Evaluation of Uncertainty-Aware Multi-Software Ensembles for Hippocampal Segmentation

Gabriel Oliveira-Stahl[1,2 [0009−0008−1489−6406]], Anna Schroder[1,2 [0009−0001−0380−0674]], James Moggridge[2,3 [0009−0002−0600−5315]], Hamza A. Salhab[2 [0000−0003−3575−7938]], Caroline Micallef[2], Josephine Barnes[6 [0000−0003−3178−025X]], M. Jorge Cardoso[5 [0000−0003−1284−2558]], Carole H. Sudre*[1,4,5 [0000−0001−5753−428X]], and Matthew Grech-Sollars*[1,2 [0000−0003−3881−4870]]

[1] Hawkes Institute, Department of Computer Science, University College London, UK
[2] Lysholm Department of Neuroradiology, National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK
[3] Department of Brain Repair and Rehabilitation, UCL Institute of Neurology, University College London, UK
[4] Unit for Lifelong Health and Ageing, Department of Population Science and Experimental Medicine, University College London, UK
[5] School of Biomedical Engineering & Imaging Sciences, King's College London, UK
[6] Dementia Research Centre, University College London, UK

* Joint senior authors

Abstract. Accurate hippocampal segmentation can be a useful tool for diagnosing and monitoring neurological conditions such as Alzheimer's disease and epilepsy. While numerous automated segmentation methods exist, their clinical adoption remains limited. Reliable uncertainty assessment can enhance trust and facilitate clinical translation. This study evaluates five heterogeneous hippocampal segmentation methods — InnerEye, ASHS, FastSurfer, HippoSeg, and FreeSurfer — across two dementia datasets and one epilepsy dataset. The sub-ensemble containing InnerEye, FastSurfer, and HippoSeg emerged as both accurate and efficient, highlighting the feasibility of balancing computational cost and performance. Additionally, ensemble-derived uncertainty quantification with sample variance, mutual information, and predictive entropy is shown to reduce inaccurate segmentations by flagging low-confidence cases, potentially providing a mechanism for automatically escalating ambiguous cases for expert assessment.

Keywords: Uncertainty estimation · Ensemble · Hippocampal segmentation · Carbon footprint.

1 Introduction

Hippocampal volume changes are associated with several neurological diseases such as Alzheimer's and epilepsy [10,5]. Accurate volumetric measurement can

serve as a biomarker for diagnosis and disease monitoring [22,3]. While manual segmentation by an expert is still considered the gold standard, its time-consuming nature and the well-documented intra- and inter-reader variability remain significant limitations [2,8]. Multiple different protocols for hippocampal segmentation exist, complicating the notion of a ground truth segmentation [8].

Numerous methods for automatic hippocampus segmentations have been proposed, ranging from atlas-based approaches [4], to deep learning methods [20]. Despite the range of available methods, clinical adoption remains limited. Key concerns hindering the clinical translation of these tools include the risk of silent failure, and a lack of robustness to unexpected distribution shifts. Reliable uncertainty estimation can play a central role in strengthening the trustworthiness of these algorithms, e.g. by escalating uncertain cases for human review, and various approaches for assessing predictive uncertainty have been explored in the literature [1]. While Bayesian neural networks offer the most principled framework for posterior approximation, they are often computationally prohibitive. Monte-Carlo Dropout and ensembles have emerged as practical methods for Bayesian approximation [14,7]. Beyond providing uncertainty estimation, ensemble methods consistently achieve superior performance over individual models, as repeatedly demonstrated by results from the Brain Tumor Image Segmentation Benchmark (BRATS) [15] and the White Matter Hyperintensity Segmentation Challenge [13]. Ensemble predictions can be aggregated through multiple strategies, with stacking of heterogeneous models proving highly effective and simple averaging providing a stable method for integrating their predictions [17]. Although deep ensembles often outperform single models, their combined outputs are not automatically well-calibrated, even when each constituent network is well-calibrated on its own [27,19]. Calibration deteriorates further under distribution shift [18], a problem that has been traced in part to insufficient diversity among ensemble members [11]. To counter this, the present study assesses the disagreement of deliberately heterogeneous segmentation methods and treats that disagreement itself as an uncertainty estimate, a strategy previously explored by Kofler et al. [12].

