MLV²-Net: Rater-Based <u>M</u>ajority-<u>L</u>abel <u>V</u>oting for Consistent <u>M</u>eningeal <u>L</u>ymphatic <u>V</u>essel Segmentation

Fabian Bongratz[a,b]   FABI.BONGRATZ@TUM.DE
Markus Karmann[a]   MARKUS.KARMANN@TUM.DE
Adrian Holz[a]   ADRIAN.HOLZ@TUM.DE
Moritz Bonhoeffer[a]   MORITZ.BONHOEFFER@TUM.DE
Viktor Neumaier[a]   VIKTOR.NEUMAIER@TUM.DE
Sarah Deli[a]   SA.DELI@TUM.DE
Benita Schmitz-Koep[a]   BENITA.SCHMITZ-KOEP@TUM.DE
Claus Zimmer[a]   CLAUS.ZIMMER@TUM.DE
Christian Sorg[a]   CHRISTIAN.SORG@TUM.DE
Melissa Thalhammer[a]   MELISSA.THALHAMMER@TUM.DE
Dennis M. Hedderich[a]   DENNIS.HEDDERICH@TUM.DE
Christian Wachinger[a,b]   CHRISTIAN.WACHINGER@TUM.DE

[a] *School of Medicine and Health, Technical University of Munich, Munich, Germany*
[b] *Munich Center for Machine Learning, Munich, Germany*

Abstract

Meningeal lymphatic vessels (MLVs) are responsible for the drainage of waste products from the human brain. An impairment in their functionality has been associated with aging as well as brain disorders like multiple sclerosis and Alzheimer's disease. However, MLVs have only recently been described for the first time in magnetic resonance imaging (MRI), and their ramified structure renders manual segmentation particularly difficult. Further, as there is no consistent notion of their appearance, human-annotated MLV structures contain a high inter-rater variability that most automatic segmentation methods cannot take into account. In this work, we propose a new rater-aware training scheme for the popular nnU-Net model, and we explore rater-based ensembling strategies for accurate and consistent segmentation of MLVs. This enables us to boost nnU-Net's performance while obtaining explicit predictions in different annotation styles and a rater-based uncertainty estimation. Our final model, *MLV²-Net*, achieves a Dice similarity coefficient of 0.806 with respect to the human reference standard. The model further matches the human inter-rater reliability and replicates age-related associations with MLV volume.

Keywords: Meningeal lymphatic vessels, Glymphatic system, Segmentation, Inter-rater variability

Data and Code Availability In Table 1, we provide an overview of the data used in this study. As no public data on MLV structures in MRI is available, we assembled a custom segmentation dataset comprising $n = 33$ labeled and $n = 22$ unlabeled 3D fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MR) images from cognitively normal subjects. To this end, four neuroanatomical experts annotated MLV structures individually along the superior sagittal sinus (SSS) in anterior, middle, and posterior brain regions, resulting in 3×7 and 1×6 annotations per rater. In addition, two images were annotated by all raters, which allows us to assess their inter-rater reliability (IRR). We keep another held-out test set comprising four more images to evaluate model accuracy. All raters annotated those images jointly for best consistency after the annotation of the 27 + 2 scans was finished. Finally, we use 22 raw images to evaluate our model indirectly by replicating known age-related associations with MLV volume. All images have a resolution of 0.5×0.5×1mm³ and were acquired using simultaneous trimodal PET-MR-EEG imaging (Del Guerra et al., 2018). We will make code and trained models publicly available at https://github.com/ai-med/mlv2-net.

Institutional Review Board (IRB) Imaging data comes from two studies. Both studies were approved by the Ethics Review Board of the Klinikum
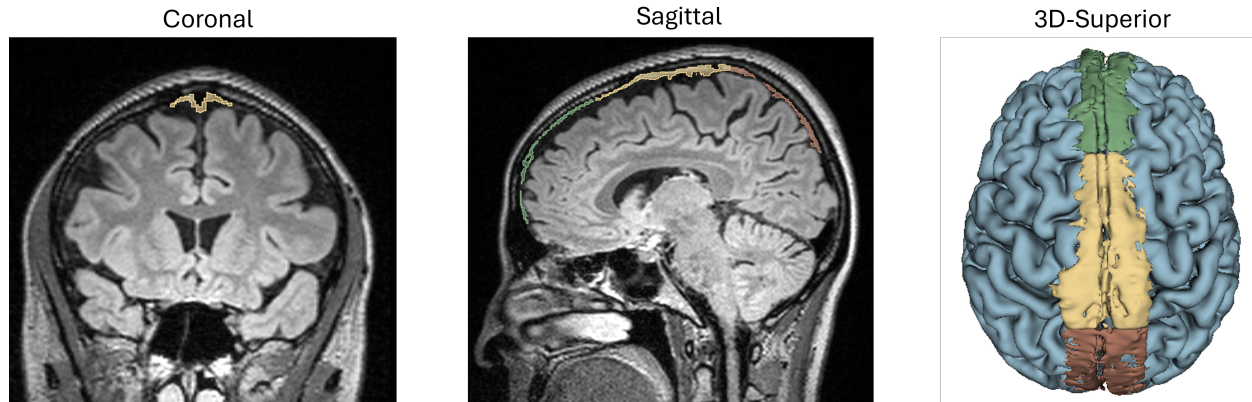
Figure 1: Exemplary segmentation of meningeal lymphatic vessels (MLVs) in coronal and sagittal planes — separated into anterior, middle, and posterior regions. On the right, we show the corresponding 3D surface representation rendered from a superior viewpoint.

