Multi-layer Aggregation as a key to feature-based OOD detection

Benjamin Lambert, Florence Forbes, Senan Doyle, Michel Dojat

Multi-layer Aggregation as a key to feature-based OOD detection

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. Deep Learning models are easily disturbed by variations in the input images that were not observed during the training stage, resulting in unpredictable predictions. Detecting such Out-of-Distribution (OOD) images is particularly crucial in the context of medical image analysis, where the range of possible abnormalities is extremely wide. Recently, a new category of methods has emerged, based on the analysis of the intermediate features of a trained model. These methods can be divided into 2 groups: *single-layer* methods that consider the feature map obtained at a fixed, carefully chosen layer, and *multi-layer* methods that consider the ensemble of the feature maps generated by the model. While promising, a proper comparison of these algorithms is still lacking. In this work, we compared various feature-based OOD detection methods on a large spectra of OOD (20 types), representing approximately 7800 3D MRIs. Our experiments shed the light on two phenomenons. First, *multi-layer* methods consistently outperform *single-layer* approaches, which tend to have inconsistent behaviour depending on the type of anomaly. Second, the OOD detection performance highly depends on the architecture of the underlying neural network.

Keywords: Uncertainty · Deep learning · Anomaly Detection · Medical images analysis

1 Introduction

Out-of-distribution (OOD) images correspond to samples that are significantly different from the ones observed during training. Deep Learning (DL) models tend to behave inconsistently for this type of inputs, making OOD image detection crucial to avoid hidden model deficiencies. It is especially required in real-world automated pipelines, where input images may not be visually inspected before running the analysis. In the context of medical-images analysis, a large variety of phenomenons in the input images can impact a model and lead to unpredictable responses: noise, artifacts, variations in the imaging acquisition protocol and device, or pathological cases that were not included in the initial training dataset. Various methods were proposed for their detection, which can roughly be divided into two different categories [3]: methods that

build a model specifically dedicated to OOD detection; and methods that rely on the uncertainty or intermediate activations of a task-specific model (e.g image segmentation) to detect abnormal inputs.

Within the first category, the most straightforward approach is to build a classifier to directly detect OOD images. For this, a Convolutional Neural Network (CNN) can be trained in a supervised manner, thus requiring the construction of an annotated dataset containing various types of real-world OOD [4]. Unsupervised Anomaly Detection (UAD) proposes to model the appearance of normal images by training an Auto-Encoder network (AE) to reconstruct in-distribution (ID) samples. At test-time, reconstruction is expected to be degraded for OOD samples, allowing for their detection [10].

Among the second category, uncertainty-based methods propose to detect OOD inputs directly from the outputs of an existing neural network. They rely on the hypothesis that the uncertainty of the deployed model should be high in the presence of a train-test mismatch, allowing its detection. A standard method consists of producing a set of diverse and plausible predictions for the same input image, with MC dropout [9], Deep Ensemble [21] or Test Time Augmentation [33] being popular approaches. Uncertainty can then be estimated by computing the variance among the predictions. Alternatively, feature-based methods propose to analyse the intermediate activations of an existing model to detect OOD inputs [28]. It is based on the assumption that the hidden activations of the model should be different for an ID image compared to an OOD image. A taxonomy of these feature-based methods is possible based on the number of layers used for OOD detection. *Single-layer* methods only target one specific convolutional layer. In the context of medical image segmentation, popular choices are the end of the encoder [11] or the penultimate convolutional layer [16,7]. *Multi-layer* methods are an extension of the former that consider the entire set of convolutional layers in the trained model for OOD detection [5]. Although gaining popularity, a proper comparison of these algorithms is still lacking. The contributions of our work are as follows:

- We develop a large MRI segmentation benchmark comprising 20 different OOD datasets of various types and strengths, representing 7796 3D MRI volumes. We use this benchmark to compare 5 different feature-based OOD detectors as well as one uncertainty baseline.
- We adapt single-layer methods to a multi-layer fashion to demonstrate the potential performance gain that can be obtained with this enhancement.