Another growing concern is the computational cost and energy consumption of the segmentation methods. Recent work has highlighted the potential environmental impact of large-scale deep learning models [23,26], emphasizing the need for sustainable solutions. High computational demands not only contribute to a significant carbon footprint, but also limit accessibility for hospitals and research institutions with limited resources. Addressing both the need for reliable uncertainty estimation and computational efficiency will be conducive for enabling broader clinical adoption.

Our contributions are three-fold. We first provide a comprehensive benchmark of five hippocampal segmentation methods, evaluating their accuracy, calibration, and carbon footprint across two dementia cohorts and one epilepsy cohort, where we identify a resource-efficient sub-ensemble: a triad of methods that matches the full ensemble's Dice performance while reducing inference-time CO_2 emissions by approximately 70%. Second, we test different methods to inte-

grate the triad's predictions, finding simple averaging to perform best. Third, we demonstrate that uncertainty estimates derived from the triad's predictive distributions can identify low-confidence cases; flagging the most uncertain segmentations successfully removes many gross segmentation failures across datasets.

2 Methods

2.1 Segmentation tools

The five chosen segmentation tools are:

- ASHS (AS) - an Atlas-based method specifically developed for segmenting substructures in the medial temporal lobe using the Penn Memory Center 3T ASHS Atlas for T1-weighted MRI [28].
- HippoSeg (HI) - an atlas-based hippocampal segmentation tool optimized for epilepsy patients [24][25].
- Freesurfer (FR) - a neuroimaging toolkit with atlas-based whole brain parcellation as a subcomponent [6].
- FastSurfer (FA) - a faster deep-learning based alternative to FreeSurfer [9].
- InnerEye (IN) - a deep-learning based hippocampal segmentation tool [21,20] trained using the InnerEye toolbox, an open-source deep-learning toolbox from Microsoft supporting medical image segmentation [16].

2.2 Data

Hippocampal segmentations were performed on three independent datasets, for which gold standard manual segmentations were defined by a clinician with neuroradiology experience, and confirmed by a consultant neuroradiologist. The definition followed an amended version of the harmonised protocol definition [2] (inclusion of the fimbria).

ADNI dataset : 30 participants of the Alzheimer's Disease Neuroimaging Initiative (ADNI) (mean age 76.8 ± 7.3 years; 70% male). One third were diagnosed with Alzheimer's disease, one third with late mild cognitive impairment, and one third were cognitively normal.

Local Epilepsy: Eight patients with epilepsy (mean age 46.8±24.9 years; 62.5% female). MRI scans were acquired using a Siemens Prisma 3T scanner with an MPRAGE sequence and a voxel size of 1.1mm×1.1mm×1.2mm. All participants provided written informed consent. The study was approved by the local ethics committee.

Local Dementia: 21 patients referred to the dementia clinic (age data unavailable; 71% male). MRI scans were acquired on a Siemens Prisma 3T scanner using an MPRAGE sequence with a voxel size of 1.1mm×1.1mm×1.2mm. The study was approved by the local ethics committee. Informed consent was not required as this was a retrospective study (Anon ethics).

2.3 Processing

Inference for the ADNI dataset was performed on a High-Performance Computing cluster. For carbon footprint benchmarking, the data was processed on a dedicated node equipped with 4 NVIDIA GTX 1080 Ti GPUs, an Intel Xeon E5-2640 v4 CPU (2.40GHz), and 64GB of RAM, with no concurrent jobs from other users. The local hospital data was processed entirely on a local MacBook Pro (Intel Core i5-5257U CPU @ 2.70 GHz, with ∼8GB of memory). For fair comparison, same software versions were installed on the cluster and on the local machine. An older version of FreeSurfer (v6.0.0) was selected to accommodate the memory constraints of the local machine. FastSurfer (v2.3.3) was run from a docker image on the local machine, and via singularity (v3.8.5-2.el7) on the reserved cluster node. ASHS (v2.0.0, July 2018) was run natively on the respective host system. In a first experiment, performance of all individual methods and possible sub-ensembles was compared, using the binary segmentation outputs of the five segmentation methods. The expected calibration error (ECE) of each sub-ensemble was approximated by interpreting the number of votes from the $T = 3$ ensemble members as discrete confidence levels. Specifically, if n models predicted the class *hippocampus* for a voxel v, the ensemble confidence was defined as $p_v = \frac{n}{T}$. For the winning triad (InnerEye + Fastsurfer + HippoSeg) minor changes to their code were introduced to enable saving of the probabilistic outputs. The relevant amended scripts are publicly available on GitHub: https://github.com/zfdiwkm/multi-software-uncertainty. Because epilepsy can cause unilateral hippocampal atrophy the left and right hippocampi were processed separately to avoid averaging out potential differences in segmentation performance. This separation allows for evaluating whether some methods underperform specifically on the atrophied side. Energy consumption and time to completion was tracked with the python package codecarbon (v2.8.3). No additional pre-processing of the scans was performed prior to segmentation with the respective tool.