Rechts der Isar, Technical University of Munich, Germany (5611/12, 338/20-S). Approvals to administer radiotracers for both studies were obtained from the Administration of Radioactive Substances (Bundesamt fuer Strahlenschutz), Germany (Z 5-22461/2 – 2014-010, Z 5- 22464/2020-199-G).

1. Introduction

The lymphatic system — part of the immune system and responsible for the drainage of waste products — stretches across the entire human body and can often be found alongside blood vessels of the circulatory system. In the brain, the glymphatic system (Iliff et al., 2012) takes a similar role in that it clears waste products. To this end, meningeal lymphatic vessels (MLVs), located alongside the dural venous sinuses, transfer interstitial fluids and macromolecules to deep cervical lymph nodes (Louveau et al., 2015). An impairment in the MLVs' functionality, potentially coupled with morphological changes such as thickening, has been linked to aging (Albayram et al., 2022) as well as to clinical conditions like Alzheimer's (Goodman et al., 2018), multiple sclerosis (Louveau et al., 2018), and Parkinson's disease (Ding et al., 2021). Yet, MLVs have only recently been described in 3D FLAIR MRI (Albayram et al., 2022), and their segmentation has only been done manually so far. However, manual annotation of MLVs is difficult and time-consuming due to their ramified structure, cf. Figure 1. Moreover, the training of automatic segmentation models on expert-annotated data is challenging due to the high inter-rater variability.

Related work Deep neural networks for medical image segmentation are commonly trained to remove this variablility (Guo et al., 2024; Hatamizadeh et al., 2022; Ronneberger et al., 2015). However, this approach does not model the reality where disagreement about the true contours of a structure often exists (Warfield et al., 2004). This issue is especially problematic for newly discovered structures, such as MLVs, which bear enormous potential for innovative findings but for which a common notion of their appearance does not (yet) exist. Notably, a few dedicated methods for rater-aware segmentation were developed (Kohl et al., 2018; Mirikharaji et al., 2021; Warfield et al., 2004; Zhang et al., 2023). These approaches yielded effective results for certain standard applications, e.g., skin lesion (Mirikharaji et al., 2021) or brain tumor segmentation (Zhang et al., 2023), but transferring them to new tasks is difficult due to the large number of hyperparameters involved. These choices are non-trivial, not reproducible, and subject to the developer's experience and preferences (Isensee et al., 2021). At the same time, the best segmentation results are typically obtained with nnU-Net (Isensee et al., 2024), which provides a versatile framework for hyperparameter selection. Unfortunately, nnU-Net cannot model the variability in segmentations provided by different raters — a functionality essential for trustworthy and comprehensible clinical predictions. We close this gap and develop a rater-based ensembling strategy for nnU-

Net that keeps its architecture intact and augments it with the ability to replicate individual raters' annotation styles.

Contribution We present the first automatic method for segmentation of MLVs from 3D FLAIR MRI. To achieve accurate and reliable segmentation of the ramified structure, we made the following technical contributions. First, we developed an innovative rater-aware training scheme for the popular nnU-Net model that takes into account the different raters involved in the creation of the training set. This enables nnUNet to learn individual raters' segmentation styles and to explicitly predict a set of plausible segmentations. In a second step, we aggregate the predictions with a weighted majority-label voting scheme for best segmentation accuracy. In addition, we obtain a rater-based uncertainty prediction from the model. Finally, since the volume of MLVs is usually of utmost importance for downstream analyses, we derive error boundaries of the model's predicted volumes with respect to the ground-truth volume.

2. Methods

2.1. MLV²-Net architecture

Figure 2 shows an overview of MLV²-Net, which stands for rater-based majority-label-voting-network for meningeal lymphatic vessels. MLV²-Net builds upon nnU-Net (Isensee et al., 2021) and takes a 3D image of shape $H \times W \times D$ as input. In addition, we incorporate a unique encoding of the rater (rater encoding) of the same shape as the image. The rationale behind this encoding is that it provides relevant information about the rater, in our case a neuroanatomical expert who provides annotations of MLVs, without any architectural changes that might derange nnU-Net's hyperparameter search strategy. As output, the network yields voxel-wise segmentation maps, separated by foreground class and rater. Eventually, all predictions are aggregated via weighted majority-label voting as shown in Figure 3. Apart from the input and output, we keep nnU-Net intact; hence, we benefit from its structured parameter selection and obtain a reproducible setup.

2.2. Rater-aware training and inference

During the training and inference of MLV²-Net, we consider the different raters in the input and the output of the model.

Rater as input To enable the network to learn the styles of different raters from the training data, we provide this information as input to the network. Technically, we assign a zero-centered one-hot-encoded ID to each rater and concatenate it as additional channels to the input image volume as depicted in Figures 2 and 3. Namely, we assign the four raters in our setting the codes [1, 0], [-1, 0], [0, 1], and [0, -1]. In general, this scheme leads to $R/2$ additional input channels for R raters, which can be well processed by nnU-Net's initial convolutional layer. Importantly, we disable the per-channel z-score normalization in nnU-Net for the rater-encoding channels, which would set them to zero and essentially erase the rater information.

