676`. Conference Name: IEEE Transactions on Medical Imaging.

Alan J Thompson, Brenda L Banwell, Frederik Barkhof, William M Carroll, Timothy Coetzee, Giancarlo Comi, Jorge Correale, Franz Fazekas, Massimo Filippi, Mark S Freedman, Kazuo Fujihara, Steven L Galetta, Hans Peter Hartung, Ludwig Kappos, Fred D Lublin, Ruth Ann Marrie, Aaron E Miller, David H Miller, Xavier Montalban, Ellen M Mowry, Per Soelberg Sorensen, Mar Tintoré, Anthony L Traboulsee, Maria Trojano, Bernard M J Uitdehaag, Sandra Vukusic, Emmanuelle Waubant, Brian G Weinshenker, Stephen C Reingold, and Jeffrey A Cohen. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2):162–173, February 2018. ISSN 1474-4422. doi: 10.1016/S1474-4422(17)30470-2. URL `https://www.sciencedirect.com/science/article/pii/S1474442217304702`.

David Vázquez, Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *Journal of Healthcare Engineering*, 2017:1–9, 2017. ISSN 2040-2295, 2040-2309. doi: 10.1155/2017/4037190. URL `https://www.hindawi.com/journals/jhe/2017/4037190/`.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An Alternative Approach to Efficient Ensemble and Lifelong Learning, February 2020. URL `http://arxiv.org/abs/2002.06715`. arXiv:2002.06715 [cs, stat].

Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. ORIGA$^{\text{-light}}$: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 3065–3068, Buenos Aires, August 2010. IEEE. ISBN 978-1-4244-4123-5 978-1-4244-4124-2. doi: 10.1109/IEMBS.2010.5626137. URL `http://ieeexplore.ieee.org/document/5626137/`.

A FULL RESULTS ON OOD DATA

A.1 POLYP SEGMENTATION

Table 3: AURC of confidence estimators across the Polyp task (lower is better). *Risk* indicates the original risk without selective prediction, which equals the AURC of the random estimator.

	Polyp (PVT)			Polyp (U-Net)		
	ColonDB	ETIS	Endoscene	ColonDB	ETIS	Endoscene
Risk	0.189	0.210	0.096	0.496	0.597	0.283
MSP	0.137	0.137	0.075	0.357	0.526	0.206
Negative Entropy	0.136	0.134	0.075	0.306	0.476	0.152
Lesion Load	0.203	0.160	0.104	0.320	0.389	0.304
SDC	0.080	0.078	0.064	0.207	0.301	0.108
Ideal	0.065	0.066	0.049	0.188	0.262	0.093

Table 4: Maximum coverage of the OOD dataset for Polyp using the designated confidence estimator that reproduces the ID risk at 100 % coverage (higher is better). Blank values (-) indicate that the corresponding ID risk could not be achieved for any coverage levels.

| | Polyp (PVT) | | | Polyp (U-Net) | | |
	ColonDB	ETIS	Endoscene	ColonDB	ETIS	Endoscene
MSP	-	-	76.7 %	-	1.5 %	76.7 %
Negative Entropy	-	-	78.3 %	1.8 %	2.0 %	76.7 %
Lesion Load	-	9.2 %	3.3 %	0.3 %	-	-
SDC	58.7 %	53.1 %	83.3 %	54.0 %	30.1 %	85.0 %
Ideal	68.9 %	67.3 %	93.3 %	58.7 %	43.4 %	85.0 %

A.2 OPTIC CUP SEGMENTATION

Table 5: AURC of confidence estimators across the Optic cup task (lower is better). *Risk* indicates the original risk without selective prediction, which equals the AURC of the random estimator.

| | Optic cup (Segtran) | |
	ORIGA	G1020
Risk	0.212	0.241
MSP	0.268	0.276
Negative Entropy	0.269	0.273
Lesion Load	0.145	0.187
SDC	0.134	0.154
Ideal	0.109	0.129

Table 6: Maximum coverage of the OOD dataset for Optic cup using the designated confidence estimator that reproduces the ID risk at 100 % coverage (higher is better). Blank values (-) indicate that the corresponding ID risk could not be achieved for any coverage levels.

| | Optic cup (Segtran) | |
	ORIGA	G1020
MSP	-	-
Negative Entropy		-
Lesion Load	-	0.3 %
SDC	1.7 %	0.9 %
Ideal	23.6 %	13.6 %