End-to-end Volumetric Segmentation of White Matter Hyperintensities using Deep Learning

Farkhani, Sadaf; Demnitz, Naiara; Boraxbekk, Carl-Johan; Lundell, Henrik; Siebner, Hartwig Roman; Petersen, Esben Thade; Madsen, Kristoffer Hougaard

Published in:
Computer Methods and Programs in Biomedicine

Link to article, DOI:
10.1016/j.cmpb.2024.108008

Publication date:
2024

Document Version
Peer reviewed version

End-to-end Volumetric Segmentation of White Matter Hyperintensities using Deep Learning

Sadaf Farkhani , Naiara Demnitz , Carl-Johan Boraxbekk , Henrik Lundell , Hartwig Roman Siebner , Esben Thade Petersen , Kristoffer Hougaard Madsen

PII: S0169-2607(24)00003-8
DOI: https://doi.org/10.1016/j.cmpb.2024.108008
Reference: COMM 108008

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 12 July 2023
Revised date: 8 December 2023
Accepted date: 3 January 2024

Please cite this article as: Sadaf Farkhani , Naiara Demnitz , Carl-Johan Boraxbekk , Henrik Lundell , Hartwig Roman Siebner , Esben Thade Petersen , Kristoffer Hougaard Madsen , End-to-end Volumetric Segmentation of White Matter Hyperintensities using Deep Learning, *Computer Methods and Programs in Biomedicine* (2024), doi: https://doi.org/10.1016/j.cmpb.2024.108008

Highlights

- Segmentation of MRI hyperintensities using an end-to-end trained 3D transformer model
- Structural information from T1-weighted MRI enable better prediction
- Performance is competitive even for healthy aging with relatively low lesions loads
- The performance of the method is comparable to inter-rater agreement

End-to-end Volumetric Segmentation of White Matter Hyperintensities using Deep Learning

Sadaf Farkhani[1], Naiara Demnitz[1], Carl-Johan Boraxbekk[1,2,3,4], Henrik Lundell[1,5], Hartwig Roman Siebner[1,2,3], Esben Thade Petersen[1,5], Kristoffer Hougaard Madsen[1,6]

[1]Danish Research Center for Magnetic Resonance, Center for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital—Amager and Hvidovre, Hvidovre, Denmark
[2]Institute for Clinical Medicine, Faculty of Medical and Health Sciences, University of Copenhagen, Denmark
[3]Department of Neurology, Copenhagen University Hospital Bispebjerg and Frederiksberg, Copenhagen, Denmark
[4]Institute of Sports Medicine Copenhagen (ISMC), Copenhagen University Hospital Bispebjerg and Frederiksberg, Copenhagen, Denmark
[5]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
[6]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

Corresponding Author: sadaff@drcmr.dk (Sadaf Farkhani), Kattegaard Alle 30, Hvidovre, Denmark.

Abstract

Background and Objectives

Reliable detection of white matter hyperintensities (WMH) is crucial for studying the impact of diffuse white-matter pathology on brain health and monitoring changes in WMH load over time. However, manual annotation of 3D high-dimensional neuroimages is laborious and can be prone to biases and errors in the annotation procedure. In this study, we evaluate the performance of deep learning (DL) segmentation tools and propose a novel volumetric segmentation model incorporating self-attention via a transformer-based architecture. Ultimately, we aim to evaluate diverse factors that influence WMH segmentation, aiming for a comprehensive analysis of the state-of-the-art algorithms in a broader context.

Methods

We trained state-of-the-art DL algorithms, and incorporated advanced attention mechanisms, using structural fluid-attenuated inversion recovery (FLAIR) image acquisitions. The anatomical MRI data utilized for model training was obtained from healthy individuals aged 62-70 years in the LIve active Successful Aging (LISA) project. Given the potential sparsity of lesion volume among healthy aging individuals, we explored the impact of incorporating a weighted loss function and ensemble models. To assess the generalizability of the studied DL models, we applied the trained algorithm to an independent subset of data sourced from the MICCAI WMH challenge (MWSC). Notably, this subset had vastly different acquisition parameters compared to the LISA dataset used for training.

Results

Consistently, DL approaches exhibited commendable segmentation performance, achieving the level of inter-rater agreement comparable to expert performance, ensuring superior quality

segmentation outcomes. On the out of sample dataset, the ensemble models exhibited the most outstanding performance.

Conclusions

DL methods generally surpassed conventional approaches in our study. While all DL methods performed comparably, incorporating attention mechanisms could prove advantageous in future applications with a wider availability of training data. As expected, our experiments indicate that the use of ensemble-based models enables the superior generalization in out-of-distribution settings. We believe that introducing DL methods in the WHM annotation workflow in heathy aging cohorts is promising, not only for reducing the annotation time required, but also for eventually improving accuracy and robustness via incorporating the automatic segmentations in the evaluation procedure.

*Keywords***:** Transformer; Segmentation; Attention mechanism; Deep learning; White matter hyperintensities

1. Introduction

White matter hyperintensities (WMH) are lesions visualized on magnetic resonance imaging (MRI) as abnormally high signal intensities in fluid-attenuated inversion recovery (FLAIR) sequences [1]. WMHs are typically located around periventricular (PWMH) and deep subcortical (DWMH) areas. WMHs are frequently observed in healthy older adults as well as patients with a high risk of cardiovascular disease [2]. A high lesion load is associated with an increased risk of cognitive impairment, stroke, and death [3]. To mitigate the high risk of cognitive impairments caused by WMHs, accurate detection and estimation of their location, shape, and volume have high importance through diagnosis and treatment monitoring [4].

Advances made in MRI, including improving dimensionality, contribute to an increased reliability in detecting and segmenting WMHs. While radiographers have a near-to-optimal knowledge of identifying WMH, manual labeling of WMHs in high-resolution images is a tedious and time-consuming task that can lead to errors due to fatigue and declining attention [5]. Hence, the utilization of autonomous segmentation tools is of great importance for supporting radiographers in managing large volumes of data, particularly in longitudinal studies, very large datasets, and personalized treatment tracking where the number of scans can escalate considerably [6].

Among the available tools for WMHs segmentation, the lesion segmentation tool (LST) [7] included in SPM and the brain intensity abnormality classification algorithm (BIANCA) [8] included in the FMRIB Software Library (FSL) have been widely used in clinical research [9–11]. Both tools mainly use the abnormal intensity features to create a probability map and apply either a global or adaptive region-based threshold to generate the segmentation map.

Deep learning (DL) algorithms show promising results in extracting informative patterns and features within images, including the segmentation of WMHs which has been attempted for example using the TrUENet architecture [12]. TrUENet is a modular DL system based on an ensemble of three 2D models trained on three orthogonal projections (sagittal, axial, and coronal). Although a 3D model increases the computational complexity and introduces additional parameters, it can be advantageous due to the enhanced capability of representing volumetric texture and context information specific to lesions. This can be of particular advantage when the data available for training is plentiful [13]. However, as TrUENet is a modular system, postprocessing in the form of combining probability maps and thresholding is needed in this architecture. While 2D models are likely better at capturing long-range spatial dependencies in

the data [14], the parameters related to the postprocessing of the images can be error-prone and may fail to generalize to data from other sites or cohorts.

Recently, transformer architectures adopted from natural language processing have shown promising improvements in grasping long-range dependencies, particularly within 3D modeling [15]. Generally, transformer-based models perform better than conventional DL methods, such as convolutional neural networks (CNN), since the transformer is able to use multiple layers of self-attention blocks to enable modeling of global long-range dependencies within a series of patches extracted from the image [16].

Recently, other work focused on a 2D transformer encoder-decoder architecture and compares the model to a fully CNN-based architecture for WMH segmentation [17]. However, the 2D transformer architecture, in comparison to CNN, demonstrated inferior performance. Given the data-intensive nature of transformer-based models, the utilization of 2D slices as input in [17] may be driven by the need to ensure sufficient training data for these architectures. Moreover, transformers are mainly effective at capturing complex dependencies which makes them more attractive for encoders than decoders [16]. By utilizing a 3D transformer-based model for encoding and a sequence of convolutional (conv) blocks for decoding, the proposed model aims to capture long-range dependencies in volumetric data by combining the strengths of both CNNs and transformers [18].