Rater as output To enforce the model to consider the rater, i.e., to coerce the network to predict the correct structure *as annotated by a certain rater*, we create rater-specific foreground labels for the loss computation (a combination of cross-entropy and Dice loss as in nnUNet). These are also the labels predicted by MLV²-Net during inference. Specifically, we use labels "Anterior MLV/Rater 1", "Anterior MLV/Rater 2", "Anterior MLV/Rater 3", and "Anterior MLV/Rater 4" instead of one label "Anterior MLV". Likewise for labels "Middle MLV" and "Posterior MLV". This results in $RF + 1$ labels, i.e., the cartesian product of R rater and F foreground labels, and the background. We do not create rater-specific background labels, as this would be redundant.

2.3. Weighted majority-label voting

To aggregate the segmentation maps in the styles of different raters, we employ a weighted majority-label voting, which we illustrate in Figure 3. As shown, we multiply the number of votes from rater-specific foreground predictions by a weight $w_{fg} > 1$. Compared to the standard majority vote ($w_{fg} = 1$), this increases the sensitivity to foreground labels, which we found to correspond best to human consensus decision-making (cf. Section 3.3). In the rare case of a tie, we choose the class with the lowest index. By default, nnU-Net also creates an ensemble of models via cross-validation and aggregates the voxel-wise mean of the predicted logits. We keep this mechanism untouched, i.e., cross-validation ensembling is part of the nnU-Net blocks in Figure 3, and compute the majority-label vote externally, treating each cross-validation ensemble as one voting model.

Table 1: Composition of the annotated and raw datasets used in this study.

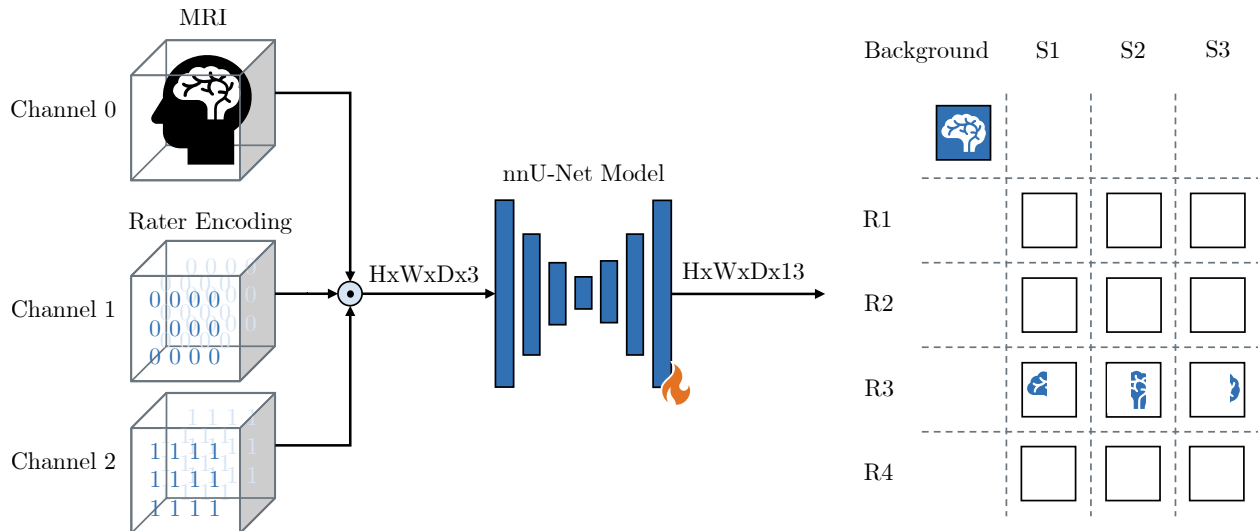| Name	Used for	Joint annot. by all raters	#Annotations per image	#Images
Training set	Training & validation	✗	1	27
IRR set	Testing (inter-rater reliability)	✗	4	2
Consensus set	Testing (accuracy)	✓	1	4
Raw set	Testing (downstream analysis)	N/A	N/A	22

Figure 2: Illustration of MLV²-Net. It augments nnU-Net with a rater-specific encoding and yields rater-aware segmentations as output. '⊙' denotes a channel-wise concatenation of inputs. In this example, we show the encoding and segmentation output figuratively for rater 3 (R3) and three foreground segmentation labels ($S1 - S3$). We train nnU-Net from scratch, as indicated by the flame.

2.4. Rater-based uncertainty

Apart from the consensus prediction, we obtain a rater-based uncertainty map in MLV²-Net. The uncertainty is based on the agreement of rater-based predictions of the model, i.e., the uncertainty is higher the more rater-based predictions speak against the majority label for a certain voxel. This provides us with an estimate of the reliability of the prediction, which renders the model more faithful as it can be used to detect potential failure cases. Unlike alternative uncertainty estimation methods, e.g., Monte Carlo Dropout (Gal and Ghahramani, 2016), MLV²-Net does not require a variational network architecture but keeps nnU-Net overall intact. Another advantage of our explicit rater-based modeling is that individual, potentially flawed, or deprecated segmentation styles can easily be ignored post hoc, i.e., without re-training. This is typically impossible

with variational approaches that implicitly model the data variability.

2.5. Boundaries on segmented volume based on Dice

The performance of segmentation models is commonly evaluated with the Dice similarity coefficient (DSC). However, in the end, the segmented volume is often of utmost importance in medical imaging. Therefore, in the following, we derieve error boundaries on the predicted volume relative to the ground-truth or reference volume.

