VesselVerse: A Dataset and Collaborative Framework for Vessel Annotation

Daniele Falcetta[1 [0009−0009−7199−5424]], Vincenzo Marciano[1,2 [0009−0000−4024−3785]], Kaiyuan Yang[3 [0000−0003−0591−2079]], Jon Cleary[4,2 [0000−0001−8836−9934]], Loïc Legris[5,6 [0000−0001−6697−4519]], Massimiliano Domenico Rizzaro[7 [0009−0005−6008−7579]], Ioannis Pitsiorlas[1 [0000−0003−0749−013X]], Hava Chaptoukaev[1 [0009−0000−0859−4059]], Benjamin Lemasson[6 [0000−0003−0446−3531]], Bjoern Menze[3 [0000−0003−4136−5690]], and Maria A. Zuluaga[1,2 [0000−0002−1147−766X]]

[1] EURECOM, Biot, France
[2] School of Biomedical Engineering & Imaging Sciences, King's College London, UK
[3] Quantitative Biomedicine, University of Zurich, Switzerland
[4] Dept. of Radiology, Guy's and St, Thomas' NHS Foundation Trust, London, UK
[5] CHU Grenoble Alpes, Grenoble, France
[6] Univ. Grenoble Alpes, Inserm U1216, Grenoble Institut Neurosciences, France
[7] Neurosurgery Department, University of Milan, Italy
maria.zuluaga@eurecom.fr

Abstract. This paper is not about a novel method. Instead, it introduces `VesselVerse`, a large-scale annotation dataset and collaborative framework for brain vessel annotation. It addresses the critical challenge of data annotation availability in supervised learning segmentation and provides a valuable resource for the community. `VesselVerse` represents the largest public release of brain vessel annotations to date, comprising 950 annotated images from three public datasets across multiple neurovascular imaging modalities. Its design allows for multi-expert annotations per image, accounting for variations across diverse annotation protocols. Furthermore, the framework facilitates the inclusion of new annotations and refinements to existing ones, making the dataset dynamic. To enhance annotation reliability, `VesselVerse` integrates tools for consensus generation and version control mechanisms, enabling the reversion of errors introduced during annotation refinement. We demonstrate `VesselVerse`'s usability by assessing inter-rater agreement among four expert evaluators.

Keywords: Brain vessel annotation · Multi-expert annotation · Collaborative framework.

1 Introduction

Image annotations are essential for developing medical image segmentation algorithms. They are critical in training supervised learning models, which typically

Table 1. Publicly available brain vessel datasets with annotations (#).

Dataset	**Modalities**	#	**Annotation Protocol**
COSTA	TOF-MRA	354	Brain arteries pixel-wise
CAS	TOF-MRA	100	Brain arteries pixel-wise
IXI from [2]	TOF-MRA	45	Brain arteries pixel-wise
SMILE-UHURA	7T TOF-MRA	14	Brain & extraparenchymal vessels pixel-wise
TubeTK	MRA	42	Brain arteries centerlines + radius
TopCoW24	CTA / TOF-MRA	125 /125	Circle of Willis pixe-wise

require a substantial amount of paired image data and segmentation masks. Furthermore, even with the increasing use of semi-supervised, weakly supervised, or unsupervised methods, annotations are still needed for validating and evaluating the techniques that have been developed. Although the medical imaging community has made advancements in acquiring and curating large datasets (e.g., [18]), fully annotated data remain scarce, particularly for complex anatomical structures such as the brain's vasculature.

The scarcity of large, annotated datasets of brain vessels can be attributed to the intricate and complex nature of the 3D cerebral vasculature. Its highly branched structure, multi-scale representation, and often tortuous paths make manual annotation challenging and time-consuming. Accurately delineating individual vessels, especially smaller ones, requires significant expertise and meticulous attention to detail. In addition to the tedious nature of the annotation process, differences in annotation protocols and the subjective interpretation of these protocols further complicate matters. Variations in how vessels are defined, the level of detail required, and the specific tools used can lead to discrepancies in vessel annotation [17], thus introducing inter-rater variability [9]. As a result, comprehensive annotated datasets of the brain vessels are limited, often employ differing annotation protocols (see Table 1) and exhibit significant variations in the resulting annotations (Figure 1).