Besides the architecture, several data characteristics play an important role in training a transformer model. For instance, the amount of data used to train the model in [17] appears inadequate (n = 60) for training transformer-based models, especially considering the sparse distribution of WMHs. In addition, a crucial aspect overlooked in previous WMH segmentation techniques is the utilization of datasets containing a high volume of lesion load. In datasets used

previously, subjects had remarkably high WMHs volume which leads to clearly visible high-contrast lesions [12,17]. However, it is important to detect WMHs even in early stages before the onset of a particular disease in which the contrast of WMHs is lower such as when investigating healthy aging cohorts, and particularly with respect to DWMH [19].

The objective of this study was to assess the state-of-the-art algorithms for the detection of WHM, specifically focusing on early detection where the lesions are notably sparse. In pursuit of this goal, we introduced an attention-based DL model, VoSHT (Volumetric Segmentation of WMH using Transformer), an end-to-end volumetric segmentation tool with self-attention via a transformer model. Our evaluation involved training and comparing state-of-the-art models, as well as the proposed attention-based model, using a dataset that incorporates subjects from a non-preselected healthy aging cohort, characterized by notably lower lesion ratios, which will typically pose a greater challenge for prediction. We assessed various models with respect to four criteria: conventional WMH segmentation tools, volumetric DL models with plain loss function, volumetric DL models with weighted loss function, and ensemble DL models. These evaluations were prompted by the challenge of comprehending the image complexity solely based on sparse occurrence of WMHs. Furthermore, we inferenced the generalizability of DL models on out-of-distribution test datasets, where the acquisition parameters of the cohort and WMH's volume differ substantially from the training dataset.

2. Material and Methods

2.1. Datasets

Two datasets with different acquisitions (scanner and MRI protocols), different lesion characteristics, and different annotators were considered in our investigations.

2.1.1. LIve active Successful Aging (LISA)

The LISA study received ethical approval from the Ethical Committees of the Capital Region of Denmark (No. H-3-2014-017) and the Danish Data Protection Agency. It complies with the declaration of Helsinki and was registered on clinicaltrials.gov (NCT02123641). LISA is a cohort of 451 older adults, initiated in 2014-2015. Figures A.4 and A.5 in the Appendix show the distribution of lesion volume among LISA training and test sets. Among all the subjects from the LISA cohort, 300 participants qualified for MRI, see [20] for details. FLAIR sequences were primarily used for the annotation of WMHs. All FLAIR scans were reoriented to MNI space before being manually annotated. An expert radiographer manually labeled the lesions over all 300 subjects. The annotator was instructed to disregard lesions that had fewer than three neighboring lesions. T1-weighted and FLAIR sequences were used in this study. The characteristics of the LISA dataset are illustrated in Table 1. Furthermore, a random subset of the LISA dataset (n = 7), was separately annotated by another expert radiographer to examine how well labels generalize across raters and to evaluate how the trained model generalizes to the other unseen annotations.

Table 1. Characteristic explanation of the three datasets used in this paper. Q1 and Q3 are the first and third quartiles, respectively.

Dataset	Scanner	Subjects	Age (y) avg ± std	Sex (F:M)	Voxel Size (mm3)		WMH (mm3) [Q1-Q3]
					FLAIR	T1-weighted	
LISA	3T Philips Achieva	300	66.46 ± 2.52	273:178	1.00×1.00×1.00	0.85×0.85×0.8	694 – 4,514
MWSC	3T Philips Ingenuity	10	-	-	1.04×1.04×0.5	0.87×0.87×1.0	3,547 – 19,130
	1.5T GE Signa	10	-	-	1.21×1.21×1.30	0.98×0.98×1.5	942 – 6,710

2.1.2. MICCAI WMH segmentation challenge (MWSC)

MWSC is a dataset consisting of 60 subjects for training and 110 subjects for the test set from three different sites: UMC Utrecht, NUHS Singapore, and VU Amsterdam. In our experiments, we selected the benchmark test set from the VU Amsterdam center that has previously been used as an unseen test set in other formal evaluations of model performance [11,16]. The MWSC test set includes 20 scans from two different scanners (Table 1).

2.2. Image Preprocessing

All images from LISA and MWSC datasets were skull-stripped and bias-corrected using BET [20], a tool from the FSL software package. All images were registered to the same space as the first session space using the FLAIR sequence as the reference using FSL FLIRT (utilizing a 6-parameters linear rigid body transformation) [21]. Subsequently, the intensity of all images in both datasets was normalized within the intervals of [0, 1]. In addition to the above-mentioned steps, a resampling step was considered just for preprocessing of MWSC test sets to ensure the same voxel resolution as the LISA dataset.

2.3. Self-Attention Deep Learning Model

Figure 1 illustrates the general framework of the proposed system, VoSHT. VoSHT's architecture is built upon the self-attention mechanism, which is derived from UNETR, a model introduced in [17]. UNETR is inspired by the structure of the UNet architecture [22]. Instead of using conv blocks for the encoder path, consecutive self-attention encoders are used which can capture spatial dependencies in the data although they are situated at considerable distances.

First, data augmentation as a key factor in training self-attention blocks is implemented on the 3D FLAIR images. Here, data augmentation plays a particularly important role in lesion segmentation tasks, where the data is limited and classes (lesion and non-lesion) are highly imbalanced. A linear projection is applied to convert 3D non-overlapping patches of augmented

data into 1D embedding represented as a sequence of tokens. To encode the location of the patch, a position embedding of the patch is added to the token. The positional embedding is computed based on

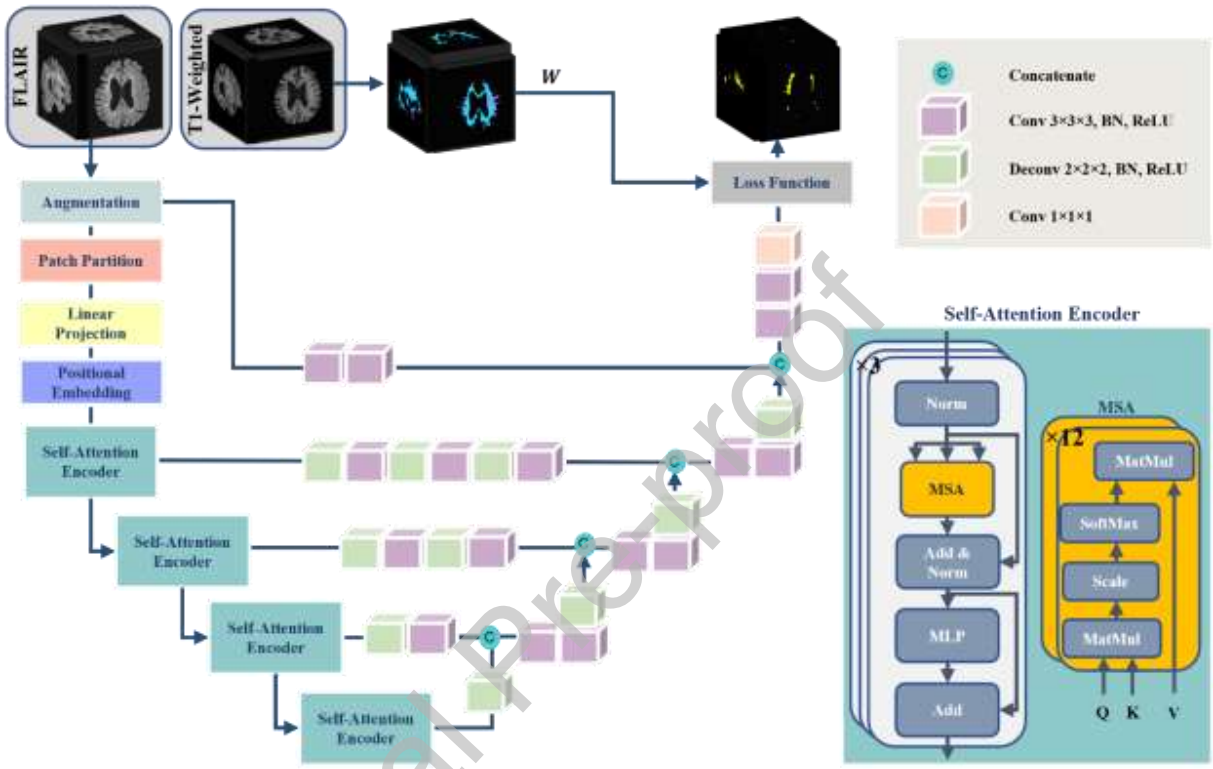

Fig. 1. Schematic depiction of the VoSHT model. The inner structure of the self-attention encoder is depicted on the bottom right. Each self-attention encoder is composed of three attention block layers, where each attention block uses 12 layers of multi-head self-attention (MSA). In MSA, MatMul denotes a multiplication operation to find similarities between query (Q), key (K), and value (V). Selected features from the T1-weighted sequence are used as inputs for the loss function.

