Proper Scoring Loss Functions Are Simple and Effective for Uncertainty Quantification of White Matter Hyperintensities

Proper scoring loss functions are simple and effective for uncertainty quantification of White Matter Hyperintensities

Ben Philps[1][0009−0006−7999−2584], Maria del C. Valdes Hernandez[2][0000−0003−2771−6546], and Miguel Bernabeu Llinares[3,4][0000−0002−6456−3756]

[1] School of Informatics, University of Edinburgh, Edinburgh, UK
B.R.Philps@sms.ed.ac.uk
[2] Department of Neuroimaging Sciences, Centre for Clinical Brain Sciences, University of Edinburgh , Edinburgh, UK
[3] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK
[4] The Bayes Centre, University of Edinburgh, Edinburgh, UK

Abstract. Uncertainty quantification is an important tool for improving the trustworthiness and clinical usefulness of medical imaging segmentation models, and many techniques exist for quantifying segmentation uncertainty. However, popular segmentation losses such as Dice loss lead to poorly calibrated models and silent failure in uncertainty maps. We compare common proper scoring rule based losses, which encourage well-calibrated models, to Dice loss and calibrated Dice loss variants, for white matter hyperintensity (WMH) segmentation in FLAIR and T1w MRI. We show that scoring rules yield strong performance (e.g., Spherical-TopK: Dice of 0.763, vs 0.717 for Dice loss) and low WMH instance detection failure rate in axial slices (Logarithmic yields 11% missing instances in the uncertainty map vs 28% for Dice). Furthermore, proper scoring rule methods do not exhibit the performance degradation in calibration error and WMH burden prediction of Dice loss in low WMH burden patients. Finally, we show temperature scaling is insufficient to overcome the drawbacks of Dice loss.

Keywords: Scoring Functions · Uncertainty Quantification · White Matter Hyperintensities.

1 Introduction

Uncertainty quantification is an important consideration for improving the trust and utility of existing AI tools, particularly for downstream clinical tasks in medical segmentation settings. However, the most common choices of loss function used to train medical imaging segmentation methods can result in poorly calibrated models [28]. Calibration measures how well a model's predictions of an outcome match the probability of that outcome occurring in the real world.

In this work, we assess the impact of loss function choice on the calibration of segmentation models for white matter hyperintensities (WMH). WMH are a clinical feature common in brain magnetic resonance imaging (MRI) of elderly individuals, and a neuroradiological feature of small vessel disease (SVD) [26]. WMH pose a difficult segmentation challenge, due to their spatial and structural heterogeneity as well as inherent aleatoric uncertainty due to unclear borders[19] and subjectivity in identifying deep isolated small WMH clusters. Therefore, various methods have been developed to segment WMH[3], with neural networks (NNs) showing strong performance[5]. Our contributions are: 1) Proper scoring functions (a measure for evaluating predictive distributions that reward calibrated probabilities) consistently yield well calibrated models and surprisingly yield stronger Dice scores for WMH segmentation than the common Dice loss. 2) We demonstrate that Dice loss is poorly calibrated and attempts to correct Dice loss are insufficient, yielding large absolute volume differences (AVD) in patients with low WMH burden. 3) Proper scoring rules yield uncertainty maps that detect a greater proportion of small WMH instances in axial slices.

2 Background

Scoring functions are a class of functions that reward a model for outputting probabilities consistent with the true event probabilities [13]. Specifically, they provide a numerical score $S(P, Q)$ for a predictive distribution P under a target distribution Q. Scoring functions are proper if $S(Q, Q) \geq S(P, Q)$ for all P, Q. Scoring functions are often used for ex-post evaluation of models to determine which model is best calibrated. However, we can use these directly as loss functions [7] to reward a well calibrated segmentation model by minimising the loss $\mathcal{L}_S = \frac{1}{N}\sum_{i=1}^{N} 1 - S(p_{ic}, y_{ic})$. Here p_{ic} is the predicted probability of voxel i belonging to class c, y_i is a one-hot encoded vector of the annotation label, and N is the number of voxels and C the number of classes. The three most common strictly proper scoring functions are the Brier score: $\frac{1}{C}\sum_{c=1}^{C}(p_{ic} - y_{ic})^2$, the Logarithmic score: $1 + \ln(p_i \cdot y_i)$ and the Spherical score: $\frac{p_i \cdot y_i}{||p_i||}$. Notably, choosing logarithmic score is equivalent to minimising cross-entropy loss. However, for segmentation, losses that aim to optimize for the intersection over union are popular, particularly in settings where there is a high class imbalance [17]. Most popular is the Dice loss function, a soft differentiable minimization of $1-$ the Dice score (Table 1). However Dice is well known to yield highly overconfident predictions [28]. One approach for improving calibration is to selectively penalize the overconfident predictions by exponentially weighting the false positive and false negative terms in the denominator of the Dice loss (Dice++ Loss[28]):

