UC Davis

UC Davis Previously Published Works

Title

Fully automated whole brain segmentation from rat MRI scans with a convolutional neural network

Permalink

https://escholarship.org/uc/item/0969v69g

Authors

Porter, Valerie A

Hobson, Brad A

Foster, Brent

et al.

Publication Date

2024-05-01

DOI

10.1016/j.jneumeth.2024.110078

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NoDerivatives License, available at https://creativecommons.org/licenses/by-nd/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Journal of Neuroscience Methods 405 (2024) 110078

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Fully automated whole brain segmentation from rat MRI scans with a convolutional neural network

Valerie A. Porter [a,b], Brad A. Hobson [a,c], Brent Foster [d], Pamela J. Lein [e], Abhijit J. Chaudhari [b,c,*,1]

[a] Department of Biomedical Engineering, University of California, Davis, CA 95616, USA
[b] Department of Radiology, University of California, Davis, CA 95817, USA
[c] Center for Molecular and Genomic Imaging, University of California, Davis, CA 95616, USA
[d] TechMah Medical LLC, 2099 Thunderhead Rd, Knoxville, TN 37922, USA
[e] Department of Molecular Biosciences, University of California, Davis, CA 95616, USA

ARTICLE INFO

Keywords:
Automated Segmentation
Machine Learning
Preclinical Neuroimaging
Rodent Brain Imaging
MRI
Skull Stripping

ABSTRACT

Background: Whole brain delineation (WBD) is utilized in neuroimaging analysis for data preprocessing and deriving whole brain image metrics. Current automated WBD techniques for analysis of preclinical brain MRI data show limited accuracy when images present with significant neuropathology and anatomical deformations, such as that resulting from organophosphate intoxication (OPI) and Alzheimer's Disease (AD), and inadequate generalizability.

Methods: A modified 2D U-Net framework was employed for WBD of MRI rodent brains, consisting of 27 convolutional layers, batch normalization, two dropout layers and data augmentation, after training parameter optimization. A total of 265 T_2-weighted 7.0 T MRI scans were utilized for the study, including 125 scans of an OPI rat model for neural network training. For testing and validation, 20 OPI rat scans and 120 scans of an AD rat model were utilized. U-Net performance was evaluated using Dice coefficients (DC) and Hausdorff distances (HD) between the U-Net-generated and manually segmented WBDs.

Results: The U-Net achieved a DC (median[range]) of 0.984[0.936–0.990] and HD of 1.69[1.01–6.78] mm for OPI rat model scans, and a DC (mean[range]) of 0.975[0.898–0.991] and HD of 1.49[0.86–3.89] for the AD rat model scans.

Comparison with existing methods: The proposed approach is fully automated and robust across two rat strains and longitudinal brain changes with a computational speed of 8 seconds/scan, overcoming limitations of manual segmentation.

Conclusions: The modified 2D U-Net provided a fully automated, efficient, and generalizable segmentation approach that achieved high accuracy across two disparate rat models of neurological diseases.

1. Introduction

Preclinical magnetic resonance imaging (MRI) is an important tool for the non-invasive assessment of brain structure, function, and pathology in models of human disease (Cunha et al., 2014; Denic et al., 2011). It has found application for detailed cross-sectional and longitudinal monitoring of the brain and therefore, assessment of disease progression and treatment response (Eed et al., 2020; Ni et al., 2021; Prescott, 2013). Furthermore, rodent brain MRI scans are often fused with image data from other modalities, such as Positron Emission Tomography (PET), to provide anatomic and complementary information (Hutchins et al., 2008; Judenhofer and Cherry, 2013). Quantitative metrics can be extracted from brain MRI and complementary modalities, and have shown utility as biomarkers for disease screening, diagnosis, prognosis, and evaluating treatment response (Fowler et al., 2022; Hobson et al., 2017; Wolf and Abolmaali, 2013).