2.4 Uncertainty

For each voxel v, uncertainty was quantified as sample variance, mutual information (MI), and predictive entropy (PE) across the winning triad's probabilistic outputs. Sample variance was calculated as $\mathrm{Var}_v = \frac{1}{T-1}\sum_{t=1}^{T}(y_{v,t} - \bar{y}_v)^2$, with the mean prediction $\bar{y}_v = \frac{1}{T}\sum_{t=1}^{T} y_{v,t}$. Mutual information was defined as $\mathrm{MI}_v = H(\bar{p}_v) - \frac{1}{T}\sum_{t=1}^{T} H(p_{v,t})$, and predictive entropy as $\mathrm{PE}_v = H(\bar{p}_v) = -\bar{p}_v \log_2 \bar{p}_v - (1-\bar{p}_v)\log_2(1-\bar{p}_v)$, where $\bar{p}_v = \frac{1}{T}\sum_{t=1}^{T} p_{v,t}$, $p_{v,t}$ is the predicted probability for voxel v from tool t, and T is the number of segmentation tools. Figure 1 presents confidence maps for all three UQ metrics for a high-dice example (a) and a low-dice example (b). Uncertainty calibration was evaluated by the ECE, with 20 bins of equal bin-width. To not have larger hippocampi have a disproportionate influence, ECE was calculated as the average of the per-hippocampus ECE values. For each hippocampus we computed

$ECE = \frac{B_m}{n} |acc(B_m) - conf(B_m)|$, where B_m represents the set of indices of voxels with confidence values falling into bin m.

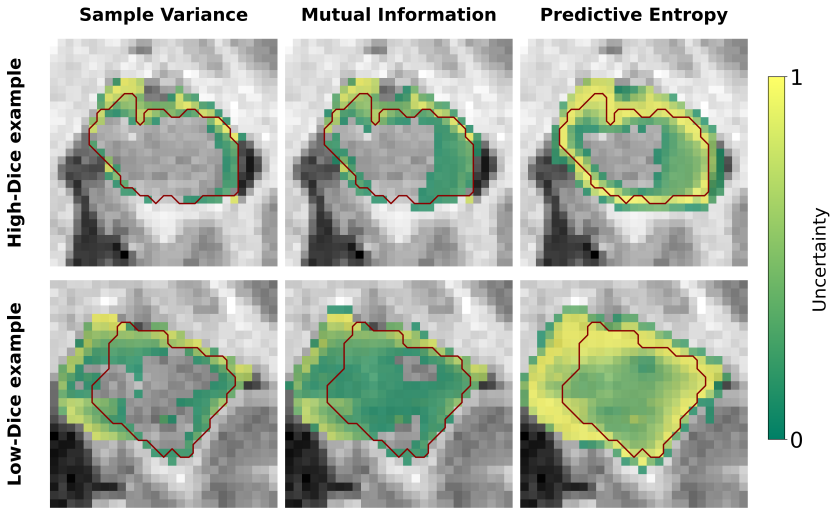

Fig. 1. Visualization of uncertainty maps. Uncertainty quantified as sample variance, mutual information, and predictive entropy, all min-max normalized. Dark red line represents the gold standard boundary. Top row: High-dice example. Bottom row: Low-dice example.