Crowdsourcing provides a promising solution for annotating large neurovascular imaging datasets while reducing annotator fatigue and errors by sharing the workload among experts. However, recent studies have raised concerns about the annotation quality achieved through crowdsourcing platforms [16]. To fully benefit from crowdsourcing, the complex task of vessel annotation requires quality control mechanisms in place. These should include version control systems for tracking individual contributions, facilitating expert reviews, and consensus mechanisms for harmonizing different annotation styles and resolving discrepancies that arise from subjective interpretations of vessel boundaries.

Although version control systems are widely adopted in software development, their usage and integration into medical imaging annotated datasets remain limited. Open-access datasets propose final segmentation masks without tracking the evolution of annotations, and very few previous efforts offer mechanisms to integrate multiple expert annotations and refinements [20]. Open-source tools like ITK-Snap [24] or 3D Slicer [10] provide advanced visualization and analysis but lack built-in version-controlled annotation workflows.

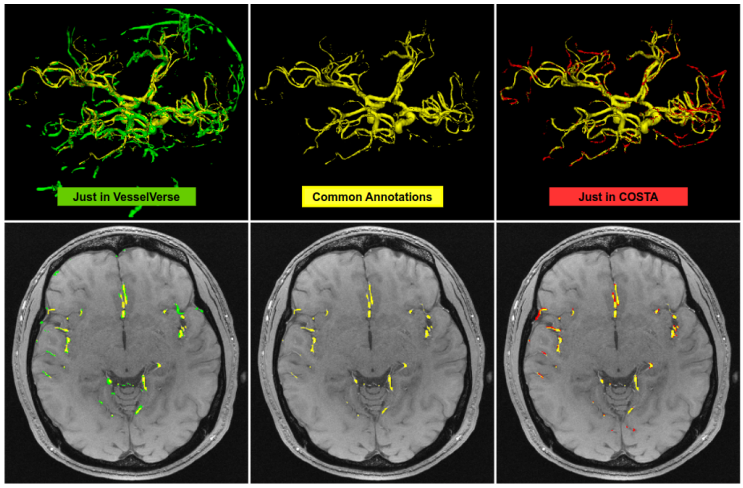

Fig. 1. Inter-rater variability in brain vessel annotations. Common labels are highlighted in yellow, VesselVerse in green, and COSTA [12] in red in TOF-MRA from IXI.

To address these limitations, we introduce `VesselVerse`, the largest release of brain vessel annotations to date, comprising three publicly available datasets totaling 950 annotated images from various neurovascular imaging modalities. `VesselVerse` also features a novel framework that systematically tracks annotation evolution, allows for multiple expert annotations and refinement, and integrates these through consensus generation. The `VesselVerse` dataset and framework can be accessed through the project's webpage[8].

2 Related Works

Neurovascular Imaging Databases. Large neurovascular imaging datasets like the UK Biobank (over 50,000 images) and OASIS-3 (about 3,000 images) lack annotations. Annotated datasets are much fewer (Table 1), and their differing annotation styles challenge their joint use for model training, something COSTA [12] has addressed partially. `VesselVerse` offers the largest collection of annotations, with multiple masks that enable users to select from different annotation protocols or create a consensus.

Crowdsourcing Initiatives. Crowdsourcing platforms have been used for large-scale medical imaging labeling, offering a rapid alternative to expert-driven methods [7,14]. However, studies show that crowd worker quality may be insufficient due to the need for domain expertise and standardized labeling guidelines[16]. `VesselVerse` implements a version-controlled repository that ensures the traceability of expert modifications. Changes of insufficient quality can be rejected, helping to maintain high standards for vessel annotation.