$$\mathcal{L}_{\text{Dice++}} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{i=1}^{N} p_{ic}y_{ic}}{2\sum_{i=1}^{N} p_{ic}y_{ic} + (\sum_{i=1}^{N}(1-p_{ic})y_{ic})^{\gamma} + (\sum_{i=1}^{N} p_{ic}(1-y_{ic}))^{\gamma}} \tag{1}$$

As we increase the focal parameter γ above 1, overconfident predictions incur larger penalties, with larger γ controlling the severity of the penalty. Note that $\gamma = 1$ is equivalent to the Dice loss. Alternatively, we can attempt to adapt $\mathcal{L}_{\mathcal{S}}$ to a highly imbalanced task by focusing our loss only on voxels that have low confidence (and therefore likely inaccurate predictions) for the target class. TopK loss [27] computes the loss for only the top $k\%$ lowest confidence voxels for the target class, removing high confience voxels:

$$\mathcal{L}_{\text{TopK}} = \frac{1}{\sum_{i=1}^{N} \mathbb{1}(y_{ic} = 1 \wedge p_{ic} < t)} \sum_{i=1}^{N} 1 - \mathbb{1}(y_{ic} = 1 \wedge p_{ic} < t) S(p_i, y_i) \quad (2)$$

where S is a proper scoring function, $\mathbb{1}$ is the indicator function and we adjust t each training batch such that we include only $k\%$ voxels.

Finally, we can treat calibration not as an ex-ante design consideration, but an ex-post correction to our model. A simple and robust approach is temperature scaling[8]. Temperature scaling utilizes a single parameter, the temperature κ, to soften (increase the entropy) of the model softmax distribution. Given a logit vector η_i from our model, we compute $p_i = \sigma(\eta_i/\kappa)$, where σ is the softmax operator. Noteably, $\lim_{\kappa \to \infty} p_i = \frac{1}{C}$. Since proper scoring functions reward correct calibration, we can tune κ on the validation data by optimizing a scoring function of our choosing, usually the logarithmic score.

Ultimately, our goal is to incorporate model uncertainty into downstream clinical tasks (such as assessing SVD severity). Numerous techniques exist to quantify the uncertainty of NN predictions [30], [1], [20], such as ensembling [25], sampling the latent space of conditional variational autoencoders [10], or modelling the covariance between voxels [21]. To assess the impact of loss function choice, we adopt a softmax entropy map over each image $\mathbb{H}(p_i)$ as a simple baseline. Softmax entropy captures the aleatoric uncertainty (the irreducible uncertainty inherint to the data) present in in-distribution samples [22]. For WMH, aleatoric uncertainty is introduced due to inter-rater disagreement and subjective definitions of WMH, such as at the borders of WMH where the exact boundary is usually unclear. At inference time we may select a threshold τ, with all voxels with uncertainty $> \tau$ flagged as uncertain during downstream tasks.

3 Materials and Methods

3.1 Dataset and Preprocessing

We validate each loss function on the WMH Segmentation Challenge 2017 dataset, [5]. This dataset provides a training dataset of 60 subjects and a test dataset of 110 subjects, collected across multiple different institutions and acquisition protocols. Supplementary A provides details. The images are provided with the following preprocessing: sagittal 3D FLAIR images are reoriented to axial and resampled to slice thickness of 3mm; T1w images are registered to the FLAIR

[5] WMH Challenge Dataset is publicly available at `https://wmh.isi.uu.nl/data/`

using Elastix [16]; Bias field inhomogeneities for FLAIR and T1w are corrected with SPM12 [2]. We further applied the following preprocessing: brain extraction using ROBEX [11]; resampling images to $1.00 \times 1.00 \times 3.00$ voxel size (using cubic spline interpolation for images and nearest neighbours for labels); Z-score normalization to brain tissue, using intensities in the 5-95th percentile to calculate the mean and variance; centre crop/pad all axial slices to 224×192 voxels. Each voxel is labeled as either Background, WMH, or Other Pathology; however the Other Pathology class is not included in evaluation statistics.

