Robust Conformal Volume Estimation in 3D Medical Images

Benjamin Lambert, Florence Forbes, Senan Doyle, Michel Dojat

▶ **To cite this version:**

Benjamin Lambert, Florence Forbes, Senan Doyle, Michel Dojat. Robust Conformal Volume Estimation in 3D Medical Images. MICCAI 2024 - 27th International Conference on Medical Image Computing and Computer Assisted Intervention, Oct 2024, Marakech, Morocco. pp.1-11, 10.1007/978-3-031-72117-5_59 . hal-04915405

HAL Id: hal-04915405
https://hal.science/hal-04915405v1

Submitted on 27 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

arXiv:2407.19938v1 [cs.CV] 29 Jul 2024

Robust Conformal Volume Estimation in 3D Medical Images

Benjamin Lambert[1,2], Florence Forbes[3], Senan Doyle[2], and Michel Dojat[1]

[1] Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut Neurosciences, 38000, FR
[2] Pixyl, Research and Development Laboratory, 38000 Grenoble, FR
[3] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, FR

Abstract. Volumetry is one of the principal downstream applications of 3D medical image segmentation, for example, to detect abnormal tissue growth or for surgery planning. Conformal Prediction is a promising framework for uncertainty quantification, providing calibrated predictive intervals associated with automatic volume measurements. However, this methodology is based on the hypothesis that calibration and test samples are exchangeable, an assumption that is in practice often violated in medical image applications. A weighted formulation of Conformal Prediction can be framed to mitigate this issue, but its empirical investigation in the medical domain is still lacking. A potential reason is that it relies on the estimation of the density ratio between the calibration and test distributions, which is likely to be intractable in scenarios involving high-dimensional data. To circumvent this, we propose an efficient approach for density ratio estimation relying on the compressed latent representations generated by the segmentation model. Our experiments demonstrate the efficiency of our approach to reduce the coverage error in the presence of covariate shifts, in both synthetic and real-world settings. Our implementation is available at `https://github.com/benolmbrt/wcp_miccai`.

Keywords: Uncertainty · Predictive Interval · Covariate Shift

1 Introduction

An important downstream application of medical image segmentation is the extraction of volume measurements for lesions or organs. Lesion volumetry plays a pivotal role in various medical scenarios, including predicting the outcome after stroke [12], grading brain tumors [5], or monitoring the progression of Multiple Sclerosis [20]. Brain volumetry can also be useful to monitor atrophy [9], and analyzing the volume of organs is useful for aging studies [27]. However, automated segmentations can be error-prone, which inevitably leads to imprecise volumetric measurements. A potential solution would be to associate predictive intervals (PIs) with the estimations to take into account this uncertainty.

Conformal prediction (CP) [19,23] is an uncertainty paradigm allowing to associate PIs with regressed scores (here, volumes). The most popular variant

of CP, Split CP [23], relies on a set-aside calibration dataset (generally a subset of the training dataset) that is used to calibrate the intervals so that they match the target coverage level on fresh test data. However, it is based on the exchangeability hypothesis, following which calibration and test data are drawn independently from the same distribution. In general, this is not the case for medical image processing applications, where domain shifts are extremely common, due to variations in the data acquisition protocol or the presence of pathologies unseen during training [29]. When calibration and test data points are not exchangeable, the accuracy of the conformal procedure collapses drastically [4,26], which hinders the relevancy of conformalized PIs in medical applications.

As a potential solution, Weighted Conformal Prediction (WCP) has been proposed to account for shifts between calibration and test distributions [4,26]. It is based on the reweighting of calibration samples according to the estimated density ratio $dP_{\text{test}}/dP_{\text{train}}$. As a result, calibration samples close to the test samples are attributed with higher importance in the conformal procedure. A flourishing literature can be found for density ratio estimation, with popular approaches including the training of a classifier to distinguish between training and test distributions [6,2], moment [14] or ratio matching [15]. More recently, Deep Learning (DL) approaches are also investigated to estimate density ratios [10,22]. However, we note that applications of WCP to medical image segmentation are still lacking, which may be due to the difficulty of estimating the density ratio for high-dimensional imaging data.