Theorem 1 *Given the Dice similarity coefficient (DSC) of a segmentation model, the predicted volume relative to the ground-truth volume is bounded by $\frac{2}{DSC} - 1$ from above and by $\frac{2}{2-DSC} - 1$ from below.*
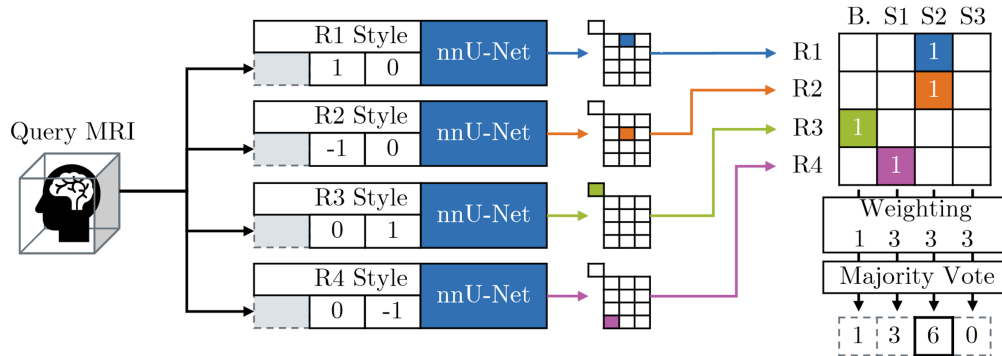
Figure 3: Illustration of the weighted majority-label voting in MLV^2-Net for four raters (R1–R4). Shown is an exemplary ambiguous prediction for a single voxel. R3 would segment the voxel as background (B), R4 as foreground segment 1 (S1), and R1 and R2 as foreground segment 2 (S2). The foreground weight in the example is set to $w_{\mathrm{fg}} = 3$.

Proof. The DSC and the relative predicted volume V^{rel} can be calculated from a confusion matrix comprising false negative (FN), true positive (TP), false positive (FP), and true negative (TN) voxels. By definition, the DSC is calculated as

$$DSC = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} = \frac{2 \cdot TP}{1 + TP + FP}, \quad (1)$$

where we utilized the fact that $TP + FN \equiv 1$ when normalized to the ground-truth (GT) volume. The relative predicted volume, i.e., the predicted relative to the GT volume, is given by

$$V^{\mathrm{rel}} = \frac{\text{Predicted volume}}{\text{GT volume}} = \frac{TP + FP}{TP + FN} = TP + FP. \quad (2)$$

Rearranging Equation (1) to $FP = \frac{2 \cdot TP}{DSC} - TP - 1$ and inserting it into Equation (2) yields

$$V^{\mathrm{rel}} = \frac{2 \cdot TP}{DSC} - 1. \quad (3)$$

From $TP \leq 1$, we obtain $V^{\mathrm{rel}} \leq \frac{2}{DSC} - 1$. Similarly, we get $V^{\mathrm{rel}} \geq \frac{2}{2-DSC} - 1$ by rearranging Equation (1) to $TP = \frac{DSC \cdot (FP+1)}{2-DSC}$ and $FP \geq 0$. ∎

3. Results

3.1. Experimental setting

We implemented MLV^2-Net into nnU-Net (v2, 3D Fullres), based on Python (v3.11), PyTorch (v2.1), and CUDA (v12.1). As there is no reference method for automatic segmentation of MLVs, we compare MLV^2-Net to a diverse set of baseline methods. Namely, we implemented a registration-based segmentation propagation algorithm (Modat et al., 2009) (SegProp) that aggregates all training references through an optimized threshold, UniverSeg (Butoi et al., 2023), a recent foundation model for medical image segmentation that we adapted for 3D images by fusing overlapping patches from all three image planes, and the standard nnU-Net configurations (2D and 3D Fullres) (Isensee et al., 2021). In addition, we implemented an ensemble of separate, rater-specific models (not to be confused with nnU-Net's cross-validation-based ensembling strategy), and we ablate the weighted majority-label voting and the rater-specific labels. We ran all methods consistently on a single Nvidia GeForce RTX 3090 graphics card with 24GB VRAM. All experiments were conducted with the data described in the initial paragraph about data and code availability.

3.2. Inter-rater reliability and rater-based uncertainty

As a measure of inter-rater reliability (IRR), we compute a Fleiss' kappa score (Fleiss, 1971) based on our IRR dataset. This dataset contains an annotation from each rater for each image. Considering all three foreground labels as a single entity, we obtain a Fleiss' kappa of $\kappa = 0.73/0.79$ for the two IRR images, respectively. The ensemble of separate, rater-specific nnU-Net models closely replicates the expert raters' IRR ($\kappa = 0.74/0.80$). With the single-model