[8] URL: https://i-vesseg.github.io/vesselverse/

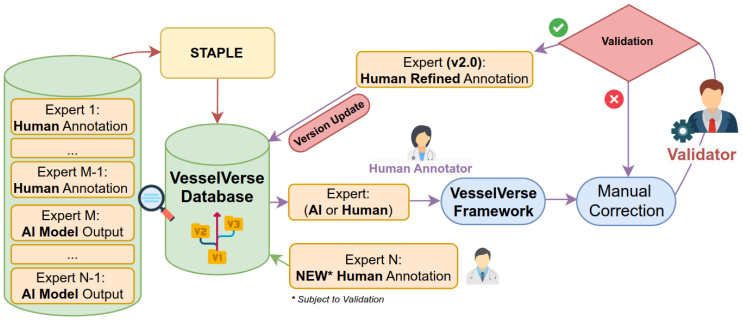

Fig. 2. Overview of the VesselVerse dataset and framework architecture.

Annotation platforms and tools. Tools like 3D Slicer [10], ITK-SNAP [24], and MITK offer visualization and annotation capabilities. 3D Slicer extensions leverage deep learning models to enhance manual annotation speed, such as MONAI Label [3], which enables interactive model-based labeling with real-time feedback. However, these frameworks lack version control, do not manage multiple annotations, and do not support consensus-based refinement. `VesselVerse` provides systematic tracking of changes, manages multi-expert annotations, and employs STAPLE-based consensus [21] generation for complete traceability.
Version Control. `VesselVerse` is akin to EXACT [11], a collaborative web-based platform for image annotation with version control capabilities. EXACT, however, is 2D-oriented, lacks 3D support and integrated consensus generation, and does not handle the 3D NIFTI format. In contrast, `VesselVerse` integrates with 3D Slicer, making it suitable for medical images while offering consensus generation, version control, and multi-expert collaboration.

3 VesselVerse Framework and Dataset

The `VesselVerse` dataset is supported by a methodological framework for managing and improving vessel annotations over time quality through a combination of automated methods and expert-validated annotations. Figure 2 presents an overview of the framework, which consists of three components: (1) Expert annotations, (2) a STAPLE-based consensus mechanism that allows to combine multiple expert annotations, and (3) a version control system for tracking annotation evolution.

3.1 The Framework

Expert Annotations. `VesselVerse` manages multiple expert annotations for a single image, allowing for various annotation protocols within the same dataset. For instance, annotation protocols may vary in several aspects, including the criteria for establishing vessel contours (e.g., specific windowing-level values to

set for visualization during annotation, inclusion/exclusion of calcifications), the types of vessels annotated (e.g., arteries only, veins only, or both), and the labeling conventions used (e.g., distinct labels for arteries and veins or a unique label). Similarly to the approach used by SMILE-UHURA [1], `VesselVerse` broadens the concept of expert annotators to encompass model-generated segmentations, thereby treating segmentations from any algorithm as expert-level annotations.

Consensus Generation. The `VesselVerse` framework provides a mechanism for generating *consensus annotations* (CA) that aggregate expert annotations from multiple sources. This functionality aims to offer an automated method for combining two or more available annotations, addressing situations where users are uncertain about which annotation set is most suitable. The consensus generation is based on STAPLE [21], a well-established expectation-maximization algorithm that, given n annotations associated with an image, computes a probabilistic estimate of the true annotation. In its E-step, STAPLE computes the sensitivity S_e and specificity S_p scores of each of the n annotations, representing the probability of correctly labeling a voxel as foreground and background, respectively. Subsequently, in its M-step, it updates a voxel-level probability map of consensus among the n annotations based on these parameters. The consensus annotation, which assigns lower weight to less accurate inputs, is ultimately obtained by thresholding the voxel values of the probability map after the EM algorithm's convergence.

Version Control. It has been widely demonstrated that dataset labels are often prone to errors, which can destabilize the results of models trained using supervised learning [13]. Public datasets currently do not provide mechanisms to report or correct identified annotation errors. `VesselVerse` addresses this by implementing collaborative refinement mechanisms that allow users to directly fix mistakes and refine annotations thanks to its integration as a 3D Slicer extension for vessel annotation revision. However, since refinement processes can inadvertently introduce new errors, `VesselVerse` also provides a version control system, making reverting to previous versions possible. Expert annotations in `VesselVerse` are systematically integrated into the version control system. Following expert correction and commit, each annotation is subjected to validation checks for annotation quality, verification of spatial consistency, metadata completeness, and version control compliance before final commitment to the repository.