In this work, we propose to investigate the use of WCP to tackle covariate shifts in medical image segmentation tasks, with the ultimate goal of computing calibrated PIs for lesion volumes. As a contribution, we propose an efficient way of computing the density ratio in high-dimensional medical images, by relying on latent representations generated by the segmentation model.

2 Conformal Prediction for volumetry in medical images

2.1 Problem definition

We consider a 3D segmentation problem with N classes where our objective is to estimate the true volumes $Y \in \mathbb{R}^{N-1}$ of each foreground class based on the predicted segmentation. Within this framework, for an estimation X of the volume, we define a predictive interval $\Gamma_\alpha(X)$ as a range of values that are constructed to contain the true volume Y with a user-defined degree of confidence $1-\alpha$ (e.g 90% or 95%). More formally, given a set of estimated volumes $X_1 \dots X_n$ and their corresponding ground truth volumes $Y_1 \dots Y_n$, $\Gamma_\alpha(\cdot)$ should be learned such that it satisfies [1]:

$$1 - \alpha \leq P(Y_{\text{test}} \in \Gamma_\alpha(X_{\text{test}})) \leq 1 - \alpha + \frac{1}{n+1} \tag{1}$$

2.2 Predictive Interval computation using a multi-head segmentation architecture

In practice, a PI associated with a volume X_i is composed of a lower bound l_i, and an upper bound u_i. For a practical estimation of these three quantities (X_i, l_i and u_i), [18] proposed to train a multi-head segmentation network that predicts 3 output masks: a restrictive one (low recall, high precision) to estimate the lower bound, a permissive one (high recall, low precision) to estimate the upper bound, and a balanced one for the estimation of the mean (Figure 1). The key element is to perform training using the Tversky loss $T_{\alpha,\beta}$ [24] allowing to control the penalties applied to false positives (FP) and negatives (FN) contained in each mask through the loss parameters α and β, respectively. Writing p_{lower}, p_{mean} and p_{upper} the outputs of each head and y the ground-truth segmentation, the loss is defined:

$$\mathcal{L} = T_{1-\gamma,\gamma}(p_{lower,y}) + T_{0.5,0.5}(p_{mean,y}) + T_{\gamma,1-\gamma}(p_{upper,y}) \tag{2}$$

where γ is a hyperparameter set to 0.2 controlling the penalties applied to FP and FN during the training of the lower and upper bound heads.

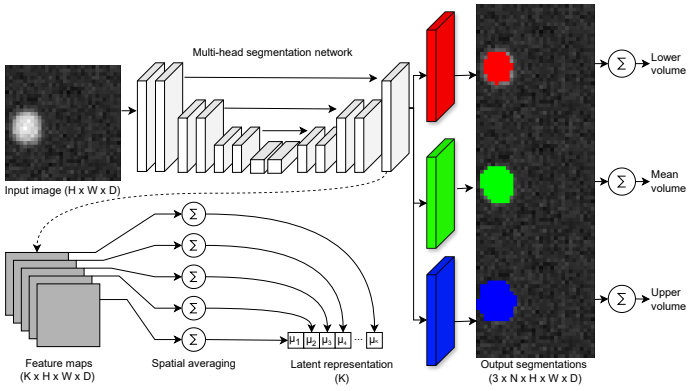


Fig. 1: Illustration of the proposed framework. A multi-head segmentation model predicts three distinct masks for each label: a restrictive mask associated with the lower bound volume (red), a permissive mask associated with the upper bound volume (blue), and a balanced mask for the average volume (green). For Weighted Conformal Prediction, a compressed latent representation is extracted from the penultimate convolution filter.