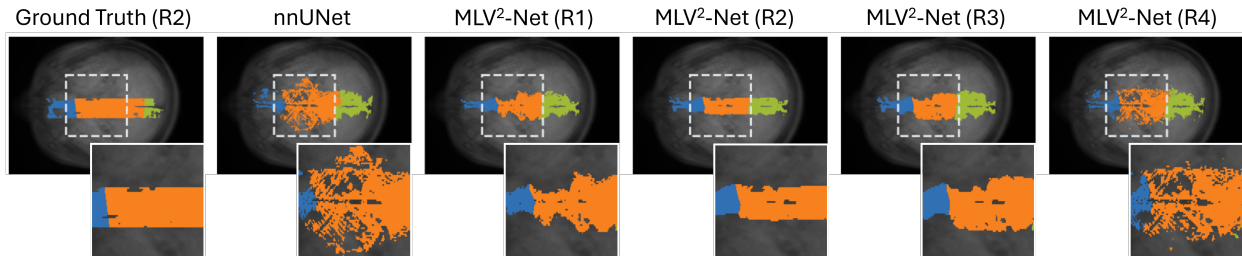| Ground Truth (R2) | nnUNet | MLV²-Net (R1) | MLV²-Net (R2) | MLV²-Net (R3) | MLV²-Net (R4) |

Figure 4: Ground-truth annotation from rater 2 (R2) and corresponding predictions from nnUNet and MLV²-Net with different rater queries (R1–R4). We show projections along the vertical axis for best visibility; note that this makes the vessels appear thicker but better displays segmentation characteristics than slices.

Table 2: Accuracy of all implemented methods in terms of Dice similarity coefficient (DSC). We report results (mean ± SD) from the cross-validation and the held-out consensus test set in anterior, middle, and posterior regions and foreground (all regions combined). **Best** and <u>second</u> results are highlighted.

| Method | Variant | Cross-validation | | | Consensus set |
		Anterior	Middle	Posterior	Foreground
MLV²-Net	$w_{\text{fg}} = 3$	<u>0.626 ± 0.108</u>	<u>0.709 ± 0.092</u>	<u>0.687 ± 0.120</u>	**0.806 ± 0.030**
MLV²-Net	Oracle	**0.638 ± 0.151**	**0.712 ± 0.094**	**0.688 ± 0.122**	-
nnU-Net	3D Fullres	0.593 ± 0.117	0.689 ± 0.098	0.682 ± 0.116	<u>0.787 ± 0.046</u>
nnU-Net	2D	0.618 ± 0.106	0.707 ± 0.099	0.683 ± 0.109	0.760 ± 0.024
UniverSeg	All planes	0.341 ± 0.106	0.459 ± 0.159	0.490 ± 0.114	0.529 ± 0.177
UniverSeg	Coronal	0.262 ± 0.090	0.437 ± 0.135	0.401 ± 0.100	0.498 ± 0.130
UniverSeg	Sagittal	0.312 ± 0.113	0.411 ± 0.148	0.413 ± 0.124	0.427 ± 0.187
UniverSeg	Transverse	0.275 ± 0.102	0.318 ± 0.128	0.366 ± 0.107	0.417 ± 0.113
SegProp	Optimized	0.350 ± 0.137	0.445 ± 0.104	0.396 ± 0.110	0.493 ± 0.079
SegProp	Standard	0.200 ± 0.078	0.291 ± 0.072	0.228 ± 0.080	0.294 ± 0.065

MLV²-Net, however, we obtain a higher agreement ($\kappa = 0.79/0.82$).

In Figure 4, we show four rater-specific predictions of MLV²-Net for an exemplary scan alongside the annotation of R2. The R2-specific prediction corresponds well to the raters' reference. It can also be seen that conditioning the model on the other raters (R1, R3, R4) yields a reasonable variability in the prediction. The vanilla nnU-Net, however, is not capable of modeling this variability and produces only a single prediction.

In Figure 5, we show the rater-based prediction uncertainty of MLV²-Net for the two images in our IRR set and compare it to the voxel-wise inter-rater variability. Qualitatively, the same boundary regions are subject to inter-rater variability and prediction uncertainty. This indicates that the rater-based model uncertainty matches the actual variability in the an-

notated data locally, thereby supporting the globally computed quantitative IRR results and the qualitative inspection from above.

3.3. Accuracy and consensus decision-making

While annotation variability among human raters is natural and unavoidable, most applications demand consistent segmentation. In our held-out consensus test set, we tried to remove the variability as much as possible through all four raters' joint annotation of the images. This is the reference standard for consensus decision-making but, unfortunately, it is only feasible for a few images. In Table 2, we report the average accuracy of all implemented methods on this consensus set and in a 5-fold cross-validation on our training set, separated by anterior, middle, and posterior regions. Qualitative predictions are in Figure 6.
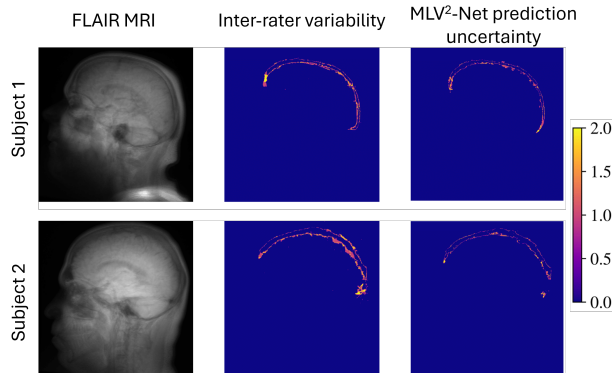
Figure 5: Inter-rater variability between human experts and in MLV²-Net. We plot the disagreement between raters, i.e., a score of one implies that at least one rater disagrees, two means it is a tie, and project voxel-wise values to the sagittal plane. Shown are the two IRR test cases for which individual annotations from all four raters are available.