3.2 Evaluation Metrics

Table 1: Evaluation metrics for assessing model performance. For **Recall** and **F1**-score, the metrics are defined across individual lesions, which are calculated as connected components in the predicted and ground truth WMH segmentations. Specifically, N refers to the number of connected components, where N_{TP} refers to the number of lesions where a predicted lesion overlaps with a true lesion by at least one voxel. TP refers to the number of true positive voxels. $V_{\hat{y}}$ and V_y are the predicted and true WMH volume respectively, **AVD** = Absolute Volume Difference. $\hat{d}(\hat{y}, y) = \max_{\hat{y}_i \in \hat{y}} \min_{y_i \in y} d(\hat{y}_i, y_i)$ where d is a distance function. **HD95** = Modified Hausdorff distance (95th percentile).

Dice(↑)	**HD95**(↓)	**AVD**(↓)	**Recall**(↑)	**F1**(↑)
$\frac{2TP}{2TP+FP+FN}$	$\max\{\hat{d}(\hat{y}, y), \hat{d}(\hat{y}, y)\}$	$\frac{\lvert V_{\hat{y}} - V_y \rvert}{V_y}$	$\frac{N_{TP}}{N_{TP}+N_{FN}}$	$\frac{2N_{TP}}{2N_{TP}+N_{FN}+N_{FP}}$

To evaluate the segmentation performance of each loss function we employ the evaluation metrics from the WMH Challenge, detailed in Table 1.To assess calibration and usefulness of the softmax entropy uncertainty map we employ three metrics (Table 2). We use Expected Calibration Error (ECE) to measure calibration performance. However, a low ECE score across the test dataset can hide highly over or under confident segmentations on individual images[14], furthermore WMH Dice and lesion F1 score degrades for images with low WMH burden [5]. Therefore we also examine how ECE varies by individual according to WMH burden. Next, the uncertainty error overlap (UEO) metric assesses overlap between uncertainty and segmentation error rewarding uncertainty that is well localized (maximises the Dice metric between error and uncertainty). The unified uncertainty score (UUS) [20] assesses the quality of the remaining segmentation after voxels with uncertainty $> \tau$ are removed. UUS rewards methods that remove incorrectly segmented voxels (with higher Dice score when computed only on the remaining voxels) while penalizing the removal of correctly segmented voxels. WMH do not represent distinct, clearly identifiable anatomic abnormalities, but instead represent foci of (often subtle) white matter changes

Table 2: Uncertainty quantification equations. **UEO** (Uncertainty Error Overlap[14]): Calculates overlap between the predicted segmentation error and the uncertainty map. Overlap is computed using Dice. We report the maximum attainable UEO score as the uncertainty threshold τ varies. $\hat{u}$ = uncertainty map, e = segmentation error. **UUS** (Modified Unified Uncertainty Score[20]): Computes the filtered Dice and filtered true/false positives/negatives as τ increases. Voxels with uncertainty $> \tau$ are removed from the computation. UUS computes the average of 5 AUC curves: filtered Dice, filtered False Positives (FFP) / Negatives (FFN) and 1 - filtered True Positives (FTP) / Negatives (FTN)). Here FTP measures the ratio of filtered (remaining true positives (TP)) to the total true positives, i.e $1 - TP_\tau/TP$. Filtered Dice (FDice) calculates the Dice metric only on filtered voxels. **ECE** (Expected Calibration Error[8]): approximates the difference between model confidence and accuracy. Predictions are placed into equal width bins B, based on their confidence. ECE compares the weighted average of the difference between the mean confidence $\text{conf}(B_w)$ and mean proportion of WMH in the ground truth $\text{acc}(B_w)$, for the voxels in each bin.