Implementation Details. The `VesselVerse` framework is implemented as a 3D Slicer extension for vessel annotation revision, taking advantage of the advanced visualization and annotation tools. The extension provides version history visualization with metadata tracking. The version control system extends DVC with specialized structures for managing vessel annotations and their associated metadata. The framework implements three key components: (1) A hierarchical storage system that organizes images, corresponding expert annotations (each identified by a unique key), and version metadata into a structured directory tree. (2) A metadata tracking mechanism that maintains a comprehensive history for

Table 2. Summary of the VesselVerse dataset and available annotations.

			Available Annotations								
Dataset	**#Images**	**#Annot.**	MA	FF	nnU-Net	A2V	S-MIP	VB	JOB-VS	SB-AL	CA
IXI TOF-MRA	600	4,822	✓	✓	✓	✓	✓	✓	✓	✓	✓
TubeTK T1-MRA	100	800	✓	✗	✓	✓	✓	✓	✓	✓	✓
TopCoW MRA	125	1000	✓	✗	✓	✓	✓	✓	✓	✓	✓
TopCoW CT	125	635	✓	✓	✓	✓	✗	✗	✓	✗	✓

each annotation, including spatial properties, provenance information, expert annotations, and data integrity. (3) A secure multi-user collaboration protocol ensuring data integrity while enabling collaborative refinement by controlling access through credentialed authentication with distinct permissions for database access and contribution uploads. A comprehensive demo of the interface and its functionality is available in the supplementary material.

3.2 The Dataset

The images. `VesselVerse` builds upon publicly available neurovascular imaging datasets. These are: **IXI** comprising 600 Time-of-Flight Magnetic Resonance Angiographies (TOF-MRA) collected across three different hospitals (GUYS, HH, IOP) in London, UK. **TubeTK** containing 100 T1–MRA image pairs from healthy patients, acquired at the University of North Carolina using a 3T Siemens Allegra scanner. Only the MRA sequence is considered. **TopCoW** [23] containing 125 Computed Tomography angiographies (CTA) and MRA pairs from the University Hospital Zurich (i.e. 430 images in total).

The expert annotations. For each image of the `VesselVerse` collection, up to 9 expert annotations are available. Along with assisted manual annotations (MA) by an expert annotator, we include labels using a **Frangi filter** (FF)[4] (15 equally spaced scales with $\sigma_{min} = 0.2$, $\sigma_{max} = 2.0$). In addition, 6 models were trained to generate expert annotations. The considered models are: 1) **nnU-Net** [8] and **A2V** [6], which correspond to the only methods of the TopCoW 2023 challenge [23] capable of handling both MRA and CTA with a single model; 3) **SPOCKMIP** (S-MIP) [15], 4) **VesselBoost** (VB) [22] and 5) **JOB-VS** [19] pre-trained models available from the SMILE-UHURA challenge [1] that were publicly available; and 6) **StochasticBatchAL** (SB-AL) [5] an active learning model that iteratively refines segmentation outputs through stochastic selection. For the MRA images, models are trained using 22 manually annotated images from IXI dataset. For the CTA images, manual annotations obtained after correcting labels generated using the Frangi filter were used for training a nnU-Net and fine-tuning an A2V. Lastly, for reproducibility, we provide the STAPLE-based **consensus annotation** (CA) used for validation (Sec. 4). Nonetheless, we highlight that consensus annotations can always be generated through the framework's tools. Table 2 summarizes the contents of the `VesselVerse` dataset.

Table 3. Average star ratings for the top two methods, quality assessment and inter-rater agreement W, Kendall's τ comparing each method's ranking to the STAPLE consensus for each sub-dataset of VesselVerse.