$$z_0 = [x_v^1 E;\ x_v^2 E; \ldots; x_v^N E] + E_{pos}, \quad (1)$$

where $E_{pos} \epsilon R^{N \times K}$ is a 1D learnable positional embedding with the hidden dimension of k that is added to embedded data (tokens) to have a track of the position. N is the total number of tokens

and $\{x_v^i;\ i = 1,2,\ldots,N\}$ is the flattened vector for the i^{th} patch. z_0 is the vector including all the embedded patches.

Afterward, the embedded patches are passed through the self-attention encoders. In each stage, 3 layers of self-attention blocks are included which each has 12 parallel layers of multi-head self-attention (MSA) followed by a multi-layer perceptron (MLP) block. MSA layers are initialized randomly to ensure diversity in learning representation subspaces. The MSA is computed by

$$MSA(z_l) = SoftMax(\frac{QK^T}{\sqrt{d_h}})V, \quad (2)$$

where Q, K, and V are query, key, and value matrices trained based on the sequence of tokens, respectively. z_l is the sequence of tokens from the l^{th} layer. The attention weight is calculated based on the similarity between the sequence and their respective Q and K representations [21]. The calculated similarity is divided by a scaling factor d_h to keep the computation and the number of parameters constant [16]. The output of the self-attention encoder is obtained as

$$z_l^{'} = z_{l-1} + MSA(Norm(z_{l-1})), l = 1,2,\ldots,L \quad (3)$$

$$z_l = z_l^{'} + MLP(Norm(z_l^{'})), l = 1,2,\ldots,L \quad (4)$$

where L is the number of parallel MSA layers considered as 12 depicted in Figure 1, and Norm is a normalization operation applied before every block in the self-attention encoder. After each MSA block, there is an MLP block with two linear layers and a Gaussian error linear unit (GELU) function at the end. The Add block in the self-attention encoder is intended to preserve the features extracted from the previous layer. The outputs of all the self-attention layers have the same dimensionality as the input embedding size.

To upsample and preserve the extracted features, a deconvolution (deconv) block is considered at the bottleneck. Also, conv and deconv layers are consecutively considered along the decoder

path in a similar way as the UNet decoder path [22]. There is a batch normalization (BN) layer and rectified linear unit (ReLU) followed by conv and deconv blocks in each layer; except for the final layer which is a 1×1×1 conv layer used to generate the final semantic segmentation map. In the skip connections, the output from self-attention encoders is upsampled and resized using consecutive deconv and conv layers.

2.4. Loss Function

Two types of loss functions were used in our investigation which both use a combination of Cross-Entropy (CE) [23] and Dice (DSC) loss functions. Both loss functions can be expressed as

$$L(G,Y) = DSC + CE \tag{5}$$

$$L(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2+\sum_{j=1}^{J} Y_{i,j}^2} - L_{CE} \tag{6}$$

$$L_{CE}(G,Y) = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}W_i\log(Y_{i,j}), \tag{7}$$

where I is the number of voxels and J is the number of classes (which is two here: the background and lesion). $Y_{i,j}$ and $G_{i,j}$ are the probability output and one-hot encoded ground truth for class j at voxel i, respectively. In the first loss function, W_i is set to one.

In the second loss function we consider, however, that the CE loss is weighted using relevant chosen structural features W_i calculated from T1-weighted images. Since WMHs are mainly located in the sub-cortical WM and close to the ventricles, the voxel distances from the ventricles and GM were calculated [12] based on the T1-weighted sequence. Consequently, the model will penalize neglected DWMH and PWMH strongly. Moreover, W_i compensates for the sparsity of labels by enhancing the model's comprehension of structural information. This method of introducing additional information avoids the large increase in the number of parameters incurred by directly incorporating T1-weighted volumes.

2.5. Implementation Details

All DL models were trained using PyTorch[1] library. Models were trained on a 24GB NVIDIA (GeForce RTX 3090) GPU. Except for nnUNet where the implementation relied directly on PyTorch, other DL models were implemented using the MONAI[2] framework. UNET- and transformer-based models were trained with a batch size of 2 using the AdamW optimizer [24] with an initial learning rate of 0.0001 for 1000 epochs. All the transformer-based models were trained using the setups provided in the UNETR article [17]. TrUENet was trained on a batch size of 8 using Adam optimizer with an initial learning rate of 0.0001 for 60 epochs. For nnUNet, the original setup reported in the original publication [27] was utilized for training.

All these methods were trained from scratch. LISA dataset was divided into three sets of training, test, and validation with a ratio of 0.8:0.1:0.1, respectively. The data were stratified based on the total lesion volume. Thus, 30 subjects from the LISA dataset were set aside as a test set. Notably, the nnUNet used a slightly different training procedure as the provided code [27] required utilizing a 5-fold cross-validation procedure during training, hence this method was able to learn the ensemble weighting from the validation set.

2.6. Performance Evaluation Metrics

To evaluate and compare the performance of different methods, nine metrics were considered: area under the curve for Precision-Recall (AUC-PR), Dice similarity coefficient (DSC), Hausdorff distance (HD), Recall, Kappa, cluster-wise false positive rate (FPR), cluster-wise false negative rate (FNR), absolute volume difference (AVD), cluster-wise F1-score, and cluster-wise true positive rate (TPR). All the above-mentioned metrics measure the agreement between the

[1] http://pytorch.org/

[2] https://monai.io/

manual segmentation (G) and the prediction (P). For the cluster-wise metrics, which include cluster-wise TPR, F1-score, FPR, and FNR, the comparison is done at the level of entire lesion clusters rather than at the voxel level. In contrast to the remaining metrics HD is a measure of the distance between the spatial maps of G and P.

Clusters were defined via a 3D neighborhood of the 26 closest voxels in our calculations. Hence, the cluster-wise TPR is the number of true positive lesion clusters predicted divided by the total number of true positive lesion clusters. For the cluster-wise F1-score, the definition is

$$\mathrm{F1} = \frac{2\times(\mathrm{TPR_C}\times\mathrm{Precision_C})}{(\mathrm{TPR_C}+\mathrm{Precision_C})}, \tag{8}$$

where C presents the cluster-wise measurement, and cluster-wise precision is calculated by dividing the number of true positive lesion clusters by the total number of detected lesion clusters. In the group-level evaluations, two metrics of FPR and FNR were used to evaluate the model performance across subjects. Here, FPR is the ratio of the number of false positives divided by false positives plus true positives, whereas FNR is the ratio of false negatives divided by false negatives plus true negatives. FPR and FNR are also calculated at cluster-level.

AUC-PR is a metric for measuring model performance with respect to *precision* (the fraction of positive lesions of P out of manually positive lesions in G) against *recall* (the fraction of positive lesions of P out of manually positive lesions and background in G) evaluated for thresholds applied to the predicted map. The higher the area under the PR curve the better the model performs. DSC is a metric for measuring the overlap between G and a binary representation of P. The R2-Score is utilized to compare the total lesion volumes between G and P.

In the HD metric, the longest distance between subsets of lesions in G and P is calculated. To suppress the outlier effects, the 95th percentile of distance point sets is considered in these calculations.