To ensure that the computed PIs will achieve the user-defined level of coverage on test data, the conformal calibration of intervals can be performed [1]. It operates by first defining a score function $s_i = \max(l_i - Y_i, Y_i - u_i)$. This score is a way to estimate the accuracy of the interval $[l_i, u_i]$ for the true quantity Y_i, with larger scores indicating larger discrepancy. The scores are computed on a

set-aside calibration dataset comprising n pairs of images and associated ground truths. It allows to compute the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-th quantile of the empiral scores: $\hat{q} = \text{Quantile}(s_1, s_2, ..., s_n; \frac{\lceil (n+1)(1-\alpha) \rceil}{n})$. In practice, $\hat{q}$ acts as a corrective factor applied to the PIs so that they encompass the desired fraction of the true volumes on the calibration dataset. At test time, the calibrated PI is computed as follows:

$$\Gamma_\alpha(X_i) = [l_i - \hat{q}, u_i + \hat{q}] \tag{3}$$

As a result, the intervals expand as $\hat{q}$ increases. Supposing the test samples are exchangeable with the calibration samples, the marginal coverage property (Equation 1) is guaranteed.

2.3 Weighted Conformal Prediction to tackle covariate shift

WCP has been proposed to take into account the non-exchangeability of calibration and test data [1,4,26]. The core concept of WCP is to reweight the calibration dataset to more accurately match the test one. This is achieved by estimating the density ratio $w = dP_{\text{test}}/dP_{\text{train}}$ for each calibration and test sample. In practice, writing $X_1, ..., X_n$ the n calibration samples and x the fresh test point, importance weights are computed as:

$$p_i^w(x) = \frac{w(X_i)}{\sum_{i=1}^{N} w(X_j) + w(x)} \tag{4}$$

Essentially, the weight is large when the calibration sample X_i is likely under the test distribution. Then, the corrective value $\hat{q}$ can be reframed as the $1-\alpha$ quantile of the reweighted distribution [1]:

$$\hat{q}(x) = \inf\left\{ s_j : \sum_{j=1}^{n} p_i^w(x) \mathbb{1}\{s_i \leq s_j\} \geq 1 - \alpha \right\} \tag{5}$$

Note that when all weights are equal to $\frac{1}{n+1}$, the standard CP procedure is recovered. A convenient way to estimate this ratio is to use an auxiliary classifier that only requires that unlabeled samples from the test distribution are available during the calibration step [26]. The idea is to train a probabilistic classification model to classify samples between the training and test distributions. That is, writing $X_1, ..., X_n$ and $X_{n+1}, ..., X_{n+m}$ the training and test data points, one can form a classification dataset composed of the pairs $\{X_i, C_i\}$ where $C_i = 0$ for $i = 1, ..., n$ and $C_i = 1$ for $i = n+1, ..., n+m$. Writing $\widehat{p}(x) = \mathcal{P}(C = 1 | X = x)$ the probability predicted by a classifier model trained on the $\{X_i, C_i\}$ dataset that the input sample x belongs to the test distribution, the weight function can be expressed as [25]:

$$\widehat{w}(x) = \frac{\widehat{p}(x)}{1 - \widehat{p}(x)} \tag{6}$$

However, this approach has several limitations. First, it requires access to a sufficient amount of calibration and test samples to allow for a supervised classification strategy. Second, training the classifier is cumbersome when dealing with high-dimensional medical images. In this setting, the dedicated classification approach would be the training of a deep learning Convolution Neural Network (CNN), requiring numerous examples of both classes (calibration and test). Moreover, the training of the auxiliary classifier should be performed during the CP procedure to allow for weight estimation. Incorporating the training of a CNN in the CP procedure is thus highly inefficient. As a conclusion, this classification task is computationally too costly when dealing with 3D medical images. Building on these limitations, we next investigate a more efficient approach making use of the latent representations extracted by the deployed segmentation model.