2.5 Evaluation metrics & experiments

Segmentation performance was evaluated using the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff distance (HD), predictive uncertainty with sample variance, mutual information, and entropy, while uncertainty calibration was evaluated in terms of expected calibration error. As initial benchmarking (Experiment 1), the five individual segmentation methods, all possible triads formed from them, and the full ensemble of five methods were compared on the ADNI dataset. For each triad and for the full ensemble, the final prediction mask was obtained via majority voting. Performance was assessed in terms of DSC, ECE, and estimated CO_2 emissions. In a second experiment, we compared different strategies for integrating the predictions of the best-performing triad to assess their impact on segmentation accuracy and uncertainty calibration. These ensemble strategies were: (1) simple averaging of the output probabilities from each ensemble member (denoted as 'Ensemble_Averaged' in the results), (2) weighting each prediction by the model's confidence ('Ensemble Weighted'), and (3) assigning each voxel to the prediction of the most confident model in a winner-takes-all approach ('Ensemble Triage'). In a third experiment then, we

evaluated whether hippocampus-wide aggregated uncertainty could serve as a predictor of low-quality segmentations. To this end, uncertainty values were aggregated across the region of interest (ROI), defined as the union of predicted segmentations with a one-voxel margin. Morphological dilation was performed once, with `scipy.ndimage.binary_dilation` with a square connectivity = 1 (6-connected kernel). The hippocampi with the highest uncertainty were excluded, and the resulting impact on segmentation performance was assessed to determine the method's suitability for preventing silent failures. The threshold for exclusion was determined through the following steps: (A) classifying all predictions into 'correct' or 'failed' segmentations, with failed segmentations defined here as DSC < 0.82 or HD95 > 2.5 mm; (B) performing a stratified split of predictions into a validation set (30%) and a test set (70%); (C) conducting a grid search to identify the optimal threshold per dataset and metric by maximizing the F1-score; and (D) applying the identified optimal threshold to the held-out test data. Due to non-normality of studied distributions, Friedman tests were performed to assess significant difference in performance. To account for multiple comparisons, p-values were adjusted with the Holm-Bonferroni method. Resource consumption was estimated in terms of duration and CO_2 expenditure for each method and each subject.

3 Results

3.1 Benchmarking cost and performance

Experiment 1 - benchmarking with binary segmentations: The ensemble consisting of InnerEye, FastSurfer, and HippoSeg produced segmentations at the same level of accuracy as an ensemble consisting of all five softwares, while substantially reducing carbon emissions (from 5.5 kg to 1.7 kg CO_2 for processing of 30 ADNI patients) and retaining a comparably well-calibrated uncertainty measure (Fig. 2). For the remainder of the paper, any mention of 'Triad' will therefore refer to this specific ensemble.

3.2 Segmentation performance

Experiment 2 - ensemble integration strategy: Simple mean averaging of the ensemble members' probabilistic outputs yielded the highest Dice performance across all datasets (Fig. 3, $p < 0.05$ after Holm-Bonferroni correction). For the HD, no significant difference was observed between the mean ensemble and other ensemble integration strategies on the epilepsy dataset. On the Local Dementia dataset, the mean ensemble did not significantly differ from InnerEye, while it significantly outperformed all other methods (Fig. 3, $p < 0.05$, Holm-Bonferroni corrected). Experiment 3 - Silent failure detection: Applying dataset-specific uncertainty thresholds demonstrated varying effectiveness across UQ metrics and datasets. For Dice-based outlier detection (Dice < 0.82), entropy achieved the highest performance with 91.7% recall and 44.2% precision, while variance and