For estimating the inter-rater reliability, Cohen's Kappa [25] score is used. The definition of Kappa score is

$$k = \frac{P_0 - P_e}{1 - P_e}, \quad (9)$$

where P_e and P_0 are the probability of annotator's agreement and the probability of random agreement, respectively.

3. Results

In this section, we first describe and compare the results gained from different models on LISA as the in-distribution test set, then we evaluate the performance on MWSC as the out-of-distribution test set for each model. Here, LST [7], BIANCA [8], 3D UNET [26], UNETR [18], TrUENet [12], and nnUnet [27] were used for performance comparisons. Except for the LST, which is an unsupervised method, the other models were trained from scratch on the LISA training set. BIANCA was additionally post-processed using a locally adaptive thresholding approach (LOCATE) [8]. To study the impact of the weighted loss function on UNET architecture, we also developed and trained a 3D UNET architecture utilizing the weighted loss function, which is shown as UNET*.

Table 2 presents the comparison between the different methods studied. Tables are divided into four categories: 1) conventional methods, 2) CNN models, 3) ensemble DL models, and 4) DL models utilizing T1-weighted features in the loss function.

3.1. In-Distribution Evaluation

To assess the consistency of DL models across various annotators, we compared nnUNet and VoSHT with the inter-rater agreement in Table 3. Interestingly, both DL models showed superior

performance compared to manual raters. We attribute this observation to DL models providing more consistent labeling, not affected by attention, fatigue, or total lesion load of the subject.

Table 2. Comparing existing methods with the UNETR model regarding the training circumstances. Rows represent different methods, arranged from top to bottom into 3 categories: intensity-based methods, DL methods trained by an unweighted loss function, and DL methods trained using a weighted loss function. Models specified with * were trained using the weighted loss function. Abbreviations: unsupervised learning (UL), supervised learning (SL), convolutional neural network (CNN), not applicable (N/A).

	Method	Type	Dimension	Image Modalities	Parameters (M)	Training Time (h)
1	**LST**	UL	1D, 3D	T1, FLAIR	N/A	-
	BIANCA	SL	1D	T1, FLAIR	N/A	<2
2	**UNET**	CNN	3D	FLAIR	48	13.3
	UNETR	CNN, transformer	3D	FLAIR	92.58	19.4
3	**TrUENet***	CNN	2D, Ensemble	T1, FLAIR	77.4	57.8
	nnUNet	CNN	2D, 3D, Ensemble	FLAIR	≈73	39.4
4	**UNET***	CNN	3D	T1, FLAIR	48	13.3
	VoSHT*	CNN, transformer	3D	T1, FLAIR	92.58	19.4

Table 3. Inter-rater comparison of the model's performance on a subset of the LISA dataset. The second annotation is considered the benchmark.

Method	DSC	HD	AUC-PR	Kappa
Inter-rater	82.22	7.24	71.62	63.33 ± 0.121
nnUNet	90.23	4.80	87.40	87.24 ± 0.062
VoSHT	85.76	13.04	71.85	71.36 ± 0.085

Table 4. Using the LISA test dataset as a benchmark in terms of average DSC, cluster-wise TPR, cluster-wise F1-score, FPR, HD, Recall, AUC-PR, and AVD. The background is included in these evaluations. While TPR and F1 are calculated at the entire lesion level, the other metrics are calculated at the voxel level. Rows represent different methods, arranged from top to bottom into 4 categories: conventional methods, standard CNN methods, ensemble CNN methods, and CNN methods trained using a weighted loss function. Models specified with * were trained using the weighted loss function.

Method	DSC (%)	TPR-cluster (%)	F1-cluster (%)	FPR (%)	HD (mm3)	Recall (%)	AUC-PR (%)	AVD
LST	71.71	54.78	37.00	0.0323	22.84	76.61	46.92	1.65
BIANCA	74.62	83.96	35.17	0.0340	20.53	87.08	57.41	2.06
UNET	84.25	**92.12**	57.09	0.0265	12.59	86.62	70.19	0.84
UNETR	81.88	83.82	63.59	0.0091	15.61	80.83	65.76	0.69
TrUENet*	80.94	80.18	70.01	**0.0070**	13.79	76.63	66.50	**0.62**
nnUnet	85.50	76.43	**78.23**	0.0092	11.61	84.32	**73.09**	0.66
UNET*	85.43	91.95	72.97	0.0132	11.12	87.48	72.41	0.63
VoSHT*	**85.59**	91.63	73.43	0.0135	**10.30**	**87.48**	72.67	**0.62**

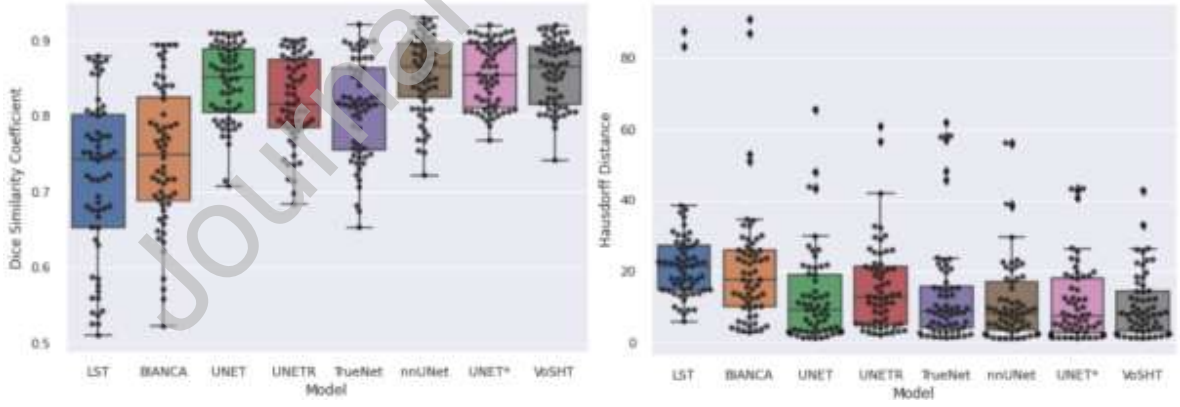

Figure 2. Comparison of DSC and HD scores among different methods on the LISA test set. The black vertical line in the boxplot shows the median. Each black point represents the relevant subject. The higher the DSC, the better the model performs. The lower the Hausdorff distance, the better the model performs.

As it is shown in Table 4 and Figure 2, DL models achieve considerably better performance compared to conventional non-DL methods, LST had the lowest performance. Furthermore, the weighted loss function slightly improves the performance of UNET-based models. For the quantitative comparison, all models exhibited comparable performance. Among models categorized with the weighted loss function, TrUENet had a 4.6% lower DSC score on average than the comparable 3D-shaped models. Interestingly, nnUNet achieved the best balance between precision and recall (AUC-PR) among all models, which is also represented in cluster-wise F1 score.

3.2. Out-of-Distribution Evaluation

To provide an assessment of the models' performance for the out-of-distribution test set, we tested all DL models, including our model, on the MWSC test set which is commonly used as a benchmark in the segmentation of WMHs. In Table 5, the results show that nnUNet outperformed other models in all metrics, except FPR and HD which UNETR and UNET* gained superior performance, respectively. Moreover, Figure 3 demonstrates that nnUNet had least variability among all the models in out-of-distribution test set.

4. Discussion

The main aim of this study was to train, evaluate, and compare various DL methods, including our proposed method, VoSHT, in healthy older participants with relatively sparse WMHs. We implemented an end-to-end trained volumetric transformer model and evaluated the performance with other state-of-the-art models. To the best of our knowledge, this is the first instance of utilizing a transformer-based model trained on a large dataset which is a prerequisite for transformer models. While this model performed relatively well, the ensemble-based nnUNet still had better performance, particularly on the external validation dataset.

Table 5. Using the MWSC set as a benchmark in terms of average DSC, cluster-wise TPR, cluster-wise F1-score, FPR, HD, Recall, AUC-PR, and AVD. The background is included in these evaluations. The background is also included as the second class in the calculations. Rows are categorized based on either the loss function, method, or both. Models specified with * were trained using the weighted loss function.