2 Compared methods

2.1 Feature-based methods

Feature-based methods rely on a trained segmentation model and follow the same principle. First, a set of feature maps $F_i \in \mathbb{R}^{N \times H_i \times W_i \times D_i}$ is collected from a ID dataset for one convolution layer i (single-layer methods) or all convolution

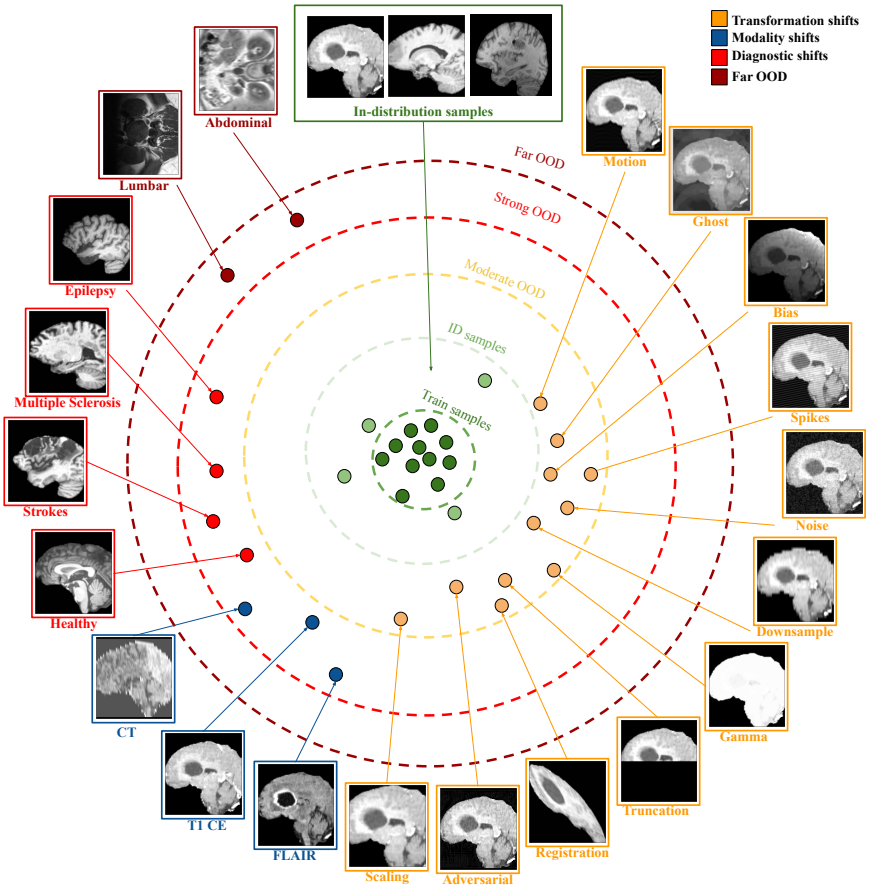

Fig. 1. Illustration of ID and OOD data used in the experiments.

layers (multi-layer methods). Here, N correponds to the number of convolutional filters in the i-th layer, and $H_i \times W_i \times D_i$ to the spatial dimensions of the feature map. Second, at inference time, a metric is computed to estimate the distance between the test features and the ID features to detect OOD samples.

Spectral signatures [16] was proposed as a single-layer method focusing on the features obtained at the penultimate convolutional layer. Features are flattened to a 2D matrix $F_{2D} \in \mathbb{R}^{N \times HWD}$, and its singular values S are calculated. The spectral signature is then taken as $\phi = \frac{log(S)}{\|log(S)\|_2}$. To detect OOD at test time, the distances d_j between the signature of the test image ϕ^i_{test} and the signatures of a set of ID samples ϕ^j_{ID} is obtained by using the Euclidean distance. The final proposed OOD score corresponds to the minimum of the d_j distances.

Prototypes [7] is a single-layer method that operates from the penultimate layer features and the segmentation masks predicted by the segmentation network. To obtain a prototype for a specific class and input image, features are multiplied with the binarized class mask, and average pooling is applied on the masked features. This yields to prototypes $P \in \mathbb{R}^{N \times C}$, C being the number of segmented classes. An average ID prototype P_{ID} is finally obtained by averaging the prototypes collected on the ID dataset. At test time, the OOD score is taken as the cosine dissimilarity between P_{ID} and the test image prototype P_j.