max UEO(↑)	**UUS**(↑)	**ECE**(↓)
$\max_{\tau\in[0,1]} \text{Dice}(\hat{u} > \tau, e)$	$\frac{1}{5}[\text{AUC}_{\text{FDice}} + (2 - \text{AUC}_{FTP} - \text{AUC}_{FTN}) + \text{AUC}_{FFP} + \text{AUC}_{FFN}]$	$\sum_{w=1}^{W} \frac{\lvert B_w \rvert}{n} \cdot \lvert \text{acc}(B_w) - \text{conf}(B_w) \rvert$

[19]. Consequently, boundaries between normal appearing white matter and hyperintense tissue are difficult to delineate and annotators may disagree about the presence of small WMH. Hence, we would like to minimize silent failure, where uncertainty maps fail to highlight WMH in the annotation that is missing from the predicted segmentation. To assess the impact of loss function choice on detection of unsegmented WMH, we calculate the size and proportion of instances missed (no voxels in the instance have uncertainty $> \tau$) in the uncertainty map as we vary τ. We define instances as 2D connected components in axial slices.

3.3 Implementation Test Bench

We utilize a simple benchmarking system for each method, training using a single U-Net architecture for each loss. We use a resnet18 [9] backbone for our U-Net encoder, provided in Segmentation-Models Pytorch[6]. We train three models per loss function, using a different random seed for initializing each model. We report the mean results across the test set for all losses, averaged over each seed. We further evaluate a TopK variant per scoring function. WMH voxels occupy less than 1% of our 3D training images, hence we assess both the performance of k=10 recommended in prior work [17] and a more restricted $k = 1$, reporting the k that achieves the highest lesion F1 score. For Dice++ we first tune the γ parameter on the validation set, choosing the gamma from $\{2, 3, 4\}$ that yields

[6] `https://github.com/qubvel/segmentation_models.pytorch`

the highest lesion F1 score. For post-hoc temperature scaling, we take the models trained on Dice loss and tune κ to maximise the logarithmic score on the validation dataset. Models are optimized using Adam [15], with a learning rate of $3e-4$, multiplying by a factor of 0.1 if validation loss does not improve within 25 epochs, until $\text{lr} = 1e-6$. Early stopping on the validation set, with a patience of 50 epochs, is used to avoid over-fitting. Each epoch consists of 125 batches. To encourage robust, generalizable results with limited training data, we apply numerous augmentations during training, following the choices in nnU-Net [12]. We make our code publicly available. [7] Due to the high proportion of background voxels, where models are typically confident and accurate, $\mathcal{L}_S$ is pulled close to zero. To counter this, we re-weight $\mathcal{L}_S$ and $\mathcal{L}_{\text{TopK}}$ by 1 over the proportion of WMH voxels in the training data (which account for 0.28% of training voxels). This re-weighting is important for yielding strong performance for the logarithmic score. We refer to the logarithmic score as CE in the results. Uncertainty maps are normalized to [0, 1] by dividing by $-\ln(\frac{1}{3})$.

4 Results

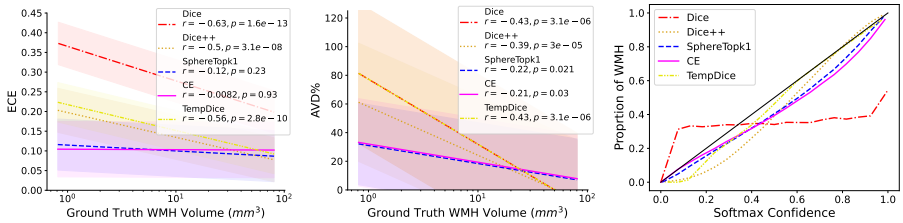
Table 3: Performance Metrics for each loss function on the test set. Best performer per metric in bold. We denote significant improvements (two-tailed t-test comparing the three runs per method, $p < 0.01$) over: both Dice++ and Dice: †; over Dice only: ∗; over Dice++ only: ‡. CE: logarithmic score. TempDice: Model trained with Dice loss, with temperature scaling applied on the validation set.