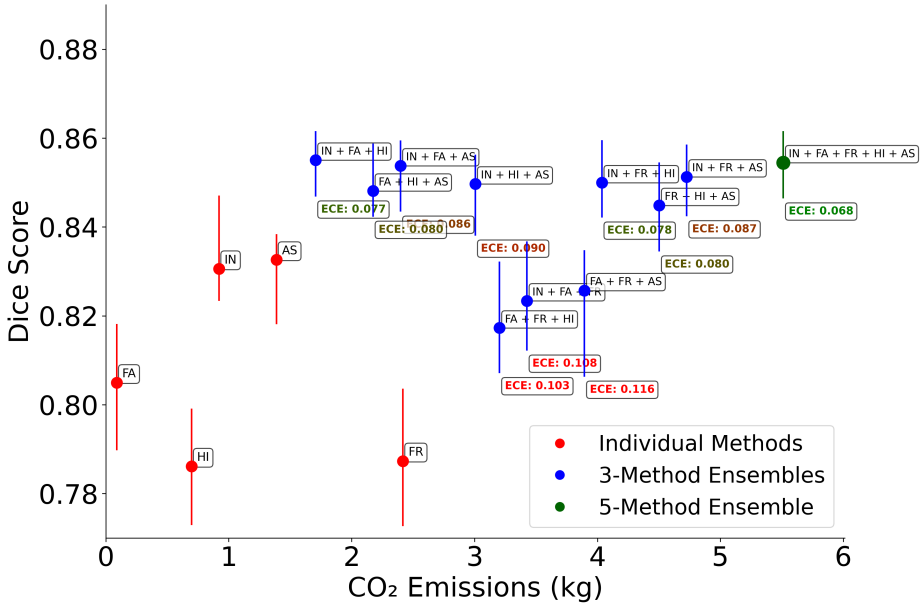

Fig. 2. Balance between accuracy, CO_2 emission, and uncertainty calibration. Benchmarking on ADNI dataset, processed on dedicated node on HPC cluster. Error bars indicate bootstrapped confidence intervals. Winning *Triad*: InnerEye + FastSurfer + HippoSeg. Abbreviations: HI = HippoSeg, FA = FastSurfer, FR = FreeSurfer, IN = InnerEye, AS = ASHS. ECE values color coded with higher values in brighter red.

mutual information both achieved 83% recall with ∼35% precision. For HD95-based outlier detection (HD95 $>$ 2.5mm), entropy again performed best with 88.9% recall and 40.6% precision, compared to variance and mutual information achieving 55.6% recall with 28.9% and 30.6% precision, respectively.

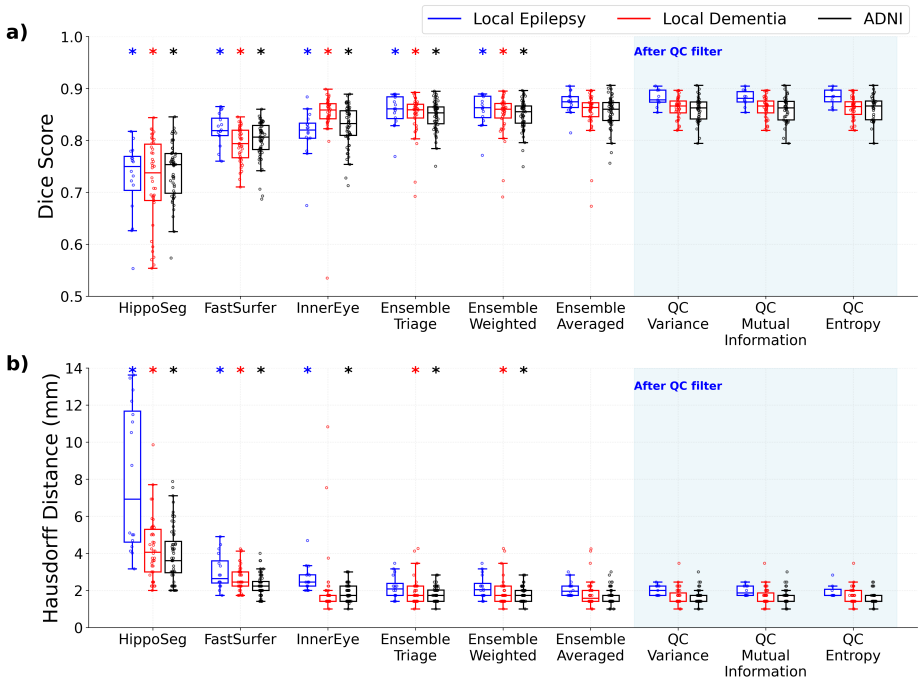

Fig. 3. Segmentation performance comparison. Ensemble = InnerEye + FastSurfer + HippoSeg. **a)** Dice score comparison. **b)** Hausdorff Distance comparison. For all quality control groups (QC) the aggregated uncertainty from the Ensemble Averaged was used to discard the most uncertain hippocampi, resulting in 35% excluded predictions when using entropy, and ∼26% exclusions with variance and mutual information (averaged across datasets). Asterisk * indicates significant difference to Ensemble Averaged; p-values in Suppl. Table 1.

3.3 Uncertainty calibration

The averaged probabilistic output by the *Triad* members produced the best calibrated confidences, with an ECE of 0.05 across all three datasets. The second most calibrated segmentation method was HippoSeg, with an ECE slightly higher than 0.05 averaged across all datasets. InnerEye (ECE $\approx$ 0.08) and FastSurfer in particular (ECE $\approx$ 0.16) were less well-calibrated.