Metric	DSC (%)	TPR-cluster (%)	F1-cluster (%)	FPR (%)	HD (mm3)	Recall (%)	AUC-PR (%)	AVD
UNET	82.12	31.49	44.41	0.0531	16.06	78.98	67.26	0.60
UNETR	78.81	27.24	34.43	**0.0084**	13.71	74.49	62.37	0.67
TrUENet*	83.45	39.05	44.41	0.0099	15.83	78.89	64.72	0.51
nnUNet	**86.16**	**46.45**	**62.28**	0.0155	10.69	**83.82**	**74.77**	**0.47**
UNET*	85.53	39.71	54.14	0.0143	**9.16**	83.74	72.38	0.53
VoSHT*	84.13	39.97	52.79	0.0127	14.02	81.49	70.20	0.56

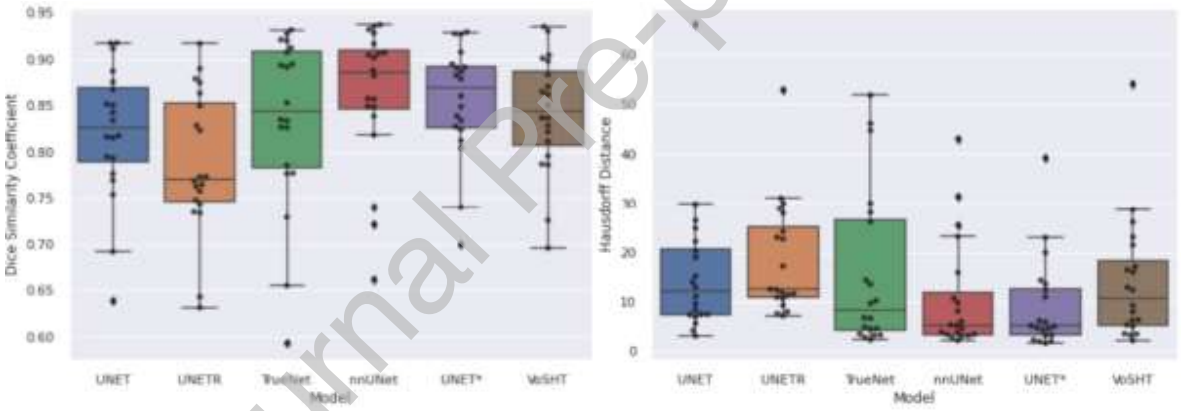

Figure 3. Comparison of DSC and HD among all supervised methods tested on the MWSC set. The black vertical line in the boxplot represents the median. Each black point is associated with a subject in the test set. The higher the DSC and the lower the HD value, the better the model performs.

In general, the performance of the DL-based models was comparable to the estimated inter-rater variability, which suggests that DL models can be used for supporting radiographers in the analysis of large high-resolution data. Additionally, we found that incorporating selected features into the loss function demonstrated an increase in the performance of certain architecture (comparing UNET and UNET* for instance), suggesting a substantial potential benefit of this

feature engineering approach in situations where lesions are sparse (Table 4). Additionally, we demonstrated that models which directly consider a volumetric representation of neuroimages achieved slightly higher performance compared to similar 2D-based ensemble models, such as TrUENet. Among all DL models, ensemble models demonstrated superior performance in generalization using an out-of-distribution test set.

Explicitly, we trained and compared state-of-the-art DL models and the most popular toolboxes used routinely for WMH segmentation: 3D UNETR [18], 3D UNET [22], TrUENet [12], nnUNet [27], BIANCA [8], and LST [7]. First, we evaluated human inter-rater performances in a subset of the LISA dataset to have a measurement of how closely our model performs with respect to different experts' interpretability. Next, we trained and tested our model using the LISA dataset as an in-distribution cohort including a wide range of lesion volumes and variations. Then, examples of the UNETR model's performance were compared in terms of the total WMH load. Last, models were tested and compared on the MWSC test set to assess performance with respect to generalizability in a completely independent dataset with different acquisition parameters.

We assessed the performance of nnUNet and our model concerning inter-rater agreements, as presented in Table 3. The results indicate that both models achieved performance superior to the human annotations, as evaluated by an independent human rater. Particularly, nnUNet gained a considerably high performance in all metrics, which may be an outcome of ensembling all folds of validation through the training procedure. Generally, the proposed system shows promising potential in accelerating the radiographer workflow when analyzing large datasets or longitudinal datasets.

When evaluating and comparing the performance of different models within the LISA test set, nnUNet, UNET*, and VoSHT achieved comparable performance (Table 4; Figure 2). Among models trained using the weighted loss function, VoSHT and UNET* outperformed TrUENet. This observation can be attributed to the importance of utilizing the volumetric analysis in WMH segmentation to capture and learn general and complex contexts of neuroimages. In addition to volumetric segmentation, UNET achieved the highest performance in terms of cluster-wise TPR. However, the considerably lower performance of UNET in terms of cluster-wise F1-score indicates that UNET has a general tendency to over-segment lesions. Particularly, DL models trained on imbalanced labeled data from scratch have only limited knowledge of the features that are important in neuroimaging [28]. While the weighted loss function significantly enhanced models' performance, it's noteworthy that nnUNet, which utilizes an ensemble model, showed similar performance even without the weighted loss function.

To emphasize the importance of which evaluation metrics are used, the correlation between the subject lesion load and their corresponding HD values using VoSHT is shown in Figure 4. Here, a negative correlation is observed between HD and the number of WMHs. Upon careful examination of the samples exhibiting a low lesion load of WMH (Figure 4(a)), it is evident that the achieved HD score (32.9mm) is relatively low for this certain case Although the visual performance shows that WMHs were segmented with high accuracy, the HD value is low since the total lesion volumes are extremely low. This can be attributed to the fact that the HD metric tends to be inflated in cases where there is little or no overlap between the prediction and the ground truth [29]. This finding holds true for the cluster-wise metrics, mainly due to the significantly imbalanced ratio of WMHs. Consequently, it is crucial to examine segmented maps for a more detailed comparison.