The Mahalanobis Distance (MD) was recently investigated in 2 distinct studies for OOD detection in 3D medical images. In [11], authors focus on the features from the end of the encoder part of the network. They apply consecutive average pooling until the number of elements M in the feature maps falls below a defined threshold of $1e4$, and then flatten it to obtain 1D vectors $z_i \in \mathbb{R}^M$, for each ID image i. From these vectors, they compute the parameters of a multivariate Gaussian: the mean $\mu \in \mathbb{R}^M$ and covariance $\Sigma \in \mathbb{R}^{M \times M}$. At test time, the MD is computed given the fitted Gaussian and the test image feature representation. We refer to this single-layer method as *MD Pool*. A similar approach is implemented in the multi-layer Free Rejection of Out-of-Distribution (FRODO) approach [5]. This work differs in two ways: first, they directly compute the average of the feature map over the spatial dimensions ($H \times W \times D$) instead of applying average poolings. Second, they fit a multivariate Gaussian independently for each convolutional layer and compute the final OOD score as the average of each layer score.

One-class SVM (OCSVM) is an unsupervised algorithm for OOD detection that can be trained using only ID samples. It aims at finding the optimal boundary around the expected (ID) data. At test time, the distance to the boundary can be used as an OOD score, with ID sample being attributed with negative distances, and OOD samples with positive distances. Similar to [34], we fit a OCSVM per convolution based on the averaged layer activations obtained from the ID images. At test time, each OCSVM produces a score, and the final OOD score is taken as the maximum of these scores.

2.2 Adapting single-layer methods to multi-layer methods

To assess the contribution of multi-layer aggregation to OOD detection, we propose to adapt single-layer methods (Spectrum, Prototypes and MD Pool) to multi-layer style. To achieve this, we replicate the OOD score computation step for each convolutional layer independently, yielding to *layer-wise* scores l_i. As in FRODO, the final multi-layer score L_{multi} is taken as the average of the scores of the individual layers:

$$L_{multi} = \frac{1}{N} \sum_{i=1}^{N} l_i \tag{1}$$

2.3 Uncertainty baseline

To compare feature-based OOD detectors with a more traditional uncertainty methodology, we implement a MC dropout baseline [9]. In MC dropout, the dropout layers of the segmentation model are kept activated at test-time and $N = 20$ predictions are repeatedly sampled for each input image. The average voxel-wise variance among the MC samples is taken as uncertainty estimate at test time.

3 Material and Method

3.1 In-distribution Datasets

Our work relies on the open-source BraTS 2021 dataset [2] containing 1251 patients. The dataset initially includes four MRI sequences for each patient with four ground truth segmentation masks: the background, the necrotic tumor core, the edematous and the GD-enhancing tumor. We choose to focus on T1w sequences as this sequence is common and sufficient for experiment with multiple OOD settings. We also simplify the prediction task by focusing on the segmentation of the *whole tumor core*, concatenation of all tumors sub-classes. The dataset is randomly split into a training fold (651), a calibration fold (200) used to fit the OOD detectors and a testing fold (400) (referred to as *Test ID* in the following). Additionally, we propose to include *control* samples in our protocol, representing images that share the same properties than the training samples (same modality, organ and pathology), but that were acquired in a different imaging center. An effective model *should* be able to generalize to these images and thus, the OOD detector should identify them as ID samples to prevent false alarms. We thus propose to use the LUMIERE glioblastoma dataset [30] as a *Control* dataset, from which we select 74 T1-w pre-operative brain MRI. Figure 1 illustrates the data used in the different experiments.

3.2 Out-of-distribution Datasets

Following the categorization of [11], we propose to investigate *Transformation*, *Diagnosis* and *Far* OODs, as well as a new proposed setting, *Modality* shifts.

Transformation shifts. Finding real images with a controlled amount of artifacts to allow evaluation of OOD detection methods is difficult. We therefore generate realistic synthetic artifacted images from the set of *Test ID* images [8,11]. We used the TorchIO Data Augmentation library [19] to generate *Bias*, *Motion*, *Ghost*, *Spikes*, *Downsample*, *Noise*, and *Scaling* artifacts. We add a set of novel transformations: the *Registration* that applies noise to the registration matrix to simulate an erroneous registration, the *Gamma* that applies extreme gamma modification to the image to mimick errors in the intensity normalization step, and the *Truncation* that crops half of the brain. Finally, we also implement *Adversarial* attacks, using the popular Fast Gradient Sign Method (FGSM) [12].

Diagnosis shifts. DL segmentation models are usually trained with images showing a single pathology (e.g brain tumor or strokes). However, once deployed, the model can be confronted with images exhibiting unseen anomalies, which can lead to incorrect predictions. To test OOD detection methods on this scenario, we use T1w brain MRI with various diseases: 170 subjects from the White Matter Hyperintensities (WMH) 2017 challenge [20], 655 subjects from the ATLAS-2 brain stroke dataset [22] and 162 subjects from the EPISURG dataset [27] containing epileptic subjects who underwent resective brain surgery. We also use 582 T1-w MRIs from healthy and young subjects from the IXI dataset [1].