Whole brain delineation (WBD), also known as skull stripping, is the removal of unwanted, non-brain signal from MR images and is an

* Corresponding author at: Department of Radiology, University of California, Davis, CA 95817, USA
E-mail address: ajchaudhari@udavis.edu (A.J. Chaudhari).
[1] ***Present Address:*** Department of Radiology, University of California Davis, 4860 Y Street, Suite 3100, Sacramento, CA 95817, USA,

https://doi.org/10.1016/j.jneumeth.2024.110078
Received 14 November 2023; Received in revised form 27 January 2024; Accepted 5 February 2024
Available online 8 February 2024
0165-0270/© 2024 Elsevier B.V. All rights reserved.

integral component of most sophisticated neuroimaging analysis toolkits available today (Ashburner, 2012; Avants et al., 2011; Fischl, 2012; Shattuck and Leahy, 2002). WBD is often a precursor to regional delineation in MRI, which in turn is essential for quantification of regional biomarkers for neurologic diseases or disorders (Jack et al., 2010). WBD also benefits multimodality image registration by excluding artifacts caused by anatomical structures such as the eyes, ears, and the jaw. Manual segmentation, where an expert outlines the brain from MRI scans, is considered the gold standard for performing WBD. While this technique can produce robust results, it is often time-consuming, subjective, and laborious (Feo and Giove, 2019). To address this issue, a range of semi- or fully automated analytical computational methods have been developed and have been implemented within commonly used brain images analysis tools (Avants et al., 2011; Fedorov et al., 2012; Fischl, 2012; Jenkinson et al., 2012; Shattuck and Leahy, 2002). Despite their wide-spread availability, these methods are often optimized for clinical MRI acquisition parameters such as spatial resolution, matrix size, and T_1-weighted image contrast (Avants et al., 2011; Fischl, 2012; Jenkinson et al., 2012; Shattuck and Leahy, 2002). These methods, therefore, perform suboptimally for small animal MRI processing where differences in MR scanner hardware, and inter-species differences in brain size and shape, lead to significant variation in spatial resolution, signal-to-noise-ratio and selection of MRI contrast mechanism (Feo and Giove, 2019). Furthermore, even when optimized for preclinical models such as rodents (Klein et al., 2010), these methods can fail in the context of severe neuropathology. In such cases, manual segmentation has remained the most practical option for image analysis.

Recently, advances in machine learning (ML) have provided candidate automated methods for WBD (Feo and Giove, 2019; Lenchik et al., 2019), including those based on convolutional neural networks (CNNs). CNNs, such as the U-Net have been shown to be effective for image segmentation (Azad et al., 2022), however, current studies are frequently limited to one or multiple healthy animal datasets (Gao et al., 2021; Hsu et al., 2020; Liang et al., 2023; Liu et al., 2020; Pontes-Filho et al., 2022), or a single neurological disease model (Chang et al., 2023). Network parameters from these approaches are not easily transferable to other experimental models without significant additional training data (Davatzikos, 2019; Weiss et al., 2020). Thus, there is a need for developing ML frameworks that are generalizable across multiple datasets, disease models, and phenotypes, with limited or no additional training.

We developed a modified 2D U-Net framework for automated WBD of MRI brain scans of two anatomically disparate rat models: (1) a Sprague-Dawley rat model of acute organophosphate intoxication (Hobson et al., 2017; Siso et al., 2017), and (2) a transgenic Fischer rat (TgF344) model of Alzheimer's Disease pathology (Cohen et al., 2013; van Oostveen and de Lange, 2021). These rodent models present a wide range of anatomic variation and neuropathology that are classic challenges for analytical segmentation methods. We improved the network's generalizability to both models through the careful optimization of network training parameters and designing of data normalization and augmentation strategies to accommodate common scenarios in multi-strain MRI-based neuroimaging. The performance of the neural network was assessed through a comparison of WBD generated by the U-Net versus manual segmentation.

2. Materials and methods

2.1. Datasets and animal models

All experiments with animals were approved by the UC Davis Institutional Animal Care and Use Committee and used facilities fully accredited by AAALAC International. All experimentation was in accordance with the National Institute of Health (NIH) Guide for The Care and Use of Laboratory Animals (NIH publication No. 8023, revised 1978). Brain MRI scans of the following two rat models were utilized.