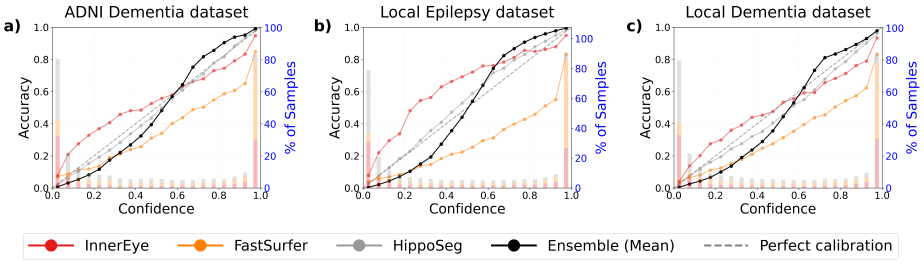

Fig. 4. Reliability diagram. **a)** ADNI dataset. **b)** Local Epilepsy dataset. **c)** Local Dementia dataset. **Black axes:** calibration curve. **Blue axis:** distribution of confidence scores. Stacked bars show the percentage of voxels in each confidence bin for the respective segmentation software. Analysis includes union of predicted segmentations with a one-voxel margin.

4 Conclusion & future work

The current study comprehensively evaluated ensembles of heterogeneous hippocampal segmentation methods, and proposes the triad of InnerEye, FastSurfer, and HippoSeg as the optimal combination for achieving high segmentation performance while minimizing resource cost. This ensemble significantly outperformed all individual methods, while also producing well-calibrated uncertainty estimates across three independent datasets. We demonstrated that aggregating ensemble-derived uncertainty values can be useful for catching low-quality predictions, highlighting the potential for reliable uncertainty estimation to reduce the risk for silent failures in a clinical setting. Future work will focus on assessing the robustness of uncertainty-based failure detection under domain shift and in lower-quality clinical scans.

Disclosure of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdar, M., et al.: A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges (2020). https://doi.org/10.48550/ARXIV.2011.06225, https://arxiv.org/abs/2011.06225, publisher: arXiv Version Number: 4
2. Boccardi, M., et al.: Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimer's & Dementia **11**(2), 126–138 (Feb 2015). https://doi.org/10.1016/j.jalz.2014.02.009
3. Briellmann, R.S., Berkovic, S.F., Syngeniotis, A., King, M.A., Jackson, G.D.: Seizure-associated hippocampal volume loss: A longitudinal magnetic resonance study of temporal lobe epilepsy. Annals of Neurology **51**(5), 641–644 (May 2002). https://doi.org/10.1002/ana.10171
4. Dill, V., Franco, A.R., Pinho, M.S.: Automated Methods for Hippocampus Segmentation: the Evolution and a Review of the State of the Art. Neuroinformatics **13**(2), 133–150 (Apr 2015). https://doi.org/10.1007/s12021-014-9243-4
5. Duan, Y., Lin, Y., Rosen, D., Du, J., He, L., Wang, Y.: Identifying Morphological Patterns of Hippocampal Atrophy in Patients With Mesial Temporal Lobe Epilepsy and Alzheimer Disease. Frontiers in Neurology **11**, 21 (Jan 2020). https://doi.org/10.3389/fneur.2020.00021
6. Fischl, B.: FreeSurfer. NeuroImage **62**(2), 774–781 (Aug 2012). https://doi.org/10.1016/j.neuroimage.2012.01.021
7. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (2015). https://doi.org/10.48550/ARXIV.1506.02142, https://arxiv.org/abs/1506.02142, version Number: 6
8. Geuze, E., Vermetten, E., Bremner, J.D.: MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed. Molecular Psychiatry **10**(2), 147–159 (Feb 2005). https://doi.org/10.1038/sj.mp.4001580
9. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M.: FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. NeuroImage **219**, 117012 (Oct 2020). https://doi.org/10.1016/j.neuroimage.2020.117012
10. Jack, C.R.: MRI-Based Hippocampal Volume Measurements in Epilepsy. Epilepsia **35**(s6) (Dec 1994). https://doi.org/10.1111/j.1528-1157.1994.tb05986.x
11. Javanbakhat, M., Hasan, M.T., Lippert, C.: Assessing Uncertainty Estimation Methods for 3D Image Segmentation under Distribution Shifts (2024). https://doi.org/10.48550/ARXIV.2402.06937, https://arxiv.org/abs/2402.06937, version Number: 1
12. Kofler, F., et al.: Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles. Frontiers in Neuroscience **15**, 752780 (Dec 2021). https://doi.org/10.3389/fnins.2021.752780, https://www.frontiersin.org/articles/10.3389/fnins.2021.752780/full
13. 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 (Nov 2019). https://doi.org/10.1109/TMI.2019.2905770, https://ieeexplore.ieee.org/document/8669968/
14. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles (2016). https://doi.org/10.48550/ARXIV.1612.01474, https://arxiv.org/abs/1612.01474, version Number: 3
15. Menze, B.H., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (Oct 2015).