Modality shifts. Medical images are usually stored in DICOM formats, whose meta-data (headers) may be incorrectly filled [13]. As a result, mismatches between the expected input modality (e.g T1w) and the test image modality (e.g CT or T2w) may be undetected. We construct 3 different *Modality* shift OOD datasets. First, we use the FLAIR and T1ce sequences corresponding to the 400 test subjects. Second, we extract 437 brain CT-scans from the CQ500 dataset, exhibiting intracranial hemorrhage or cranial fractures.

Far OOD corresponds to images that show little to no similarity with the ID samples. We use 2 non-brain T1w MRI datasets, respectively 80 abdominal MRI from the CHAOS dataset [17] and 515 images from the Lumbar Spine MRI dataset [25], as far OOD samples.

3.3 Influence of the segmentation model architecture

Feature-based OOD detection methods rely on the hypothesis that the activations of the trained segmentation models are representative of the conformity of the input sample. To verify if this holds true for any segmentation model, we use the MONAI library [6] to train 6 different segmentation models: an Attention UNet (AttUNet) [26], a Residual UNet (ResUNet) [18], a Dynamic UNet (DynUNet) [15], a UNet++ [35], a VNet [24] and a Transformer-based model, namely the UneTR [14]. All models are trained with the Dice loss [24], instance normalization [31], a 3D dropout [29] rate of 20%, and a batch size of 1, using the ADAM optimizer [19] with a learning rate of $2e-4$.

3.4 Evaluation Setting

We cast OOD detection as a binary classification problem, where ID samples correspond to the positive class and OOD samples to the negative class. Each

Table 1. Number of model parameters (in millions) as well as the average segmentation performance (Dice) on the *Test ID* dataset, for each segmentation model.

	AttUNet	ResUNet	VNet	UNet++	DynUNet	UNETR
Nb. parameters	5.0	4.8	11.4	7.0	16.5	102
Test ID Dice ↑	**.83**	.81	.81	.81	.82	.76

Table 2. OOD detection performance (AUROC) for each dataset, obtained with the AttUNet segmentation model for the features and uncertainty approaches. We also report the number of samples in each dataset (N) and Dice score when the ground truth for brain tumors is available. S: single-layer. M: multi-layer.

			Spectrum		Prototypes		MD Pool		Frodo	SVM	MC
Dataset	N	Dice (↑)	S	M	S	M	S	M	M	M	-
Test ID	400	.83	-	-	-	-	-	-	-	-	
Control	74	.85	0.56	0.84	0.36	0.48	0.42	0.40	0.45	0.38	0.50
Motion	400	.82	0.66	**0.97**	0.48	0.49	0.76	0.71	0.80	0.60	0.56
Ghost	400	.80	0.66	0.57	0.46	0.48	**0.88**	0.85	0.85	0.60	0.63
Bias	400	.78	**1.00**	**1.00**	0.86	0.82	**1.00**	0.98	**1.00**	**1.00**	0.77
Spikes	400	.79	0.90	**1.00**	0.69	0.82	0.86	**1.00**	**1.00**	**1.00**	0.63
Noise	400	.81	0.84	**1.00**	0.57	0.61	0.80	**1.00**	**1.00**	**1.00**	0.62
Downsample	400	.82	0.69	**0.97**	0.48	0.50	0.55	0.55	0.73	0.60	0.53
Gamma	400	.31	**1.00**	**1.00**	0.95	0.96	0.98	0.98	**1.00**	**1.00**	0.97
Truncation	400	.61	0.99	0.99	0.86	0.83	**1.00**	0.99	**1.00**	0.99	0.63
Registration	400	.01	**1.00**	**1.00**	0.93	0.88	**1.00**	**1.00**	**1.00**	**1.00**	0.84
Adversarial	400	.58	0.77	0.69	0.42	0.50	0.89	0.95	**1.00**	0.97	0.80
Scaling	400	.77	0.99	**1.00**	0.91	0.91	0.99	0.99	**1.00**	**1.00**	0.72
Transform	**4400**	-	0.86	0.92	0.70	0.71	0.89	0.92	**0.95**	0.89	0.70
FLAIR	400	.10	0.99	0.99	0.40	0.91	0.99	0.97	**1.00**	0.99	**1.00**
T1Ce	400	.69	**0.97**	0.96	0.80	0.78	0.92	0.92	**0.97**	0.90	0.90
CT	437	-	0.99	**1.00**	0.36	0.89	0.99	0.98	**1.00**	0.99	**1.00**
Modality	**1237**	-	0.98	0.98	0.54	0.86	0.97	0.96	**0.99**	0.97	0.97
Healthy	577	-	0.55	0.99	0.89	**1.00**	0.12	0.57	0.97	0.98	0.92
Strokes	655	-	0.58	0.88	0.82	**0.97**	0.37	0.54	0.86	0.82	0.81
WMH	170	-	0.73	0.96	0.89	**0.98**	0.30	0.54	0.96	0.93	0.76
Epilepsy	162	-	0.68	0.94	0.89	**0.99**	0.41	0.60	0.88	0.83	0.82
Diagnosis	**1564**	-	0.59	0.92	0.86	**0.98**	0.28	0.58	0.92	0.90	0.85
Lumbar	515	-	**1.00**	**1.00**	0.86	0.96	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Abdominal	80	-	**1.00**	**1.00**	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Far OOD	**595**	-	**1.00**	**1.00**	0.88	0.96	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Overall	**7796**	-	0.84	0.93	0.72	0.81	0.79	0.86	**0.95**	0.91	0.80