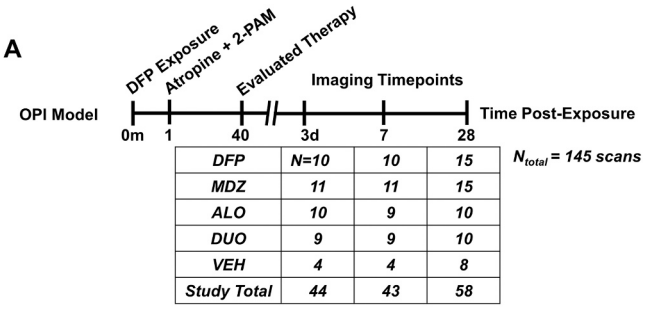
2.1.1. Rat model of acute organophosphate intoxication (OPI)

Adult male Sprague Dawley rats were acutely intoxicated with the organophosphate cholinesterase inhibitor, diisopropylfluorophosphate (DFP), as a part of a drug development study (DAlmeida et al, 2024) evaluating novel neuroprotective therapies as previously described (Dhir et al., 2020). Rats received a single subcutaneous injection of DFP (4 mg/kg; Sigma Chemical Company, St. Louis, MO, USA), or vehicle control (phosphate-buffered saline, 3.6 mM Na2HPO4, 1.4 mM NaH2PO4, 150 mM NaCl; pH 7.2), followed 1 minute later by a combined injection (im) of atropine (2 mg/kg,; Sigma; >97% purity) and 2-PAM (25 mg/kg, Sigma; >99% purity) to increase survival. Forty minutes after injection, DFP-exposed animals were further randomized into one of four interventional groups: no intervention (DFP), midazolam (MDZ, 1.8 mg/kg, im), allopregnanolone (ALO, 24 mg/kg, im), or combined MDZ and ALO (DUO), administered intramuscularly (DAlmeida et al, 2024). Vehicle control animals (VEH) were not intoxicated with DFP. MRI brain scans were acquired at three timepoints (3-, 7-, and 28-days post-DFP) for each group (Fig. 1**A**).

This rat model has been shown to present with severe neuropathology as a consequence of OP-induced *status* epilepticus (Siso et al., 2017), including significant regional brain atrophy that causes large changes in brain structure (Hobson et al., 2017). Intervention with therapy, such as MDZ, has been shown to acutely attenuate the severity of neuropathology (Supasai et al., 2020). Thereby, MR data from these animals provide a broad range of imaging features, such as lesions or infarcts, for neural network training.

2.1.2. Rat model of Alzheimer's disease (AD)

The TgF344-AD rat model is a double knock-in gene model of familial AD that expresses two human AD risk genes, APPswe (K670N, M671L) and PS1ΔE9 (deletion of the exon 9 Presenilin 1 gene) (Cohen et al., 2013; Saré et al., 2020). Phenotypic characterization of the TgF344-AD rat suggests that it recapitulates the progression and pathological hallmarks of human AD, including amyloid plaque formation, neurofibrillary tau tangles, and age-dependent neuronal loss and cognitive decline that is not typically seen in mouse models of AD (Saré



DFP	*N=10*	*10*	*15*	N_{total} = 145 scans
MDZ	*11*	*11*	*15*	
ALO	*10*	*9*	*10*	
DUO	*9*	*9*	*10*	
VEH	*4*	*4*	*8*	
Study Total	*44*	*43*	*58*	

	N=48	N=36	N=24	N=12	N_{total} = 120 scans
TG	*24*	*18*	*12*	*6*	
WT	*24*	*18*	*12*	*6*	
Study Total	*48*	*36*	*24*	*12*	

Fig. 1. : Schematic illustrating the experimental study design of the rat models of: (**A**) OPI, and (**B**) AD. (A) OPI paradigm, where DFP is administered to each animal, followed by the initial rescue therapy, atropine and 2-PAM, 1 minute later. The therapy (MDZ, ALO, or DUO) is administered 40 minutes post-injection of DFP. T_2-weighted MRI scans are captured at each timepoint (3-, 7-, and 28-days post-exposure); 55 unique rats were imaged. (B) The AD rat model was imaged with T_2-weighted MRI at 7, 9, 11, and 13 months of age; 48 unique rats were imaged. At each timepoint, six animals from each group, transgenic (TG) vesus wildtype (WT), were euthenized for histology. The tables below the imaging timelines indicate the number scans captured by timepoint and group.

et al., 2020).

Female Fischer 344 rat were scanned in a prospective longitudinal imaging study to characterize the spatiotemporal evolution of AD. Animal were separated into two genotypic groups, transgenic (TG) and matched congenic controls (WT). MRI brain scans were acquired at four timepoints (Rat ages: 7-, 9-, 11- and 13-months) for both groups (Fig. 1**B**).