https://doi.org/10.1109/TMI.2014.2377694, http://ieeexplore.ieee.org/document/6975210/
16. Microsoft: InnerEye Deep Learning Toolbox, https://github.com/microsoft/InnerEye-DeepLearning
17. Muller, D., Soto-Rey, I., Kramer, F.: An Analysis on Ensemble Learning Optimized Medical Image Classification With Deep Convolutional Neural Networks. IEEE Access **10**, 66467–66480 (2022). https://doi.org/10.1109/ACCESS.2022.3182399, https://ieeexplore.ieee.org/document/9794729/
18. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift (2019). https://doi.org/10.48550/ARXIV.1906.02530, https://arxiv.org/abs/1906.02530, version Number: 2
19. Rahaman, R., thiery, a.: Uncertainty Quantification and Deep Ensembles. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 20063–20075. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/a70dc40477bc2adceef4d2c90f47eb82-Paper.pdf
20. Schroder, A., Moggridge, J., Wu, J., Salhab, H.A., Vos, S., Bristow, M., Pérez-García, F., Alvarez-Valle, J., Yousry, T.A., Thornton, J.S., others: InnerEye as a Tool for Accurate Hippocampal Segmentation. Proceedings of 2023 ISMRM Annual Meeting & Exhibition (ISMRM). (2023)
21. Schroder, A., Salhab, H.A., Moggridge, J., Micallef, C., Wu, J., Vos, S., Bristow, M., Pérez-García, F., Alvarez-Valle, J., Yousry, T.A., others: Clinical Validation of the InnerEye Hippocampal Segmentation Tool. Proceedings of 2024 ISMRM Annual Meeting & Exhibition (ISMRM). (2024)
22. Schuff, N., Woerner, N., Boreta, L., Kornfield, T.: MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain **132**(4), 1067–1077 (May 2008). https://doi.org/10.1093/brain/awp007, https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awp007
23. Strubell, E., Ganesh, A., McCallum, A.: Energy and Policy Considerations for Modern Deep Learning Research. Proceedings of the AAAI Conference on Artificial Intelligence **34**(09), 13693–13696 (Apr 2020). https://doi.org/10.1609/aaai.v34i09.7123, https://ojs.aaai.org/index.php/AAAI/article/view/7123
24. Vos, S.B., et al.: Hippocampal profiling: Localized magnetic resonance imaging volumetry and T2 relaxometry for hippocampal sclerosis. Epilepsia **61**(2), 297–309 (Feb 2020). https://doi.org/10.1111/epi.16416, https://onlinelibrary.wiley.com/doi/10.1111/epi.16416
25. Winston, G.P., Cardoso, M.J., Williams, E.J., Burdett, J.L., Bartlett, P.A.: Automated hippocampal segmentation in patients with epilepsy: Available free online. Epilepsia **54**(12), 2166–2173 (Dec 2013). https://doi.org/10.1111/epi.12408, https://onlinelibrary.wiley.com/doi/10.1111/epi.12408
26. Wu, C.J., et al.: Sustainable AI: Environmental Implications, Challenges and Opportunities. In: Proceedings of Machine Learning and Systems. vol. 4, pp. 795–813 (2022)
27. Wu, X., Gales, M.: Should Ensemble Members Be Calibrated? (Jan 2021). https://doi.org/10.48550/arXiv.2101.05397, http://arxiv.org/abs/2101.05397, arXiv:2101.05397 [cs]
28. Xie, L., et al.: Accounting for the Confound of Meninges in Segmenting Entorhinal and Perirhinal Cortices in T1-Weighted MRI. In: Ourselin, S. (ed.) MICCAI 2016, vol. 9901, pp. 564–571. Cham (2016)