2.4 Efficient density ratio estimation using latent representations

As training the auxiliary classifier directly from the input images is too costly, more efficient approaches have to be investigated. One idea would be to use a compressed representation of the input image that still preserves important structural information. A lead in this direction is the use of low-dimensional latent representations generated by the segmentation model during the inference process, which has been proven to be a highly efficient summary allowing the detection of out-of-distribution images [7,13,3,28]. Therefore, using compressed latent representations in place of the high-dimensional 3D images seems promising as our end goal is to address covariate shifts. To test this framework, we collect the activations of the penultimate convolution layer. The feature maps have a shape of $K \times H \times W \times D$, where H, W, and D are the spatial dimensions of the 3D image and K the number of kernels in the layer. This feature map is reduced to a compressed vector z of dimension K by performing an averaging over the spatial dimensions (see Figure 1). This approach allows training the auxiliary classifier on compressed representations of the input MRIs, which can be performed efficiently during the WCP procedure.

3 Experiments

3.1 Synthetic dataset with controlled covariate shift

To prove the relevancy of the proposed approach, we first rely on a synthetic setting allowing us to control covariate shift precisely. The task that we propose here is the segmentation of spheres inside cubic volumes of shape $32 \times 32 \times 32$, with the end goal of computing a PI for the volume of each sphere. The covariate of interest here is the signal-to-noise ratio (SNR) between the background of the image and the foreground spheres. A total of 4000 synthetic images are generated. We then split this dataset into an in-distribution (ID) split (3000 images) containing images with high SNRs, and a shifted test dataset (1000

images) containing images with lower SNRs. The ID dataset is further split into training, calibration, and ID test parts, with 1000 images each. Several examples of synthetic images with varying SNRs are presented in Figure 2, along with the densities of SNR in the ID and shifted datasets.

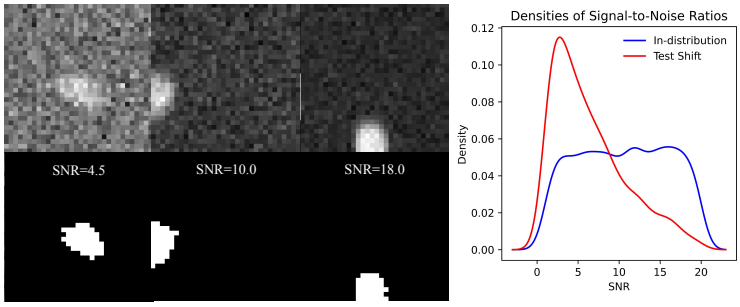


Fig. 2: Left: Examples of synthetic images with varying Signal-to-Noise ratios (SNRs) and associated ground truths. Right: Distribution of SNRs in the in-distribution and shifted synthetic datasets.

3.2 Real-world covariate shift in brain tumor segmentation tasks

To test the framework on real-world medical image data, we address multi-class tumor segmentation in brain MRI. Our dataset consists of glioblastoma and meningioma subjects gathered from the open-source BraTS 2023 datasets [17,21]. Each subject has four MRI sequences: T1-weighted, T2-weighted, FLAIR, and T1 with contrast enhancement. Ground truth masks include necrosis, edematous, and gadolinium-enhancing tumor classes. For covariate shift analysis, we divide subjects into an ID dataset (320 glioblastoma subjects, 748 meningioma subjects, 30%-70% repartition) and a shifted test dataset (873 glioblastoma subjects, 196 meningioma subjects, $82\% - 18\%$ repartition). The covariate shift thus corresponds to the difference in frequencies of each subtype of tumor in the ID and shifted datasets. The ID dataset is further divided into training (568), calibration (250), and ID test (250) subsets.

3.3 Experimental Protocol and Metrics

We use MONAI's [8] Dynamic U-Net [11] as segmentation backbone, modified to have three output heads. The penultimate convolution layer contains 64 kernels, meaning that the extracted latent representations will also have a dimension of 64. The models are trained using Equation 2 and the ADAM optimizer [16] with a learning rate of 2×10^{-4}. After training, PIs are calibrated on the calibration dataset, with a target coverage of 95% for the synthetic task, and 90% for brain tumors as the segmentation is more challenging. Three variants of CP are further compared:

- **Standard** CP corresponds to the setting where calibration samples are associated with identical weights, thus not taking into account potential covariate shifts.
- Weighted CP using Oracle covariates (**W-Oracle**) uses the ground truth covariates to train the auxiliary classifier: SNR for the synthetic images, tumor subtype for brain tumors (0 for glioblastoma, 1 for meningioma).
- Weighted CP using latent representations (**W-Latent**) leverages the compressed latent representations to train the auxiliary classifier.