2.2. MR acquisition protocol

MRI brain scans were acquired on a Biospec 70/30 (7 T) preclinical MR scanner running Paravision 6.0 (Bruker BioSpin MRI, Ettlingen, Germany), equipped with a 116-mm internal diameter (ID) B-GA12S gradient (450 mT/m, 4500 T/m/s), and a 72 mm ID volume coil and 20 mm ID surface coil for signal transmission and reception, respectively. Rats were anesthetized with isoflurane/O2 (Piramal Healthcare, Bethlehem, Pennsylvania) using 2.0%–3.0% vol/vol to induce and 1.0%–2.0% vol/vol to maintain anesthesia, stereotactically mounted to the MR scanner bed, and imaged as described previously (Hobson et al., 2017). For the present study, multi-slice, T_2-weighted, Rapid Acquisition with Repeated Echoes (RARE) axial images were collected using the following parameters: repetition time (TR) = 6100 ms; echo time (TE) = 15 ms; RARE factor 8; averages = 4; field of view (FOV) = 35 × 25 mm^2, with an in-plane data matrix of 280 × 200, resulting in a data set resolution of 0.125 × 0.125 mm^2; 59 slices with a 0.5 mm thickness spanning approximately the posterior aspect of the eyes through the anterior aspect of the spinal cord, 11 mm to −18.5 mm bregma. Subsets of animals received additional MR scans for T_2 parametric mapping and PET to assess neuroinflammation. These data are the subject of separate publications and, therefore, will not be discussed further herein.

2.3. Manual segmentation

A protocol was developed to manually outline the brain on 2D MRI slices from the image volume based on the Paxinos and Watson's *The Rat Brain in Stereotaxic Coordinates* (Paxinos and Watson, 2007). Scans were manually segmented from approximately 6 mm to −14 mm bregma, by an observer blinded to the group assignments. The trigeminal nerves, olfactory bulb, pituitary, and brainstem were not included in the segmentations. Manual segmentation accuracy was assessed additionally by an expert in rodent neuroanatomy with greater than 10 years of experience, and discrepancies were resolved by consensus upon rigorous re-evaluation of the scans.

2.4. Image pre- and post-processing

Scans included in the present study contained biological and technical variation reflected in T_2-weighted image contrast, signal intensity, and spatial localization of the brain within the field-of-view. Therefore, to normalize input image data and improve the U-Net training performance we created an image pre-processing pipeline that applies bias correction, standardized image cropping, resampling, and normalization. In this pipeline, we applied N4ITK bias correction with the SimpleITK python library (Beare et al., 2018) to reduce signal drop off and homogenize signal intensities of brain tissue. We center-cropped the MRI scans and their matching manually segmented images (from [280×200] to [200×200]), removing areas to the left and right side of the brain without brain signal. We then resampled each volume with bicubic interpolation to [128×128] to enhance convergence of the U-Net. Lastly, we performed intensity normalization with a range between 0 and 1. The training and testing data were randomly selected to ensure representation across treatment groups and timepoints to provide a similar number of scans across treatment groups for training and validation.

Once each scan had been processed through the neural network, the label volumes were converted to original MRI data space dimensions by reversing the pre-processing steps. That is, U-Net-generated label volume [128×128] were up-sampled to [200×200] using nearest-neighbor interpolation, zero padding was performed to add empty voxels to the lateral edges of the image to convert it back to a matrix size of [280×200], and finally, we concatenated the 2D label images for each rat into a 3D label volume that matched the dimensions of the original MR data.

2.5. Training and test dataset composition

To train the U-Net, we utilized 145 OPI rat scans. These data were subdivided into training and test datasets for the U-Net consisting of 125 randomly chosen scans (DFP = 31, MDZ = 33, ALO = 25, DUO = 24, VEH = 12) and the remainder 20 scans (DFP = 4, MDZ = 4, ALO = 4, DUO = 4, VEH = 4), respectively. A validation dataset consisted of a single randomly selected scans within the training dataset. This OPI-rat-trained U-Net and weights were then utilized for WBD of the AD rat dataset (TG = 60, WT = 60) without any additional training.