Loss	UUS	ECE	Dice	F1	AVD%	HD95 mm	Recall	max UEO	prop. miss.
Brier	0.689*	0.076*	0.757‡	0.683	20.7‡	6.97*	0.637	0.427*	0.140*
Spherical	0.688*	0.076*	0.760†	0.700*	20.3‡	6.53*	0.673	0.433*	0.124*
CE	0.682*	0.074*	0.760†	**0.727**†	20.6‡	6.19*	**0.710**	**0.437**†	0.110†
Brier-TopK10	**0.695***	0.083*	0.760†	0.713†	20.9‡	6.43*	0.707	0.430*	0.108†
Sphere-TopK1	0.682*	**0.065***	**0.763**†	0.710†	**19.6**‡	6.37*	0.700	0.433*	0.114†
CE-TopK10	0.685*	0.079*	0.760†	0.720†	19.9‡	**6.06**†	**0.710**	0.435*	**0.106**†
Dice	0.560	0.236	0.717	0.627	35.9	8.61	0.597	0.150	0.275
Dice++	0.684*	0.085*	0.740	0.660	28.0	6.60*	0.695	0.426*	0.144*
TempDice	0.527	0.092*	0.717	0.627	35.9	8.61	0.597	0.233	0.241

The mean performance metrics per loss are shown in Table 3. Surprisingly, all scoring function variants yield significant improvements in Dice score over Dice/Dice++ loss, with the combination of Spherical and TopK yielding the highest Dice and best calibration. CE, or TopK variants of Spherical and Brier score, also show significant improvements over Dice/Dice++ loss for F1 score and in the proportion of instances missed. Furthermore, all scoring functions yield ECE scores less than that of Dice (0.236), or Dice++ (0.085) vs CE (0.068) or Spherical TopK (0.065), while temperature scaling makes only modest improvements (0.092). Figure 1c shows calibration curves for various loss functions. CE,

[7] Code repository: `https://github.com/BenjaminPhi5/Scoring_Functions_WMH`

SphericalTopk1, Dice++ and TempDice yield slightly overconfident distributions, with curves close to the optimum calibration. Dice loss yields arbitrarily accurate predictions regardless of confidences above zero, making differentiating model accuracy based on confidence difficult. Furthermore, examination of the ECE score when compared to ground truth WMH volume reveals Dice loss yields poor calibration at low volumes. Figure 1a shows correlation between individual ECE score and log WMH burden. Dice++ improves calibration over all volumes, but still degrades with log volume ($r = -0.5, p = 3.1e-08$), while both SphericalTopk1 ($r = -0.12, p = 0.23$) and CE ($r = -0.0082, p = 0.93$) retain low ECE at low volumes. Similarly for AVD, while Dice++ improves performance over the Dice model, each scoring function substantially reduces the AVD score. Crucially, scoring functions yield less pronounced degradation in AVD performance for low burden patients (Figure 1a); for CE mean ECE and AVD is almost half that of Dice or Dice++ in images with the lowest WMH burden. While high sample standard deviation (10.4) in average Dice loss AVD scores yields non-significant p-values when compared to scoring functions (Table 3), all scoring functions yield total separation in average AVD results compared to Dice loss.

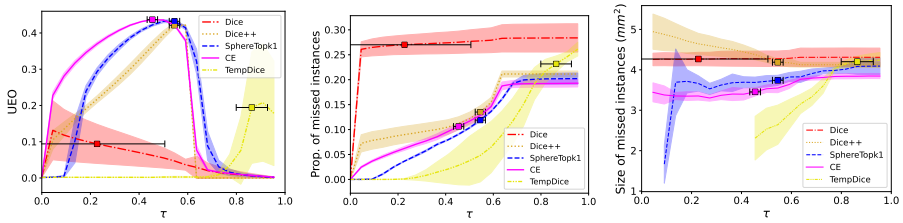


(a) ECE vs WMH Vol. (b) AVD vs WMH Vol. (c) Calibration curves

Fig. 1: (a), (b): Pearson correlation between (a) ECE or (b) AVD with log WMH volume for Dice, Dice++, SphericalTopk1, CE and TempDice. ECE and AVD scores are averaged for each individual over three runs. Shaded area: the standard deviation of the residuals, (c) Average calibration curves across three runs.