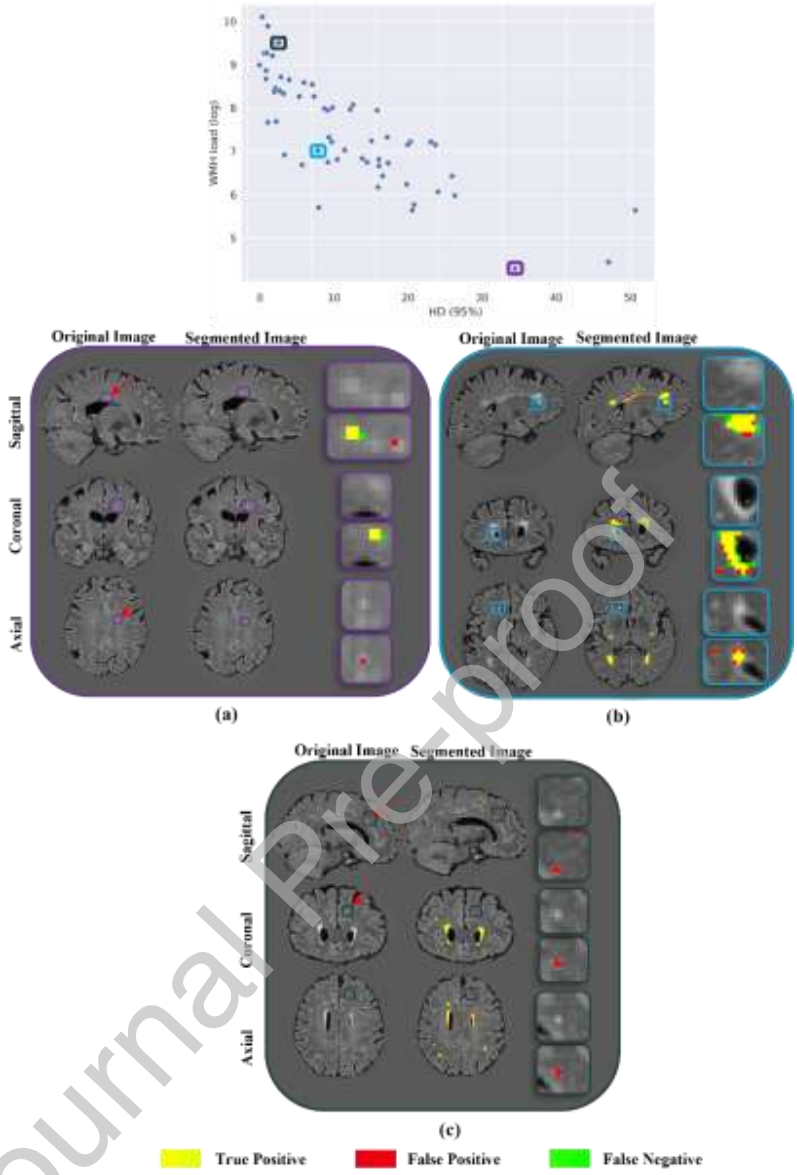

Figure 4. Three examples from the LISA test set comparing the manual segmentation with segmentation maps estimated via the VoSHT model. WMH volumes are 86, 8085, and 1366 for (a), (b), and (c) respectively. The overlap between the manually labeled map and the prediction map is color-coded by red, green, and yellow representing FP, FN, and TP, respectively. On the right side, the magnified cropped sections of original and segmented images are highlighted.

FP voxels indicated by red arrows in Figures 4(a) and 4(c) were missed by the radiographer, yet the model was able to segment them. Consequently, these dismissed lesions (FP voxels) have a

great negative impact on the evaluation metrics. To implicitly explore the performance of our model in relation to FP and FN voxels, two instances from the LISA test set with high FPR and FNR are shown in Figure A.2. Notably, VoSHT provides segmentation of the WMHs that are not annotated by the radiographer (also with reference to selection criteria stated in Section 2.1 and Figure A.2 panel a). This again emphasizes the importance of using a reliable automated segmentation tool to assist radiographers.

Providing a cohort-level comparison, Figure A.3 in the Appendix shows the accumulated FP and FN voxels across test subjects using BIANCA, nnUNet, and VoSHT. Interestingly, nnUNet demonstrated no missing lesion clusters (FN) among all test subjects. Although BIANCA has a sparse FN map across the LISA test set, the relatively dense FP map reveals that BIANCA has a tendency to over-segment lesions. In comparison, the FN map for VoSHT in Figure A.3 shows that the model occasionally misses small WMH clusters, especially near the ventricles.

In the evaluation of the MWSC test set, nnUNet achieved comparably higher performance than all the other DL models. In Figure 3, the highest performances were ranked for nnUNet, UNET*, and VoSHT, respectively. These observations underscore the significance of ensemble models, offering a beneficial balance between individual models, each with their distinct advantages. As indicated in Table 1, the MWSC has an anisotropic resolution within the FLAIR image. This characteristic potentially affects 3D-based methods like UNET* and VoSHT to a greater extent compared to 2D-based methods such as TrUENet and nnUNet. On the other hand, volumetric-based methods achieved a comparable performance to 2D-based methods in Table A.1 and Fig.

The implementation of the structurally penalized weighted loss function led to notable advancements in the performance of VoSHT compared to UNETR (Table 5). Similarly, there were significant improvements in the performance of UNET* compared to UNET. Therefore,

integrating T1-weighted data in this manner enhances the model's performance while maintaining a consistent number of parameters It also has the additional advantage that the information can readily be replaced by a template if T1-weighted data is unavailable which will then effectively act as prior information. In terms of using the attention-based mechanism, VoSHT demonstrated a wider variability than UNET-based models among test subjects in Figure 3. In practice, there is always a tradeoff between the capability of modeling specific aspects of the data and overfitting, and we expect the disadvantage of the more complex model to disappear with better data availability or transfer learning approaches [30].

In Figure 5, the HD score of the VoSHT model is depicted against the WMH volume on the MWSC test set to provide insight into the performance of attention-based models in an out-of-distribution setting. Examining the two instances with different lesion loads from Figure 5, VoSHT mainly showed FP disagreements with the manual labels near the borders of lesions. This effect can potentially be attributed to the subsampling deployed in the preprocessing of MWSC. Additionally, the FP voxels magnified in Figure 5(a) may also be due to the non-adaptive thresholding deployed after the MWSC resampling. Figure 5(b) highlights instances where VoSHT faced challenges in identifying WMHs, denoted by red arrows. The majority of FN voxels are observed in the borders of lesions or regions connecting larger lesions.

To also explore the correlation between DSC obtained from VoSHT and the total lesion load in the LISA test set, a scatter plot is shown in Figure 6, where the age of each subject is represented by the size of the data point. Aligned with the previous results, a positive correlation was observed between the DSC value and total lesion load. This emphasizes the importance of attending to multiple aspects of the data when attempting to interpret the results. Particularly, this is of importance for the interpretation of subjects with low lesion load where even the

misprediction of one voxel can have a great impact on the statistical results. In our sample of older adults (60 – 72 years), we did not observe any meaningful correlation between age and the model performance.

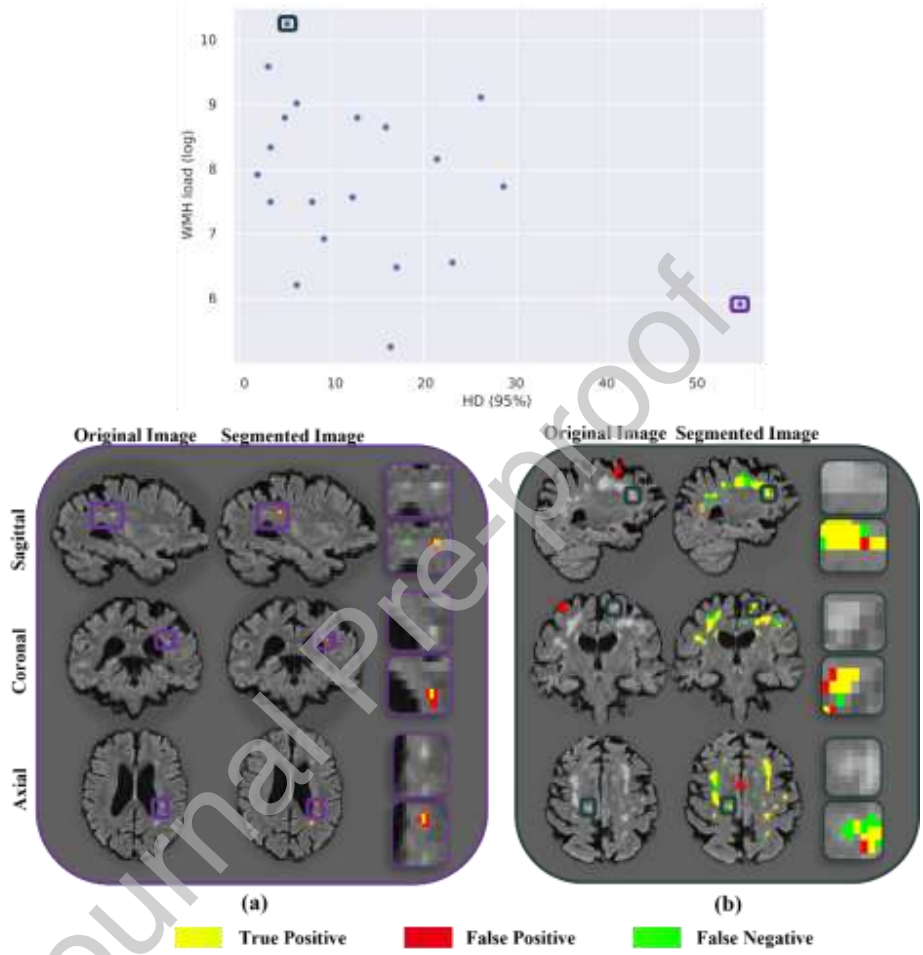

Figure 5. Examples of the MWSC set for comparing the manual segmentation with the VoSHT prediction map. The number of WMH voxels is 191 and 28284 for (a) and (b), respectively. FP, FN, and TP voxels are color coded with red, green, and yellow respectively. On the right side, the magnified cropped sections of original and segmented images are highlighted.