Table 3: Ablation study of training and consensus-finding strategies. Using our consensus set, we report the mean $\pm$ SD Dice similarity coefficient (DSC).

Configuration	DSC
Standard majority vote ($w_{fg} = 1$)	0.790 ± 0.030
Weighted Majority Vote ($w_{fg} = 4$)	0.790 ± 0.043
No rater-specific labels ($w_{fg} = 3$)	0.800 ± 0.032
Rater-specific models ($w_{fg} = 3$)	0.804 ± 0.029
Weighted majority vote ($w_{fg} = 2$)	0.805 ± 0.031
Weighted majority vote ($w_{fg} = 3$)	0.806 ± 0.030

The models evaluated on the consensus set are ensembles of the five cross-validation models.

First, it stands out that MLV²-Net yields the highest accuracy during cross-validation and on the consensus set, followed by the vanilla nnU-Net. Surprisingly, UniverSeg, which was pre-trained on more than 22K medical scans, is not competitive with supervised learning from scratch on our comparably small training set — even after deliberate optimization. Visually, SegProp also produces reasonable MLV segmentations but is not competitive in terms of quantitative measures. All models achieve higher Dice scores on the consensus set and sacrifice accuracy in the cross-validation, where the best accuracy is obtained with the respective rater as an input to MLV²-Net (oracle). This is likely due to the higher annotation variability in the training set compared to the consensus set. Nonetheless, it is noteworthy that MLV²-Net learns to produce highly accurate and consistent segmentations from training data with non-neglectable annotation variability. In Figure 7, we plot the accuracy of MLV²-Net ($w_{fg} = 3$) and the relative segmented volume in all annotated test samples and from the cross-validation. All results lie within the theoretically derived error bounds. On average, the relative predicted volume can be assumed to be between 0.67 and 1.49 on the consensus set (mean $DSC = 0.806$).

3.4. Ablation study

From the ablation study in Table 3, we infer that the most important design choice in MLV²-Net is the weighted majority-label voting. It makes the model more sensitive to foreground voxels than the standard, equally weighted majority vote. Other methodological choices, such as using a single-model approach and rater-specific labels, seem to have only a minor positive effect on segmentation accuracy.

For our setting with four raters, we can deduce explicit segmentation thresholds in dependence of the foreground weight w_{fg}. In words, $w_{fg} = 3$ means that a voxel is segmented as foreground if no more than two out of the four rater-specific predictions anticipate it to be background (with three background votes, the voxel is predicted as background due to our policy to choose the lower-label index in case of a tie, cf. Section 2.3). With $w_{fg} = 2$, two votes on the background can only be overruled if the other two votes are on the same MLV sub-label (anterior, middle, posterior), which makes it slightly less foreground-sensitive than $w_{fg} = 3$. Increasing the sensitivity to foreground votes further by setting $w_{fg} = 4$ reduces the performance to the standard majority vote. Thus, we deduce that a moderately increased foreground sensitivity emulates human consensus-finding best based on the given data.

3.5. Downstream analysis of MLV volume

Finally, we apply our model in a downstream analysis of MLV volume using unlabeled imaging data. Recently, Albayram et al. (2022) found a positive association of age with MLV volume based on manual annotations. Using our MLV²-Net model, we can replicate this finding based on a group of adults ($n = 4$, age 51-62) and a larger young reference cohort ($n = 18$,
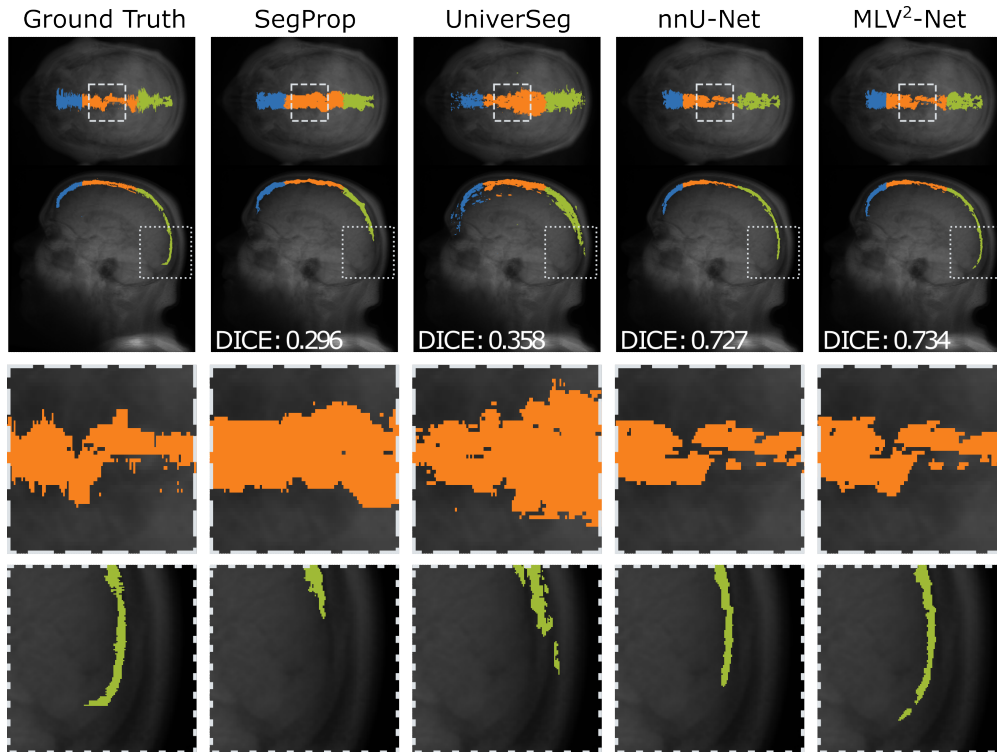
Figure 6: Predictions of all implemented methods based on an image from our consensus test set. The first row shows the orthogonal projections onto axial (top) and sagittal (bottom) planes. We reduced the brightness of the FLAIR image to highlight the segmentation details.
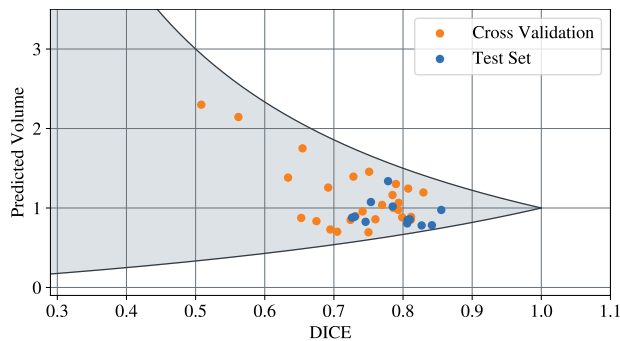