2.6. U-net architecture and data augmentation

We used the 2D U-Net architecture (Ronneberger et al., 2015), with modifications (Fig. 2). The left half of the U-Net is the analysis path, where the network performs down-sampling convolutions that analyze the image based on important features for segmentation. Each convolution creates feature maps used to determine the classification of each pixel into brain and non-brain tissue. Feature maps are then utilized in the synthesis path, the right side of the U-Net, by performing up-convolutions to create a segmentation label map for the image. We modified the architecture by using padded convolutions with a 3×3 kernel for each layer and used the leaky rectified linear unit (ReLU) as the activation function. We used padded convolutions since our data input is already preprocessed to the correct size, and we used leaky ReLU to speed up the training time and improve weight training. We added a batch normalization layer to the first step and a drop-out layer (rate = 0.2) to the fourth and fifth steps in the U-Net architecture. The batch normalization was used to normalize the batched input data between 0 and 1, and we added the drop-out layers to reduce overfitting of the weights during training. Lastly, we trained the network with images down-sampled to 128×128 pixels. In total the network had 23 convolutional layers, 4 max pooling layers with zero padding, and 2 dropout layers.

The output labels are a probability label map that are binarized by using a threshold value (T), where if the label voxel value is greater than or equal to T, the voxel is set to 1, and vice versa. During training, the MR volume scans were picked randomly from the training dataset and loaded as 2D slices. For each slice that has a corresponding non-zero label, data augmentation was applied to enhance the model's generalization by artificially increasing training images to emulate the distribution of the test data. The network optimized weights based on the non-zero label MR slices. The output of the network was a 2D label image that was concatenated to create a 3D label map volume. When applying the weights to test data, a T value was applied to the label map to convert label pixels into an integer mask. The final output was a 3D label that delineated the whole brain from the MR scan. This framework was trained with an Intel Core i7–6800 K CPU equipped with 128 GB of GDDR4 memory operating at 3.40 GHz and an NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of dedicated VRAM.

Data augmentation was performed with Medical Open Network for Artificial Intelligence (MONAI)(The MONAI Consortium, 2020), a python library, to improve the training performance of the U-Net. The augmentation function parameters we utilized were random shift by [−40, 40] pixels in the x- and y- direction to account for the different locations of the brain in the field-of-view, rotation between [−45, 45] degrees to increase the diversity of head and brain poses, flipping the image horizontally and vertically with a probability of 0.5 to increase