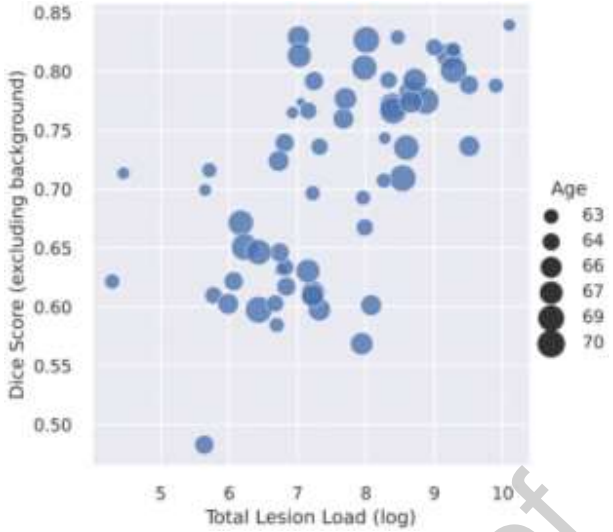

Figure 6. The correlation between total lesion voxels at the subject level and the logarithm of DSC values for VoSHT. The DSC score is only calculated for the lesion voxels and the background is excluded in this plot. The size of the points indicates the age.

To further investigate the effects of introducing structural information in the loss function, regression plots between the real and predicted lesion load for UNET and UNETR on the LISA test set are depicted in Figure 7. The introduction of structural information appears to have less impact on the transformer models (UNETR and VoSHT) than the models not utilizing a transformer structure (UNET and UNET*), we expect that this difference could be caused by the transformer models' ability to capture long-range dependencies in the data already without the used of additional information. While the UNETR trained with the non-weighted loss function has a tendency to under-segment lesions in general, the VoSHT method tends to over-segment. Here, we speculate that this behavior might actually be desired in some cases where the human annotation may be incomplete. In particular, in settings where the lesion load for a subject is high, it is perfectly reasonable for a human annotator to ignore very small lesions, as they will have little or no impact on the final classification or scores of the subject. However, from the

classification models perspective such labels may be very difficult to accurately predict as they have no special treatment in the cost function.

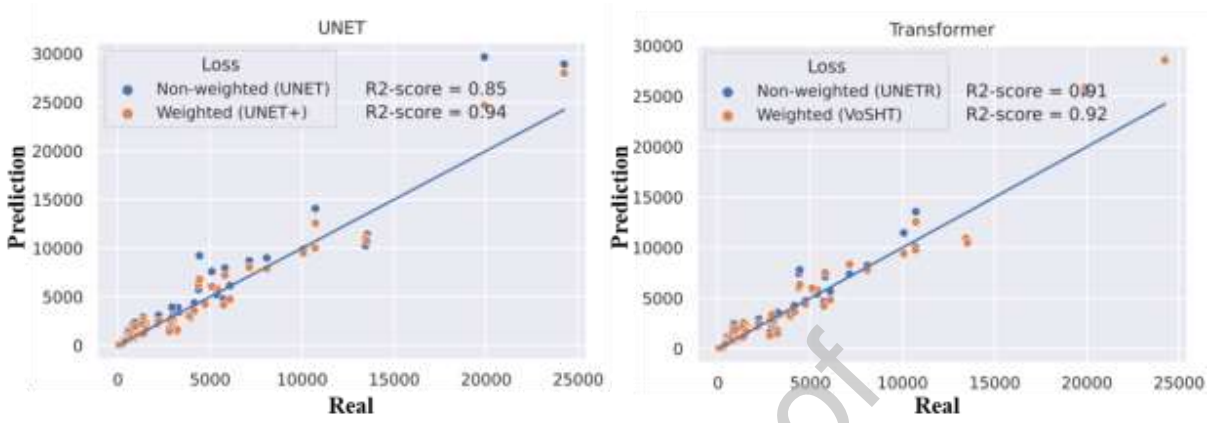

Figure 7. Scatter plot of manual WMH volume against predicted WMH volume on the LISA test set for UNET and UNETR. Each point on the plot is associated with the total number of lesions voxels at the subject level. The blue and orange color codes present the non-weighted and the weighted loss functions, respectively.

Regarding the weighted loss function, it is important to mention that the specific weighting scheme is motivated by prior knowledge about the expected location of WMH, this means that a model trained with the weighted loss function will highly penalize non-WM voxels. However, in other settings, lesions may be more likely to appear in non-WM tissues. Therefore, we do not expect the model to immediately generalize well to such populations. In multiple sclerosis (MS), for instance, the importance of mapping lesions on GM is substantial [31]. Therefore, it is important to mention that the utilized weighted loss function should be modified regarding the recurring location of lesions in MS or other relevant applications.

In this study, the following insights were gained: (1) All DL models performed comparably among in-distribution test set. (2) In an out-of-distribution test set, nnUNet with incorporation of 2D and volumetric segmentation gained slightly superior performance. (3) The utilization of T1-weighted sequences as input leads to a significant increase in the number of parameters, whereas

the weighted loss function notably enhances performance without relying on T1-weighted input directly. (4) Ensemble DL models demonstrate superior performance across both cluster-wise and voxel-wise metrics, underscoring their effectiveness in enhancing segmentation accuracy and robustness. (5) It is important to mention that the introduction of selected structural features here is tailored specifically for WMH, and it might be of interest to investigate other ways to introduce structural information in other settings such as the detection of lesions in MS. (6) DL models have a promising potential to assist radiographers and minimize the efforts needed for segmentation and improving the accuracy by allowing the radiographer to pay more attention to specifically difficult cases [33].

Based on the insights gained from this study, further investigation of methods that utilize a combination of attention mechanisms, anatomical information and ensemble weighting would be interesting. Additionally, it would be of interest to explore different methods of generalization, such as domain adaptation techniques [32], which may provide advantages when multi-site MR imaging data is considered. Further, exploring the potential of pretraining the model using self-supervision models [33] to improve its performance is a promising direction for future research. Based on the results obtained from the cohort-level study, explainable blocks [34] could also be utilized as post-hoc modules to direct the attention of human annotators and thereby enhance the prediction, especially FN. Finally, the combination of human expertise and automated models has immense potential to create a synergistic effect, improving the identification and achieving reliable segmentation tools [5,35].

5. Conclusion

In summary, we trained and evaluated several DL-based segmentation frameworks for WMH detection. Generally, all DL models performed at a comparable level when infering lesions in the

in-distribution test set. Additionally, we found that the nnUNet which relies on an ensemble of 2D and volumetric models generalized best to the external validation dataset.

Declaration of Competing Interests

HRS has received honoraria as speaker from Sanofi Genzyme, Denmark, Lundbeck AS, Denmark, and Novartis, Denmark, as consultant from Sanofi Genzyme, Denmark, Lophora, Denmark, and Lundbeck AS, Denmark, and as editor-in-chief (Neuroimage Clinical) and senior editor (NeuroImage) from Elsevier Publishers, Amsterdam, The Netherlands. He has received royalties as book editor from Springer Publishers, Stuttgart, Germany, and from Gyldendal Publishers, Copenhagen, Denmark. The remaining authors declare that they have no known conflicts of interest.

Acknowledgments

We would like to thank Sascha Gude and Jasmin Merhout from the Reader Center at DRCMR for their invaluable contribution in providing the manual lesions drawing.