Figure 7: Dice similarity coefficient vs. relative predicted volume for the cross-validation and annotated test sets. We also show the theoretically derived boundaries on the predicted volume.
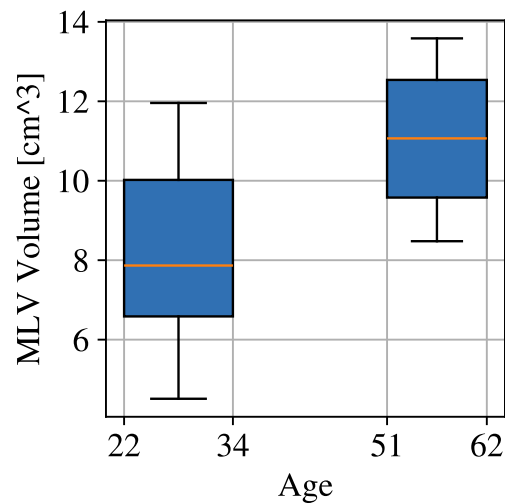

Figure 8: Box plots of the MLV volume predicted by MLV²-Net for two age groups. Lines indicate the median, boxes span the inter-quartile range (IQR), and whiskers extend to 1.5×IQR.

age 22-34), see Figure 8. The difference is significant based on $p < 0.05$ (two-sided t-test). This result indirectly confirms the accuracy of our model and proves its applicability for analyzing real-world study data.

4. Discussion

MLV segmentation task The task of MLV segmentation is new and challenging due to the ramified structure, the thin diameter, and the high inter-rater variability among experts on the voxel level. The difficulty of segmenting these structures is reflected in the high inter-rater variability in the expert-annotated data (Fleiss' kappa of $\kappa = 0.73/0.79$), cf. Section 3.2. According to Landis and Koch (1977), this corresponds to a *substantial agreement* ($0.6 < \kappa \leq 0.8$), which is inferior to *perfect agreement* ($\kappa > 0.8$). Nonetheless, we achieved a high accuracy of $DSC = 0.806$ on our consensus test set. As no established baselines exist for this task, we tried to cover various methods ranging from segmentation propagation over supervised learning to recent foundation models, cf. Section 3.3. Yet, future research should investigate and compare alternative model architectures and training paradigms to draw a more complete picture of the task at hand.

Dataset size We are aware that the number of annotated scans ($n = 33$) used in this study is comparably small, especially when compared to recent segmentation datasets with annotations of thousands of anatomies (Wasserthal et al., 2023). Yet, datasets with around 30 annotated scans are not uncommon for 3D medical image segmentation (Antonelli et al., 2022). In fact, creating much larger segmentation datasets manually for MLV structures is impossible due to their complex shape, the low contrast even in FLAIR imaging, and the required high resolution of 0.5mm in sagittal and vertical axes. In our experiments, we tried to account for the small dataset size by tuning hyperparameters based on extensive cross-validation on the training set. Moreover, we indirectly assessed our model's performance on $n = 22$ unannotated scans (cf. Section 3.5) by replicating known age-related associations with MLV volume. Finally, we put particular effort into assessing the inter-rater reliability (cf. Section 3.2) and created a consensus test set to ensure the used annotations are of high quality.

Foreground bias in ensemble decision-making In our ablation study in Section 3.4, we found a foreground weight of $w_{\text{fg}} = 3$ to work best for the given MLV datasets. This essentially creates a bias toward foreground labels, which seems to mimic human consensus decision-making to a certain degree in our case. However, it is unclear how this observation generalizes to other data, structures, and rater groups. Albeit an analysis of this relation is out of scope for this paper, it could be an interesting starting point for follow-up research to investigate the observed foreground annotation bias and its implications.

5. Conclusion

In summary, we presented the first automatic method for MLV segmentation from 3D FLAIR imaging. Our model, MLV^2-Net, outperformed state-of-the-art baselines by embracing the styles of all annotators involved in the creation of the training set. In contrast to most segmentation methods, MLV^2-Net provides a rater-based uncertainty estimation. Together with the derived theoretical bounds on the segmented volume, we expect MLV^2-Net to be a valuable tool for clinical researchers that study the glymphatic system. Yet, the technical contributions and code are generic and could be beneficial for other applications as well.