For W-Oracle and W-Latent, we use a Logistic Regression (LR) model as the auxiliary classifier, trained in a 20-fold cross-validation setting. The probabilities predicted by LR are clipped in the range $[0.01, 0.99]$ to avoid infinite weights (see Equation 6). To estimate the performance of the CP procedures, we monitor the empirical coverage on the test datasets (ID and shifted) as well as the average interval width. We also report the segmentation performance using Dice scores, and the accuracy of the auxiliary LR classifier for W-Oracle and W-Latent. The experiments are reproduced for $R = 250$ trials by shuffling the ID calibration and test datasets. The shifted test dataset is kept identical in each trial.

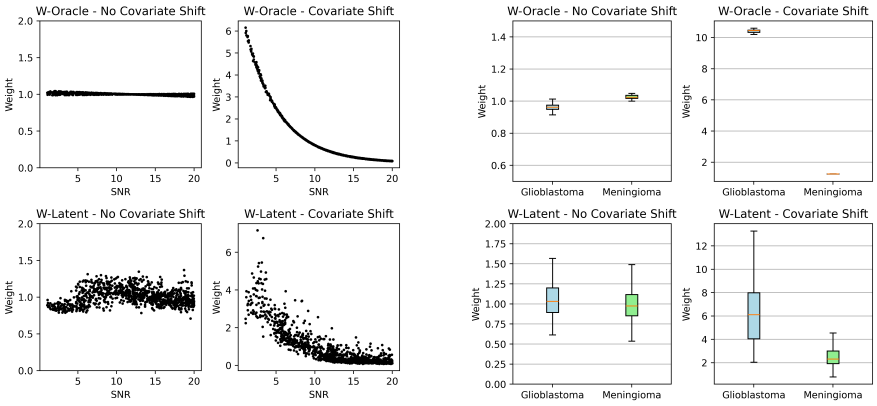
4 Results and Discussion

Tables 1 and 2 present the performance of each CP variant on the synthetic and brain tumor datasets, respectively. In the absence of covariate shifts (ID datasets), W-Oracle and W-Latent closely mimic Standard CP, achieving target coverages with great accuracy (95% for synthetic data, 90% for brain tumors). However, in Shift datasets, Standard CP exhibits miscoverage, with empirical coverages lower than the target level, revealing its inability to handle non-exchangeable data points. W-Oracle and W-Latent alleviate this issue, with W-Oracle recovering the exact target coverages on the synthetic task and the necrosis and edematous brain tumor classes. W-Latent also reduces the coverage gap, although it doesn't exactly recover the target coverages. It can be noticed that this increase robustness is linked with an increase in the average interval width to achieve the target coverage on shifted test data.

A deeper dive into the functioning of WCP is presented in Figure 3. It presents the calibration weights' behavior with and without covariate shifts. When there are no shifts, all weights are close to the unit, mimicking the standard CP procedure. When a covariate shift is observed, higher weights are assigned to calibration samples resembling test samples. For the synthetic task, higher weights are attributed to calibration weights with low SNRs, which are similar to the shifted test samples. For tumor segmentation, higher weights are attributed to glioblastoma subjects, which indeed represent the majority of the shifted test subjects. W-Oracle and W-Latent provide similar trends, although W-Latent is more noisy than the Oracle version.

In conclusion, our WCP framework is effective in tackling covariate shifts in medical image analysis, by addressing covariate shifts either directly or through

latent representations, ensuring the robustness of predictive intervals. However, one limitation of the presented WCP framework is that it can only account for moderate covariate shifts. Otherwise, if the covariate shift is too important between calibration and test samples, the weights will likely diverge (see Equation 6 when $\hat{p}(x)$ converges to 1) which would undermine the accuracy of the WCP procedure.