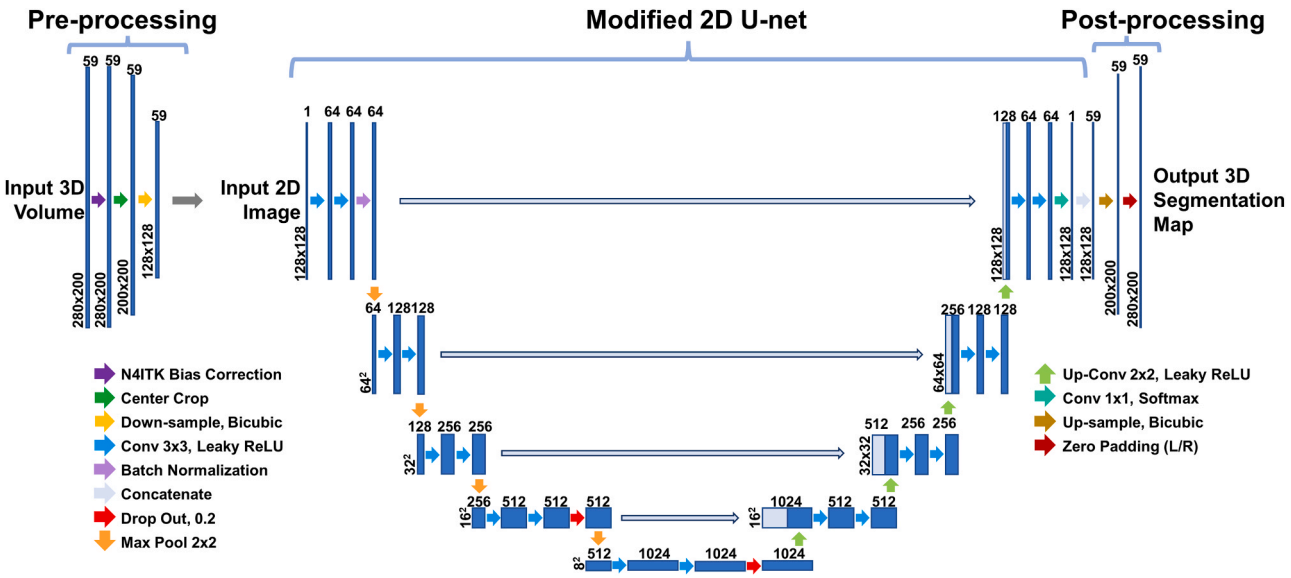


Fig. 2. : Architecture of the segmentation pipeline utilized; Each blue box is an image volume, where the x- and y-dimensions are denoted in the lower left of the box, and the number of slices (z-dimension) is denoted above the box. The arrows indicate different operations in the pipeline and the order in which operations are applied to the image. For the modified 2D U-Net architecture, each slice of the scan is processed through the neural network individually, and the z-axis indicates the number of feature maps generated. The light blue boxes represent concatenated feature maps from previous layers. Post-processing converts the U-Net-generated segmentation map to the original size of the input scan.

the variety of image orientation, random scaling with a factor between [−0.3, 0.3] to increase variation in size, shifting the brightness levels by ±0.5 to accommodate image contrast variety, and adding gaussian noise (mean=1, st. dev.=0.25). These augmented data were included with the original OPI rat dataset to create the full training dataset.

2.7. Parametric selection for the U-Net

We assessed the impact of four U-Net hyper-parameters, namely, training dataset size (TDS), learning rate (LR), epoch (E), and probability map threshold (T) on segmentation performance. This was performed to increase the generalizability of the network while mitigating the impact of overfitting. Each hyper-parameter was varied, as described below, and upon network training, the performance was assessed on OPI rat test dataset and compared with manual segmentations.

2.7.1. Training dataset size assessment

We evaluated multiple training data sizes while maintaining similar dataset composition across treatment groups (Table 1) and secondarily by timepoint, except for the N=1 case. For the latter case, an MDZ Day 28 rat was used because that treatment group and timepoint was determined to be a median representation of disease pathology between VEH (no detectable pathology) and DFP (severe pathology) groups. Each training dataset size was trained with LR=0.002, E=150, and T= 0.85. More details of the performance of the U-Net are reported in supplementary materials in **Supplementary Table S1**.

Table 1
Composition of training datasets at each TDS evaluated by treatment group. Each TDS was distributed equally across treatment groups. The VEH group had 12 scans total. Thus, for TDS>50, training dataset makeup prioritized MDZ and DFP groups to provide more training data with moderate to severe neuropathology, while maintaining similar numbers of scans across treatment group and timepoint. TDS at 125, used all remaining scans and did not prioritize any treatment group or timepoint.

Treatment Groups	Training Dataset Sizes (TDS)						
	1	10	25	50	75	100	125
DFP	0	2	5	10	15	23	25
MDZ	1	2	5	10	18	23	31
ALO	0	2	5	10	15	21	24
DUO	0	2	5	10	15	21	33
VEH	0	2	5	10	12	12	12

2.7.2. Learning rate evaluation

The learning rate for the U-Net was varied by factors of 10, ranging from 2×10^{-3} to 2×10^{-7}, while keeping the other parameters constant, TDS=100, E=150, and T= 0.85.

2.7.3. Epoch assessment

With TDS = 100, LR = 0.0002, and T= 0.85, the U-Net was trained over 300 epochs and repeated for a total of ten training sessions. The minimum moving average error (Eq. 1) over a range of ten data points for each training metric (training accuracy, training loss, validation accuracy, and validation loss) and its corresponding epoch value were calculated for each training session. The epoch values across all ten training sessions were averaged to determine average number of epochs for each training metric. The moving average error is given by:

$$Moving\ Average\ Error = \frac{\left| \frac{\sum_{i-10}^{i-5} X_i}{N} - \frac{\sum_{i-5}^{i} X_i}{N} \right|}{\frac{\sum_{i-5}^{i} X_i}{N}} \quad (1)$$

where i is the epoch number and $i \geq 10$, provided the range of the epoch value where the U-Net was stable.

Each metric's epoch value was compared to the epoch value where overfitting occurred during training across all ten sessions. The metric with an epoch value that was less than but closest to the overfitting epoch value was considered as the suitable epoch value.

2.7.4. Probability map threshold selection

A range of T values from 0 to 1 at 0.05 intervals were evaluated to binarize the U-Net-generated label volumes. The training parameters utilized are based on the parameters found via the methods mentioned above: TDS=100, LR= 0.0002, and E=150.