Acknowledgments

References

Mehmet Sait Albayram, Garrett Smith, Fatih Tufan, Ibrahim Sacit Tuna, Mehmet Bostancıklıoğlu, Michael Zile, and Onder Albayram. Non-invasive mr imaging of human brain lymphatic networks with connections to cervical lymph nodes. *Nature communications*, 13(1):203, 2022.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar,

Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), July 2022.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. *International Conference on Computer Vision*, 2023.

Alberto Del Guerra, Salleh Ahmad, Mihai Avram, Nicola Belcari, Arne Berneking, Laura Biagi, Maria Giuseppina Bisogni, Felix Brandl, Jorge Cabello, Niccolò Camarlinghi, Piergiorgio Cerello, Chang-Hoon Choi, Silvia Coli, Sabrina Colpo, Julien Fleury, Vito Gagliardi, Giuseppe Giraudo, Karsten Heekeren, Wolfram Kawohl, Theodora Kostou, Jean-Luc Lefaucheur, Christoph Lerche, George Loudos, Matteo Morrocchi, Julien Muller, Mona Mustafa, Irene Neuner, Panagiotis Papadimitroulas, Francesco Pennazio, Ravichandran Rajkumar, Cláudia Régio Brambilla, Julien Rivoire, Elena Rota Kops, Jürgen Scheins, Rémy Schimpf, N. Jon Shah, Christian Sorg, Giancarlo Sportelli, Michela Tosetti, Riccardo Trinchero, Christine Wyss, and Sibylle Ziegler. Trimage: A dedicated trimodality (pet/mr/eeg) imaging tool for schizophrenia. *European Psychiatry*, 50:7–20, 2018.

Xue-Bing Ding, Xin-Xin Wang, Dan-Hao Xia, Han Liu, Hai-Yan Tian, Yu Fu, Yong-Kang Chen, Chi Qin, Jiu-Qi Wang, Zhi Xiang, Zhong-Xian Zhang, Qin-Chen Cao, Wei Wang, Jia-Yi Li, Erxi Wu, Bei-Sha Tang, Ming-Ming Ma, Jun-Fang Teng, and Xue-Jing Wang. Impaired meningeal lymphatic drainage in patients with idiopathic parkinson's disease. *Nature Medicine*, 27(3):411–418, January 2021.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382, November 1971.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

James R. Goodman, Zachariah O. Adham, Randall L. Woltjer, Amanda W. Lund, and Jeffrey J. Iliff. Characterization of dural sinus-associated lymphatic vasculature in human alzheimer's dementia subjects. *Brain, Behavior, and Immunity*, 73:34–40, October 2018.

Xiayu Guo, Xian Lin, Xin Yang, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. Uctnet: Uncertainty-guided cnn-transformer hybrid networks for medical image segmentation. *Pattern Recognition*, 152: 110491, August 2024.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

Jeffrey J. Iliff, Minghuan Wang, Yonghong Liao, Benjamin A. Plogg, Weiguo Peng, Georg A. Gundersen, Helene Benveniste, G. Edward Vates, Rashid Deane, Steven A. Goldman, Erlend A. Nagelhus, and Maiken Nedergaard. A paravascular pathway facilitates csf flow through the brain parenchyma and the clearance of interstitial solutes, including amyloid β. *Science Translational Medicine*, 4(147), August 2012.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2 2021.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jäger. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15009. Springer Nature Switzerland, October 2024.

Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous

images. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NeurIPS'18, page 6965–6975, Red Hook, NY, USA, 2018. Curran Associates Inc.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, March 1977.

Antoine Louveau, Igor Smirnov, Timothy J. Keyes, Jacob D. Eccles, Sherin J. Rouhani, J. David Peske, Noel C. Derecki, David Castle, James W. Mandell, Kevin S. Lee, Tajie H. Harris, and Jonathan Kipnis. Structural and functional features of central nervous system lymphatic vessels. *Nature*, 523 (7560):337–341, June 2015.

Antoine Louveau, Jasmin Herz, Maria Nordheim Alme, Andrea Francesca Salvador, Michael Q. Dong, Kenneth E. Viar, S. Grace Herod, James Knopp, Joshua C. Setliff, Alexander L. Lupi, Sandro Da Mesquita, Elizabeth L. Frost, Alban Gaultier, Tajie H. Harris, Rui Cao, Song Hu, John R. Lukens, Igor Smirnov, Christopher C. Overall, Guillermo Oliver, and Jonathan Kipnis. Cns lymphatic drainage and neuroinflammation are regulated by meningeal lymphatic vasculature. *Nature Neuroscience*, 21(10):1380–1391, September 2018.

Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. D-lema: Deep learning ensembles from multiple annotations - application to skin lesion segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1837–1846, 2021.

Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Comput Methods Programs Biomed*, 98(3):278–284, October 2009.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pages 234–241. Springer International Publishing, 2015.

S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, July 2004.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), September 2023.

Le Zhang, Ryutaro Tanno, Moucheng Xu, Yawen Huang, Kevin Bronik, Chen Jin, Joseph Jacob, Yefeng Zheng, Ling Shao, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. Learning from multiple annotators for medical image segmentation. *Pattern Recognition*, 138:109400, 2023.