Funding

This study was supported by the Capital Region of Denmark. ND is supported by funding from the Lundbeck Foundation (Grant Nr. R380-2021-1269). HRS holds a 5-year professorship in precision medicine at the Faculty of Health Sciences and Medicine, University of Copenhagen which is sponsored by the Lundbeck Foundation (Grant Nr. R186-2015-2138).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) occasionally used ChatGPT[3] in order to improve the readability and coherence of the text. After using this tool/service, the author(s)

[3] https://chat.openai.com/chat

reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1. Gebeily S, Fares Y, Kordahi M, Khodeir P, Labaki G, Fazekas F. Cerebral white matter hyperintensities (WMH): an analysis of cerebrovascular risk factors in Lebanon. Int J Neurosci. 2014;124: 799–805. doi:10.3109/00207454.2014.884087

2. Chutinet A, Rost NS. White matter disease as a biomarker for long-term cerebrovascular disease and dementia. Curr Treat Options Cardiovasc Med. 2014;16: 292. doi:10.1007/s11936-013-0292-z

3. Zhu W, Huang H, Zhou Y, Shi F, Shen H, Chen R, et al. Automatic segmentation of white matter hyperintensities in routine clinical brain MRI by 2D VB-Net: A large-scale study. Front Aging Neurosci. 2022;14: 915009. doi:10.3389/fnagi.2022.915009

4. Hamilton OKL, Cox SR, Okely JA, Conte F, Ballerini L, Bastin ME, et al. Cerebral small vessel disease burden and longitudinal cognitive decline from age 73 to 82: the Lothian Birth Cohort 1936. Transl Psychiatry. 2021;11: 376. doi:10.1038/s41398-021-01495-4

5. Sorantin E, Grasser MG, Hemmelmayr A, Tschauner S, Hrzic F, Weiss V, et al. The augmented radiologist: artificial intelligence in the practice of radiology. Pediatr Radiol. 2022;52: 2074–2086. doi:10.1007/s00247-021-05177-7

6. Pozorski V, Oh JM, Okonkwo O, Krislov S, Barzgari A, Theisen F, et al. Cross-sectional and longitudinal associations between total and regional white matter hyperintensity volume and cognitive and motor function in Parkinson's disease. Neuroimage Clin. 2019;23: 101870. doi:10.1016/j.nicl.2019.101870

7. Schmidt P. Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. Text.PhDThesis, Ludwig-Maximilians-Universität München. 2017. doi:10.5282/edoc.20373

8. Sundaresan V, Zamboni G, Le Heron C, Rothwell PM, Husain M, Battaglini M, et al. Automated lesion segmentation with BIANCA: Impact of population-level features, classification algorithm and locally adaptive thresholding. Neuroimage. 2019;202: 116056. doi:10.1016/j.neuroimage.2019.116056

9. Hotz I, Deschwanden PF, Liem F, Mérillat S, Malagurski B, Kollias S, et al. Performance of three freely available methods for extracting white matter hyperintensities: FreeSurfer, UBO Detector, and BIANCA. Hum Brain Mapp. 2022;43: 1481–1500. doi:10.1002/hbm.25739

10. Huang P, Zhang R, Jiaerken Y, Wang S, Yu W, Hong H, et al. Deep white matter hyperintensity is associated with the dilation of perivascular space. J Cereb Blood Flow Metab. 2021;41: 2370–2380. doi:10.1177/0271678X211002279

11. Boutzoukas EM, O'Shea A, Albizu A, Evangelista ND, Hausman HK, Kraft JN, et al. Frontal White

Matter Hyperintensities and Executive Functioning Performance in Older Adults. Front Aging Neurosci. 2021;13: 672535. doi:10.3389/fnagi.2021.672535

12. Sundaresan V, Zamboni G, Rothwell PM, Jenkinson M, Griffanti L. Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. Med Image Anal. 2021;73: 102184. doi:10.1016/j.media.2021.102184

13. Avesta A, Hossain S, Lin M, Aboian M, Krumholz HM, Aneja S. Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-Segmentation. Bioengineering (Basel). 2023;10. doi:10.3390/bioengineering10020181

14. Baldeon Calisto M, Lai-Yuen SK. AdaEn-Net: An ensemble of adaptive 2D-3D Fully Convolutional Networks for medical image segmentation. Neural Netw. 2020;126: 76–94. doi:10.1016/j.neunet.2020.03.007

15. Yan Q, Liu S, Xu S, Dong C, Li Z, Shi JQ, et al. 3D Medical image segmentation using parallel transformers. Pattern Recognit. 2023;138: 109432. doi:10.1016/j.patcog.2023.109432

16. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR. 2021. Available: https://openreview.net/forum?id=YicbFdNTTy&utm_campaign=f86497ed3a-EMAIL_CAMPAIGN_2019_04_24_03_18_COPY_01&utm_medium=email&utm_source=Deep%20Learning%20Weekly&utm_term=0_384567b42d-f86497ed3a-72965345

17. Viteri JA, Piguave BV, Pelaez E, Loayza FR. Automatic Brain White Matter Hyperintensities Segmentation with Swin U-Net. 2022 IEEE ANDESCON. 2022. pp. 1–6. doi:10.1109/ANDESCON56260.2022.9989775

18. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: Transformers for 3D Medical Image Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022. Available: http://arxiv.org/abs/2103.10504

19. Tsuchida A, Boutinaud P, Verrecchia V, Tzourio C, Debette S, Joliot M. Early detection of white matter hyperintensities using SHIVA-WMH detector. bioRxiv. 2023. p. 2023.02.03.526961. doi:10.1101/2023.02.03.526961

20. Gylling AT, Eriksen CS, Garde E, Wimmelmann CL, Reislev NL, Bieler T, et al. The influence of prolonged strength training upon muscle and fat in healthy and chronically diseased older adults. Exp Gerontol. 2020;136: 110939. doi:10.1016/j.exger.2020.110939

21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017. pp. 5998–6008. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

22. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Medical Image Computing and Computer-Assisted Intervention – MICCAI. 2016. Available: https://link.springer.com/chapter/10.1007/978-3-319-46723-8_49#citeas

23. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA. 2016. Available: https://ieeexplore.ieee.org/document/7785132

24. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. ICLR. 2018. Available: https://openreview.net/forum?id=Bkg6RiCqY7

25. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist Assoc Comput Linguist. 2008;34: 555–596. doi:10.1162/coli.07-034-r2

26. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. Bildverarbeitung für die Medizin, Springer. 2019. Available: https://link.springer.com/book/10.1007/978-3-658-25326-4

27. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.

28. Ben Hamida A, Devanne M, Weber J, Truntzer C, Derangère V, Ghiringhelli F, et al. Deep learning for colon cancer histopathological images analysis. Comput Biol Med. 2021;136: 104730. doi:10.1016/j.compbiomed.2021.104730

29. Bogoya JM, Vargas A, Schütze O. The Averaged Hausdorff Distances in Multi-Objective Optimization: A Review. Sci China Ser A Math. 2019;7: 894. doi:10.3390/math7100894

30. Zoetmulder R, Gavves E, Caan M, Marquering H. Domain- and task-specific transfer learning for medical segmentation tasks. Comput Methods Programs Biomed. 2022;214: 106539. doi:10.1016/j.cmpb.2021.106539

31. Ontaneda D, Raza PC, Mahajan KR, Arnold DL, Dwyer MG, Gauthier SA, et al. Deep grey matter injury in multiple sclerosis: a NAIMS consensus statement. Brain. 2021;144: 1974–1984. doi:10.1093/brain/awab132

32. Lu Y, Perer A. An Interactive Interpretability System for Breast Cancer Screening with Deep Learning. arXiv [eess.IV]. 2022. Available: http://arxiv.org/abs/2210.08979

33. Tang Y, Yang D, Li W, Roth H, Landman B, Xu D, et al. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. 2022. Available: https://ieeexplore.ieee.org/document/9879123

34. Moradi M, Samwald M. Post-hoc explanation of black-box classifiers using confident itemsets. Elsevier, Expert Systems with Applications. 2021. Available: https://www.sciencedirect.com/science/article/pii/S0957417420307302

35. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. 36th Conference on Neural Information Processing Systems (NeurIPS). 2022. Available: https://openreview.net/pdf?id=TG8KACxEON

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Hartwig Roman Siebner has received honoraria as speaker from Sanofi Genzyme, Denmark, Lundbeck AS, Denmark, and Novartis, Denmark, as consultant from Sanofi Genzyme, Denmark, Lophora, Denmark, and Lundbeck AS, Denmark, and as editor-in-chief (Neuroimage Clinical) and senior editor (NeuroImage) from Elsevier Publishers, Amsterdam, The Netherlands. He has received royalties as book editor from Springer Publishers, Stuttgart, Germany, and from Gyldendal Publishers, Copenhagen, Denmark. The remaining authors declare that they have no known conflicts of interest.