Figure 2 shows the uncertainty map performance as τ varies for different loss functions. Fig. 2b shows scoring functions fail to detect lower proportions of WMH instances than Dice or Dice++ at any setting of τ. Noteably, while choosing a τ below the max UEO point ($\hat{\tau}$) can improve the detection of missing instances, both SphericalTopK and CE miss less instances while retaining high UEO scores (Fig. 2a). For Dice++, UEO decreases sharply compared to CE as τ goes to zero, hence choosing a lower τ will yield uncertainty that is poorly localized to the segmentation error, increasing the true positives and negatives incorrectly identified as uncertain. Temperature scaling improves the maximum UEO Dice, but at $\hat{\tau}$ max UEO is less than half that of any other method except Dice, while failing to detect double the instances of CE. Nonetheless, temperature scaling improves the useability of the uncertainty map, permitting meaningful

tuning of τ to set the silent failure rate. Fig 2c shows the mean size of undetected lesions. Regardless of τ, Dice and Dice++ fail to capture some of the smallest instances, while SphericalTopK and CE permit trading UEO score for detection of very small instances. Supplementary B provides uncertainty map examples.



(a) (b) (c)

Fig. 2: Uncertainty map metrics, as uncertainty threshold τ increases. (a) UEO score, (b) Proportion of undetected (intersection over union = 0) WMH instances in segmentation and uncertainty map, (c) Size of undetected WMH instances in segmentation and uncertainty map. Bold line: mean, shaded area: standard deviation, square: mean τ that maximises UEO score, error bar: standard deviation in location of max τ.

5 Discussion

We have shown that proper scoring functions can be effective and well calibrated loss functions for WMH segmentation. Spherical and Brier scores are understudied as segmentation losses, and when combined with TopK to counter class imbalance they can yield competitive performance. The TopK variants yield the same or improved results for Dice, HD95, Recall and the proportion of missed lesions at $\hat{\tau}$ across all scoring function losses. These variants could be fitted into existing compound losses that generalize well across datasets [29], and tuning of cutoff k could further enable generalization to other tasks with high class imbalance. Furthermore, alternative scoring rules such as generalized spherical or winklers score offer parameterized scoring functions that may be adjusted to the task at hand [7]. Greater attention should be paid to AVD divergence and poor calibration in low WMH burden patients, such individuals are poorly represented in average lesion-wise scores (e.g lesion F1) or ECE scores due to the small lesion number and volume that they contribute overall. We find scoring functions yield models that are less prone to this degradation than Dice based methods, with no appreciable degradation in ECE at low volumes. This is especially important for downstream tasks such as identifying patients at risk of developing further manifestations of cerebral small vessel disease [24]. While Dice++ consistently outperforms Dice loss, performance still degrades at low WMH burdens, albeit

to a lesser extent. Small WMH instances can often be missed by a segmentation model; however they often indicate areas where tissue microstructure is damaged, precedent of wider damage [6] and hence are important to detect - regardless of their size [18],[26]. ECE score is not indicative of the most useful uncertainty maps, with BrierTopk and CETopK yielding the highest UUS and UEO respectively, despite higher ECE than other scoring function variants. Finally, while temperature scaling is effective and robust for improving the calibration of a given model [25], it alone is insufficient to overcome the drawbacks of Dice loss, yielding no significant improvement in any other metric. UEO and silent failure rates see only modest improvements, with ECE still degrading for low WMH burden. Global temperature scaling is not suitable in segmentation tasks (which commonly display spatial and heteroscedastic aleatoric uncertainty [21]) such as WMH segmentation. Dice loss encourages high confidence in both WMH instance centres and at ambiguous areas (edges and small WMH); hence the temperature parameter must vary spatially and per image. A separate model for predicting the temperature parameter [23] can achieve this, however this increases complexity and amount of validation data required [4] while still unable to improve over the segmentation performance of Dice loss.

References

1. Abdar, M., Pourpanah, F., Hussain, S., et.al: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (Dec 2021). https://doi.org/10.1016/j.inffus.2021.05.008
2. Ashburner, J., Friston, K.J.: Voxel-Based Morphometry—The Methods. NeuroImage **11**(6), 805–821 (Jun 2000). https://doi.org/10.1006/nimg.2000.0582
3. Balakrishnan, R., Hernández, M.d.C.V., Farrall, A.J.: Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data–a systematic review. Computerized Medical Imaging and Graphics **88**, 101867 (2021)
4. Balanya, S.A., Maroñas, J., Ramos, D.: Adaptive temperature scaling for robust calibration of deep neural networks. arXiv preprint arXiv:2208.00461 (2022)
5. Gaubert, M., et.al: Performance evaluation of automated white matter hyperintensity segmentation algorithms in a multicenter cohort on cognitive impairment and dementia. Frontiers in psychiatry **13**, 2928 (2023), publisher: Frontiers
6. Ge, Y., Grossman, R.I., Babb, J.S., et.: Dirty-appearing white matter in multiple sclerosis: volumetric MR imaging and magnetization transfer ratio histogram analysis. American Journal of Neuroradiology **24**(10), 1935–1940 (2003), publisher: Am Soc Neuroradiology
7. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association **102**(477), 359–378 (2007), publisher: Taylor & Francis
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1321–1330. PMLR (Jul 2017), iSSN: 2640-3498
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

10. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 137–145. Springer (2019)
11. Iglesias, J.E., Cheng-Yi Liu, Thompson, P.M., Zhuowen Tu: Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods. IEEE Transactions on Medical Imaging **30**(9), 1617–1634 (Sep 2011). https://doi.org/10.1109/TMI.2011.2138152
12. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, number: 2 Publisher: Nature Publishing Group
13. Jose, V.R.: A characterization for the spherical scoring rule. Theory and Decision **66**, 263–281 (2009), publisher: Springer
14. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. Frontiers in Neuroscience **14** (2020)
15. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (Jan 2017). https://doi.org/10.48550/arXiv.1412.6980, arXiv:1412.6980 [cs]
16. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Transactions on Medical Imaging **29**(1), 196–205 (Jan 2010). https://doi.org/10.1109/TMI.2009.2035616, conference Name: IEEE Transactions on Medical Imaging
17. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (Jul 2021). https://doi.org/10.1016/j.media.2021.102035
18. MacLullich, A.M., Ferguson, K.J., Reid, L.M., et.al: Higher systolic blood pressure is associated with increased water diffusivity in normal-appearing white matter. Stroke **40**(12), 3869–3871 (2009), publisher: Am Heart Assoc
19. Maillard, P., Fletcher, E., Harvey, D., Carmichael, O., Reed, B., Mungas, D., DeCarli, C.: White matter hyperintensity penumbra. Stroke **42**(7), 1917–1922 (2011)
20. Mehta, R., Filos, A., Baid, U., et.al: QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation–Analysis of Ranking Metrics and Benchmarking Results. arXiv preprint arXiv:2112.10074 (2021)
21. Monteiro, M., Le Folgoc, L., et.al: Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty. In: Advances in Neural Information Processing Systems. vol. 33, pp. 12756–12767. Curran Associates, Inc. (2020)
22. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P.H., Gal, Y.: Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. arXiv preprint arXiv:2102.11582 (2021)
23. Ouyang, C., Wang, S., Chen, C., Li, Z., Bai, W., Kainz, B., Rueckert, D.: Improved post-hoc probability calibration for out-of-domain mri segmentation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. pp. 59–69. Springer (2022)
24. Prins, N.D., Scheltens, P.: White matter hyperintensities, cognitive impairment and dementia: an update. Nature Reviews Neurology **11**(3), 157–165 (2015), publisher: Nature Publishing Group UK London
25. Rahaman, R., et al.: Uncertainty quantification and deep ensembles. Advances in Neural Information Processing Systems **34**, 20063–20075 (2021)

26. Wardlaw, J.M., Valdés Hernández, M.C., Muñoz-Maniega, S.: What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. Journal of the American Heart Association **4**(6), e001140 (2015), publisher: Am Heart Assoc
27. Wu, Z., Shen, C., Hengel, A.v.d.: Bridging category-level and instance-level semantic image segmentation. arXiv preprint arXiv:1605.06885 (2016)
28. Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C.B., Yang, G.: Calibrating the Dice Loss to Handle Neural Network Overconfidence for Biomedical Image Segmentation. Journal of Digital Imaging **36**(2), 739–752 (Apr 2023). https://doi.org/10.1007/s10278-022-00735-3
29. Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L.: Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized Medical Imaging and Graphics **95**, 102026 (Jan 2022). https://doi.org/10.1016/j.compmedimag.2021.102026
30. Zou, K., Chen, Z., Yuan, X., et.al: A review of uncertainty estimation and its application in medical imaging. arXiv preprint arXiv:2302.08119 (2023)