method produces a score for each OOD sample, which is compared with the scores obtained on the ID data in order to compute AUROC classification scores. We also report the segmentation performance using the Dice score, when the ground truth for brain tumors is available.

4 Results and Discussion

The average segmentation performance of each tested segmentation backbone is presented in Table 1, with AttUnet being the best performer. OOD detection performances of each method are presented in Table 2 using the AttUnet as backbone for feature and uncertainty-based methods. Finally, Figure 2 presents the *Overall* OOD detection performance with respect to the neural architecture.

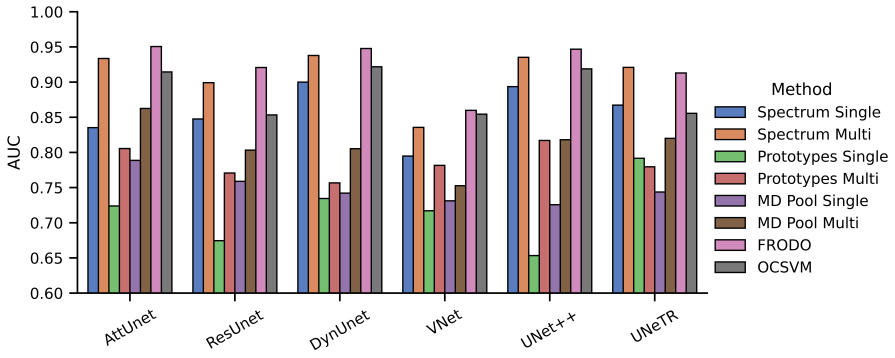

Fig. 2. *Overall* OOD detection performance (AUROC score) of feature-based approaches with respect to the segmentation model architecture.

All OOD detectors achieve high detection accuracy on *Far OOD* with mixed performances on other OOD types, showing that restricting to extreme OOD examples is insufficient to robustly validate a method. The best performer is FRODO, achieving a perfect detection of non-conform inputs (AUROC=1.00) in 12 out of 20 settings, followed by the multi-layer implementation of Spectrum, and OCSVM. Overall, multi-layer methods outperform their single-layer version. For the AttUnet, this enhancement allows an increase of the *Overall* AUROC score of 10.7% for Spectrum, 12.5% for Prototypes and 8.9% for MD Pool. Single-layer methods exhibit variable performances depending on the OOD type, in accord with observations on 2D image classification [34]. Finally, the MC dropout baseline has mixed performance in our benchmark, being outperformed by all multi-layer feature-based detectors.