Dataset	Best 2 Models	Star Rating (Avg ⭐)	Quality Score (Avg)	Kendall's W (Mean ± Std)	Kendall's τ (Mean ± Std)
IXI-GUYS	nnUNet STAPLE	4.50 4.42	3.75/5	0.760 ± 0.135	0.578 ± 0.050
IXI-HH	nnUNet STAPLE	4.33 4.08	3.87/5	0.609 ± 0.157	0.450 ± 0.312
IXI-IOP	nnUNet STAPLE	4.33 4.08	3.5/5	0.774 ± 0.149	0.656 ± 0.309
TopCoW-MR	STAPLE nnUNet	3.69 3.00	3.56/5	0.511 ± 0.074	0.450 ± 0.068
TopCoW-CT	JoB-VS STAPLE	4.03 3.34	3.33/5	0.552 ± 0.331	-0.407 ± 0.419
TubeTK	nnUNet SPOCKMIP	4.33 4.22	3.22/5	0.585 ± 0.105	0.311 ± 0.319

4 Validation

Setup. A neuroradiologist (JC), a neurologist (LL), a neurosurgery resident (MDR), and an expert annotator (MAZ) from 4 different institutions across 3 different countries (France, Italy and UK) were asked to evaluate annotations associated with 20 randomly selected images of `VesselVerse`. Precisely, five `VesselVerse` expert annotations were presented to the evaluators in a blinded manner, i.e., the source of each annotation was unknown. Each evaluator was asked to assign a 5 star-based rating to the presented annotations. These ratings reflect their impression of the overall quality of the annotations. Then, evaluators were asked to assign to their best-ranked annotation a separate 1–5 quality score (where 5 denotes higher quality) guided by anatomical correctness, completeness, and clinical validity. The experiment is repeated twice. In the first experiment, only annotations from models are included (**Experiment 1**). In the second experiment, the set of annotations included one manual annotation with the remaining ones obtained from models (**Experiment 2**). The experiment was conducted in IXI, thus we excluded those models that used IXI for training (nnUnet and A2V). A ranking of the annotations was extracted from the answers of the first round to establish a performance hierarchy. Inter-evaluator agreement of the rankings was assessed using Kendall's Coefficient of Concordance, *Kendall's W*. Additionally, we assessed the correlation between evaluators' rankings and the rankings obtained from the STAPLE consensus procedure (using S_e and S_p) using *Kendall's* τ coefficient (**Experiment 1**).

Experiment 1. Table 3 summarizes the results obtained in Experiment 1. We observe that the best two expert annotations, nnUnet, and the STAPLE consensus, are consistent across the subset of images in IXI (i.e., the different centers). This is also reflected in their higher average ratings and the highest Kendall's W coefficient values across all datasets. Additionally, at least one of them appears in the top 2 ranking for the remaining datasets. These results are in line with the

literature, highlighting the effectiveness of ensemble approaches like STAPLE and state-of-the-art architectures like nnU-Net.

Interestingly, as the inter-rater agreement drops in TopCoW and TubeTK, there is a drop in the quality scores. This may be an indicator of the higher difficulty of the task. For instance, E1 highlighted the variability of the quality of the TOF-MRA data as a major factor influencing the assessment. Another major factor influencing quality assessment involves the variations in the annotation protocol. SPOCKMIP, VesselBoost and JOB-VS are pre-trained models from the SMILE-UHURA dataset. As it can be seen in Table 1, SMILE-UHURA follows a different annotation protocol that includes extraparenchimal vessels. Evaluators that favored annotated vessels, in general, (E3 and E4) gave higher scores to those models, whereas those that favored standard annotation protocols (E1 and E2) penalized more those models. These findings highlight the relevance of having multi-expert annotations that can cover the needs from different users.

The assessment of STAPLE's capacity to reflect expect judgment, as indicated by Kendall's τ reveal a predominantly positive alignment, with average τ coefficients showing moderate to strong positive correlations across most datasets, indicating that STAPLE successfully reflects expert evaluation criteria. The negative correlation observed for TopCoW-CT can be attributed to this being the most challenging dataset with a low-quality score (3.33/5), high inter-rater variability, reflected in the Kendall's W score, and limited input with only three model-generated segmentations available. Overall, the consistent positive correlation in MRA datasets validates that STAPLE's consensus-based approach effectively captures the collective expertise of radiologists, validating its relevance as a consensus tool withing our framework.

Experiment 2. Manual annotations received consistently high ratings (4.00 ⭐) exceeding or comparable to automated methods. Only SPOCKMIP achieved a higher ranking (4.50 ⭐), with all other models ranking below the manual annotations. We explain this by the fact that SPOCKMIP extracts very small vessels that the human annotator misses, leading to higher scores from the evaluators. The latter points to one of `VesselVerse`'s strong points: the inclusion of annotations provided by users and models. Although we did not assess it, we hypothesize that generating a consensus agreement between the manual labels and those from SPOCKMIP should lead to higher-quality annotated images.