(a) Synthetic dataset (b) Brain tumor dataset

Fig. 3: Weights of calibration samples estimated by W-Oracle and W-Latent, with and without covariate shift, according to the value of the covariate.

Table 1: Comparison of standard and weighted Conformal Prediction on the synthetic task, for a target coverage of 95%. The mean and standard deviation over 250 trials are presented.

Setting	**CP version**	**Accuracy**	**Coverage** (%)	**Width** (mm^3)	**Dice**
Calib ID	Standard	-	95.11 ± 0.93	86.18 ± 3.00	
vs.	W-Oracle	0.50 ± 0.02	95.01 ± 1.01	86.06 ± 2.95	0.92 ± 0.07
Test ID	W-Latent	0.50 ± 0.01	95.03 ± 0.93	86.45 ± 3.79	
Calib ID	Standard	-	87.47 ± 1.08	94.76 ± 2.77	
vs.	W-Oracle	0.73 ± 0.01	95.22 ± 1.57	153.28 ± 23.52	0.89 ± 0.11
Test Shift	W-Latent	0.72 ± 0.01	93.39 ± 0.90	128.89 ± 11.34	

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)

Table 2: Comparison of standard and weighted Conformal Prediction on multi-class tumor volume estimation for a target coverage of 90%. The mean and standard deviation over 250 trials are presented.

Class	**Setting**	**CP version**	**Accuracy**	**Coverage** (%)	**Width** (mL)	Dice
Necrosis	Calib ID	Standard	-	90.40 ± 2.62	4.1 ± 0.7	0.63 ± 0.33
	vs.	W-Oracle	0.50 ± 0.04	90.09 ± 2.67	3.9 ± 0.7	
	Test ID	W-Latent	0.50 ± 0.03	90.20 ± 2.61	4.0 ± 0.7	
	Calib ID	Standard	-	80.19 ± 2.28	6.2 ± 1.0	0.71 ± 0.28
	vs.	W-Oracle	0.70 ± 0.01	89.53 ± 2.45	13.0 ± 2.6	
	Test Shift	W-Latent	0.81 ± 0.00	88.88 ± 2.88	12.4 ± 3.0	
Edematous	Calib ID	Standard	-	90.48 ± 2.77	18.9 ± 1.8	0.81 ± 0.22
	vs.	W-Oracle	0.50 ± 0.04	90.19 ± 2.85	18.6 ± 1.9	
	Test ID	W-Latent	0.50 ± 0.03	90.29 ± 2.81	18.7 ± 2.0	
	Calib ID	Standard	-	80.56 ± 2.35	26.2 ± 2.2	0.80 ± 0.20
	vs.	W-Oracle	0.70 ± 0.01	89.58 ± 2.44	39.7 ± 5.4	
	Test Shift	W-Latent	0.81 ± 0.00	85.52 ± 3.91	32.9 ± 6.0	
GD-enhancing	Calib ID	Standard	-	90.54 ± 2.57	5.0 ± 0.4	0.88 ± 0.20
	vs.	W-Oracle	0.50 ± 0.04	90.29 ± 2.64	4.9 ± 0.4	
	Test Shift	W-Latent	0.50 ± 0.03	90.36 ± 2.55	4.9 ± 0.4	
	Calib ID	Standard	-	81.29 ± 1.76	7.0 ± 0.4	0.85 ± 0.18
	vs.	W-Oracle	0.70 ± 0.01	87.25 ± 3.39	9.0 ± 1.5	
	Test Shift	W-Latent	0.81 ± 0.00	86.19 ± 3.90	8.6 ± 1.6	

2. Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: The Twelfth International Conference on Learning Representations (2024)
3. Anthony, H., Kamnitsas, K.: On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging pp. 136–146 (2023)
4. Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J.: Conformal prediction beyond exchangeability. The Annals of Statistics **51**(2), 816–845 (2023)
5. Baris, M.M., Celik, A.O., et al.: Role of mass effect, tumor volume and peritumoral edema volume in the differential diagnosis of primary brain tumor and metastasis. Clinical neurology and neurosurgery **148**, 67–71 (2016)
6. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning for differing training and test distributions. In: Proceedings of the 24th international conference on Machine learning. pp. 81–88 (2007)
7. Calli, E., Van Ginneken, B., Sogancioglu, E., Murphy, K.: Frodo: An in-depth analysis of a system to reject outlier samples from a trained neural network. IEEE Transactions on Medical Imaging **42**(4), 971–981 (2022)
8. Consortium, T.M.: Project MONAI, `https://doi.org/10.5281/zenodo.4323059`
9. Contador, J., Pérez-Millán, A., et al.: Longitudinal brain atrophy and csf biomarkers in early-onset alzheimer's disease. NeuroImage: Clinical **32**, 102804 (2021)
10. Ding, X., Wang, Z.J., Welch, W.J.: Subsampling generative adversarial networks: Density ratio estimation in feature space with softplus loss. IEEE Transactions on Signal Processing **68**, 1910–1922 (2020)
11. Futrega, M., Milesi, A., Marcinkiewicz, M., Ribalta, P.: Optimized u-net for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 15–29. Springer (2021)
12. Ghoneem, A., Osborne, M.T., et al.: Association of socioeconomic status and infarct volume with functional outcome in patients with ischemic stroke. JAMA Network Open **5**(4), e229178–e229178 (2022)

13. González, C., Gotkowski, K., Fuchs, M., Bucher, A., Dadras, A., Fischbach, R., Kaltenborn, I.J., Mukhopadhyay, A.: Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. Medical image analysis **82**, 102596 (2022)
14. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B., et al.: Covariate shift by kernel mean matching. Dataset shift in machine learning **3**(4), 5 (2009)
15. Kanamori, T., Hido, S., Sugiyama, M.: A least-squares approach to direct importance estimation. The Journal of Machine Learning Research **10**, 1391–1445 (2009)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. LaBella, D., Adewole, M., Alonso-Basanta, M., Altes, T., Anwar, S.M., Baid, U., Bergquist, T., Bhalerao, R., Chen, S., Chung, V., et al.: The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. arXiv preprint arXiv:2305.07642 (2023)
18. Lambert, B., Forbes, F., Doyle, S., Dojat, M.: Triadnet: Sampling-free predictive intervals for lesional volume in 3d brain mr images. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 32–41. Springer (2023)
19. Lei, J., Rinaldo, A., Wasserman, L.: A conformal prediction approach to explore functional data. Annals of Mathematics and Artificial Intelligence **74**, 29–43 (2015)
20. Mattiesing, R.M., Gentile, G., et al.: The spatio-temporal relationship between white matter lesion volume changes and brain atrophy in clinically isolated syndrome and early multiple sclerosis. NeuroImage: Clinical **36**, 103220 (2022)
21. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
22. Nam, H., Sugiyama, M.: Direct density ratio estimation with convolutional neural networks with application in outlier detection. IEICE TRANSACTIONS on Information and Systems **98**(5), 1073–1079 (2015)
23. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13. pp. 345–356. Springer (2002)
24. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International workshop on machine learning in medical imaging. pp. 379–387. Springer (2017)
25. Sugiyama, M., Suzuki, T., Kanamori, T.: Density ratio estimation: A comprehensive review (statistical experiment and its related topics). RIMS Kokyuroku **1703**, 10–31 (2010)
26. Tibshirani, R.J., Foygel Barber, R., Candes, E., Ramdas, A.: Conformal prediction under covariate shift. Advances in neural information processing systems **32** (2019)
27. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
28. Woodland, M., Patel, N., Al Taie, M., P. Yung, J., J. Netherton, T., B. Patel, A., K. Brock, K.: Dimensionality reduction for improving out-of-distribution detection in medical image segmentation. International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging pp. 147–156 (2023)

29. Xue, Z., Yang, F., Rajaraman, S., Zamzmi, G., Antani, S.: Cross dataset analysis of domain shift in cxr lung region detection. Diagnostics **13**(6), 1068 (2023)