The rankings of OOD methods is roughly the same with different segmentation architecture (Figure 2), with FRODO and Spectrum Multi-layer being the two top-performers. Interestingly, converting single-layer methods to multi-layer methods is beneficial for all architectures, with a gain on the AUROC for 6 out of 6 backbones for Spectrum and MD Pool, and 5 out of 6 backbones for Prototypes. However, the global *Overall* OOD detection performance is variable depending on the segmentation architecture, e.g. FRODO achieves an AUROC score of 0.95 or of 0.86 when implemented on an AttUnet or VNet respectively. This indicates that certain popular medical image segmentation architectures are more prone to *feature collapse* [32], mapping OOD images to ID feature representations. Several strategies have been proposed in the context of 2D image classification to alleviate this issue, such as adding Gradient Penalty [32], Lipschitz constraints [23], or a reconstruction term in the loss [28]. These methods aims at enforcing a discriminative feature space for OOD detection, requiring changes in the training paradigm of the model, possibly resulting in sub-optimal predictive performances. To summarize, our main findings are:

– Feature-based methods monitoring the activation of *all* convolution layers are more performant and robust than methods only targeting a single layer, whose performance is highly variable depending on the type of OOD.
– The performance of these methods is dependent on the underlying segmentation architecture, with some of them being more prone to feature collapse, undermining the sensibility of OOD detection.

References

1. The ixi brain dataset. `https://brain-development.org/ixi-dataset/`
2. Baid, U., Ghodasara, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Berger, C., et al.: Confidence-based out-of-distribution detection: a comparative study and analysis. Unsure 2021 pp. 122–132 (2021)
4. Bottani, S., et al.: Automatic quality control of brain t1-weighted magnetic resonance images for a clinical data warehouse. Medical Image Analysis **75**, 102219 (2022)
5. Çallı, E., Van Ginneken, B., et al.: 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)
6. Cardoso, M.J., Li, W., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
7. Diao, Z., et al.: A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity. Knowledge-Based Systems **246**, 108739 (2022)
8. Fuchs, M., Gonzalez, C., Mukhopadhyay, A.: Practical uncertainty quantification for brain tumor segmentation. Medical Imaging with Deep Learning (MIDL) (2021)
9. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. ICML **48**, 1050–1059 (2016)
10. Gong, D., Liu, L., et al.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. ICCV pp. 1705–1714 (2019)
11. González, C., Gotkowski, K., et al.: Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. Medical Image Analysis **82**, 102596 (2022)
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. 3rd International Conference on Learning Representations, ICLR 2015 (2015)
13. Gueld, M.O., Kohnen, M., et al.: Quality of dicom header information for image categorization. Medical imaging 2002: PACS and integrated medical information systems: design and evaluation **4685**, 280–287 (2002)
14. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
15. Isensee, F., Jaeger, P.F., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

16. Karimi, D., Gholipour, A.: Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Transactions on Artificial Intelligence (2022)
17. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
18. Kerfoot, E., Clough, J., et al.: Left-ventricle quantification using residual u-net. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges Workshop, Held in Conjunction with MICCAI 2018 pp. 371–380 (2019)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Kuijf, H.J., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging **38**(11), 2556–2568 (2019)
21. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems 30 pp. 6402–6413 (2017)
22. Liew, S.L., Lo, B.P., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
23. Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., Lakshminarayanan, B.: Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Advances in Neural Information Processing Systems **33**, 7498–7512 (2020)
24. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV) pp. 565–571 (2016)
25. Natalia, F., Meidia, H., et al.: Development of ground truth data for automatic lumbar spine MRI image segmentation. HPCC/SmartCity/DSS 2018 pp. 1449–1454 (2018)
26. Oktay, O., Schlemper, J., et al.: Attention u-net: Learning where to look for the pancreas. Medical Imaging with Deep Learning (MIDL) (2018)
27. Pérez-García, F., Rodionov, R., et al.: Simulation of brain resection for cavity segmentation using self-supervised and semi-supervised learning. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020. pp. 115–125. Springer (2020)
28. Postels, J., et al.: On the practicality of deterministic epistemic uncertainty. ICML **162**, 17870–17909 (2022)
29. Srivastava, N., Hinton, G., et al.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
30. Suter, Y., Knecht, U., et al.: The lumiere dataset: Longitudinal glioblastoma mri with expert rano evaluation. Scientific data **9**(1), 768 (2022)
31. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
32. Van Amersfoort, J., et al.: Uncertainty estimation using a single deep deterministic neural network. International conference on machine learning pp. 9690–9700 (2020)
33. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)

34. Wang, H., Zhao, C., et al.: Layer adaptive deep neural networks for out-of-distribution detection. Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference pp. 526–538 (2022)
35. Zhou, Z., Siddiquee, R., et al.: Unet++: A nested u-net architecture for medical image segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA) 2018, Held in Conjunction with MICCAI 2018 pp. 3–11 (2018)