5 Conclusion

In this paper, we presented `VesselVerse`, the largest public release of brain vessel annotations to date, with 950 images across three public datasets and various imaging modalities, which offers multiple annotations per image to accommodate diverse annotation protocols. This approach acknowledges annotation subjectivity while providing access to diverse annotation styles in a unified framework. Notably, `VesselVerse` is not just limited to a dataset. It encompasses a collaborative framework supporting collaborative refinement and version control, as well as consensus agreement tools to combine the available set of annotations. In

this way, we transform static data collections into dynamic, living datasets, enabling continuous refinement while maintaining complete traceability of changes and providing quality control without sacrificing the benefits of crowdsourcing approaches. While the current release utilizes a single label despite the dataset containing both veins and arteries, `VesselVerse`'s dynamic formulation enables future updates, allowing for multi-label annotations through community collaboration, thus further enhancing its utility.

Finally, `VesselVerse`'s modality-agnostic architecture with multi-expert annotation, consensus, and version control has the potential to be used for other medical imaging annotation tasks, making it a versatile resource for the broader community. Thus, we believe `VesselVerse` represents a step towards addressing the constant need for data annotation.

Acknowledgments. This work is partly funded by the ANR JCJC project I-VESSEG (22-CE45-0015-01), the 3IA Côte d'Azur Investments project (ANR-23-IACL-0001), and ERC CoG CARAVEL (101171357). KY and BM acknowledge support from the Helmut Horten Foundation. LL and BL are partly funded by France Life Imaging (ANR-11-INBS-0006).

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

References

1. Chatterjee, S., Mattern, H., Dörner, M., Sciarra, A., Dubost, F., Schnurre, H., Khatun, R., Yu, C.C., Hsieh, T.L., Tsai, Y.S., et al.: SMILE-UHURA Challenge–Small Vessel Segmentation at Mesoscopic Scale from Ultra-High Resolution 7T Magnetic Resonance Angiograms. arXiv preprint arXiv:2411.09593 (2024)
2. Chen, Y., Jin, D., Guo, B., Bai, X.: Attention-assisted adversarial model for cerebrovascular segmentation in 3D TOF-MRA volumes. IEEE Transactions on Medical Imaging **41**(12), 3520–3532 (2022)
3. Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., et al.: MONAI label: A framework for AI-assisted interactive labeling of 3D medical images. Medical Image Analysis **95**, 103207 (2024)
4. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Medical image computing and computer-assisted intervention—MICCAI. pp. 130–137 (1998)
5. Gaillochet, M., Desrosiers, C., Lombaert, H.: Active learning for medical image segmentation with stochastic batches. Medical Image Analysis **90**, 102958 (2023)
6. Galati, F., Falcetta, D., Cortese, R., Casolla, B., Prados, F., Burgos, N., Zuluaga, M.A.: A2V: A Semi-Supervised Domain Adaptation Framework for Brain Vessel Segmentation via Two-Phase Training Angiography-to-Venography Translation. In: 34th British Machine Vision Conference 2023, BMVC (2023)
7. Heim, E., Roß, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., Sommer, G., Sauter, A.W., et al.: Large-scale medical image annotation with crowd-powered algorithms. Journal of Medical Imaging **5**(3), 034002–034002 (2018)

8. Isensee, F., Jaeger, P.F., Kohl, S.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 (2021)
9. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in CT. European radiology **29**, 1391–1399 (2019)
10. Kikinis, R., Pieper, S.D., Vosburgh, K.G.: 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support. In: Intraoperative imaging and image-guided therapy, pp. 277–289 (2013)
11. Marzahl, C., Aubreville, M., Bertram, C.A., Maier, J., Bergler, C., Kröger, C., Voigt, J., Breininger, K., Klopfleisch, R., Maier, A.: EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. Scientific reports **11**(1), 4343 (2021)
12. Mou, L., Yan, Q., Lin, J., Zhao, Y., Liu, Y., Ma, S., Zhang, J., Lv, W., Zhou, T., Frangi, A.F., et al.: COSTA: A Multi-center TOF-MRA Dataset and A Style Self-Consistency Network for Cerebrovascular Segmentation. IEEE transactions on medical imaging (2024)
13. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks (2021)
14. Ørting, S.N., Doyle, A., van Hilten, A., Hirth, M., Inel, O., Madan, C.R., Mavridis, P., Spiers, H., Cheplygina, V.: A survey of crowdsourcing in medical image analysis. Human Computation **7**, 1–26 (2020)
15. Radhakrishna, C., Chintalapati, K.V., Kumar, S.C.H.R., Sutrave, R., Mattern, H., Speck, O., Nürnberger, A., Chatterjee, S.: SPOCKMIP: Segmentation of Vessels in MRAs with Enhanced Continuity using Maximum Intensity Projection as Loss. arXiv preprint arXiv:2407.08655 (2024)
16. Rädsch, T., Reinke, A., Weru, V., Tizabi, M.D., Heller, N., Isensee, F., Kopp-Schneider, A., Maier-Hein, L.: Quality Assured: Rethinking Annotation Strategies in Imaging AI. In: European Conference on Computer Vision. pp. 52–69 (2024)
17. Renard, F., Guedria, S., Palma, N.D., Vuillerme, N.: Variability and reproducibility in deep learning for medical image segmentation. Scientific Reports **10**(1), 13724 (2020)
18. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al.: UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine **12**(3), e1001779 (2015)
19. Valderrama, N., Pitsiorlas, I., Vargas, L., Arbeláez, P., Zuluaga, M.A.: JOB-VS: Joint brain-vessel segmentation in TOF-MRA images. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023)
20. van Walsum, T., Schaap, M., Metz, C.T., van der Giessen, A.G., Niessen, W.J.: Averaging centerlines: mean shift on paths. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 900–907 (2008)
21. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)
22. Xu, M., Ribeiro, F.L., Barth, M., Bernier, M., Bollmann, S., Chatterjee, S., Cognolato, F., Gulban, O.F., Itkyal, V., Liu, S., Mattern, H., Polimeni, J.R., Shaw, T.B., Speck, O., Bollmann, S.: VesselBoost: A Python Toolbox for Small Blood Vessel Segmentation in Human Magnetic Resonance Angiography Data. Aperture Neuro **4** (2024)

23. Yang, K., Musio, F., Ma, Y., Juchler, N., Paetzold, J.C., Al-Maskari, R., Höher, L., Li, H.B., Hamamci, I.E., Sekuboyina, A., Shit, S., Huang, H., Waldmannstetter, D., Kofler, F., Navarro, F., Menten, M., Ezhov, I., Rueckert, D., Vos, I., Ruigrok, Y., Velthuis, B., Kuijf, H., Hämmerli, J., Wurster, C., Bijlenga, P., Westphal, L., Bisschop, J., Colombo, E., Baazaoui, H., Makmur, A., Hallinan, J., Wiestler, B., Kirschke, J.S., Wiest, R., Montagnon, E., Letourneau-Guillon, L., Galdran, A., Galati, F., Falcetta, D., Zuluaga, M.A., Lin, C., Zhao, H., Zhang, Z., Ra, S., Hwang, J., Park, H., Chen, J., Wodzinski, M., Müller, H., Shi, P., Liu, W., Ma, T., Yalçin, C., Hamadache, R.E., Salvi, J., Llado, X., Estrada, U.M.L.T., Abramova, V., Giancardo, L., Oliver, A., Liu, J., Huang, H., Cui, Y., Lin, Z., Liu, Y., Zhu, S., Patel, T.R., Tutino, V.M., Orouskhani, M., Wang, H., Mossa-Basha, M., Zhu, C., Rokuss, M.R., Kirchhoff, Y., Disch, N., Holzschuh, J., Isensee, F., Maier-Hein, K., Sato, Y., Hirsch, S., Wegener, S., Menze, B.: TopCoW: Benchmarking Topology-Aware Anatomical Segmentation of the Circle of Willis (CoW) for CTA and MRA. arXiv:2312.17670 (2024)
24. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006)