2.8. U-Net performance evaluation metrics

To evaluate performance accuracy, the Dice coefficient (DC) of volumetric overlap (Zou et al., 2004), and the Hausdorff distance (HD) were calculated between the U-Net-generated WBD labels and those from manual segmentation. These are commonly used evaluation metrics for image segmentation methods (Crum et al., 2006; Eelbode et al., 2020; Karimi and Salcudean, 2020; Müller et al., 2022; Taha and Hanbury, 2015), including those based on neural networks (Hsu et al., 2021, 2020; Liang et al., 2023). DC is given by:

$$DC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2}$$

where X is the manual segmentation and Y is the U-Net-generated segmentation. The HD was calculated as:

$$HD(X, Y) = \max\{h(X, Y), h(Y, X)\} \tag{3}$$

where $h(X, Y) = \max_{x \in X}\{\min_{y \in Y}\{d(x, y)\}\}$, and $d(x,y)$ is the distance between point cloud X (from manual segmentation) and Y (from U-Net-generated segmentation).

The DC ranges from 0 (no overlap between the U-Net-generated label and manual segmentation) to 1 (perfect match between the two). The HD evaluates the distance between the U-Net-generated and manual segmentation point clouds and the lower this metric, the better concurrence is between the point clouds. In comparison to DC, HD has the advantage that it takes voxel location into consideration.

The accuracy of each segmentation was further evaluated by calculating the true positive rate (TPR) and the false positive rate (FPR) for each voxel (Taha and Hanbury, 2015), with manual segmentation as the ground truth. TPR and FPR were graphed in a receiver operating characteristic (ROC) for threshold optimization. TPR and FPR are defined as follows for the U-Net generated segmentation:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

where TP is the number of true positive pixels, TN is the number of true negative pixels, FP is the number of false positive pixels, and FN is the number of false negative pixels.

For all of these metrics, the values are reported as [median[range]] for the (N=20) OPI rat test dataset and [mean ± sd] for the (N=120) AD rat test dataset due to the size of the test datasets. The DC and HD metrics were calculated within the 3D Slicer program (Fedorov et al., 2012) segment comparison module (Slicer Wiki contributors, 2020) and the TPR and FPR metrics were calculated with Numpy library in Python 3.8 (Harris et al., 2020). Additionally, ROC curves were generated for threshold value analysis in RStudio 4.2.2 (R Core Team, 2022) with R packages ggplot2 (Wickham, 2016) and ggbeeswarm (Clarke et al., 2023).

Lastly, the neural network was timed during training and testing of the method for each dataset and averaged across both datasets. Manual segmentation was timed for comparison with the modified U-Net.

3. Results

3.1. Training dataset size assessment

The median DC and HD [metric: median[range]] reached their highest values at TDS=100 [DC: 0.984 [0.936,0.990]; HD: 1.69 mm [1.01,6.78]] (**Supplementary Table S1**). The median TPR achieved the maximum value at TDS=25 [TPR: 0.991 [0.786,0.994]], while FPR achieved its minimum value [FPR: 1.8×10^{-3} [1.0×10^{-3},4.8×10^{-3}]] at TDS=100. A notable finding was that as TDS increased, each median metric value improved until TDS=25, slightly worsened until achieving their highest value at TDS=100, or their second-most highest value for the TPR. For TPR and FPR, the difference between the maximum and minimum values for each metric were lowest at TDS=100 (Fig. 3).

3.2. Learning rate evaluation

As the learning rate increased to 2×10^{-4}, the median DC [metric: median[range]] improved greatly (see **Supplementary Table S2**). However, at higher learning rates, LR$\geq2\times10^{-3}$, the network did not produce any label maps, a likely byproduct of not finding an optimal solution for the weights (Takase et al., 2018). All metrics achieved their best performance values at LR=2×10^{-4} [DC: 0.984 [0.936,0.990]; HD: 1.69 mm [1.01,6.78]; TPR: 0.989 [0.885,0.993]; FPR: 1.8×10^{-3} [1.0×10^{-3},4.8×10^{-3}]]. For TPR and FPR, the difference between the maximum and minimum values for each metric were lowest at TDS=100 (Fig. 4).

3.3. Epoch assessment

Across 10 training runs, the average epoch value based on training accuracy metric was 179 epochs [accuracy: 0.990, minimum moving average error: 9.09×10^{-6}] (Table 2). For training loss metric, it was 284 epochs [loss: 5.72×10^{-3} and minimum moving average error of 1.44×10^{-3}] (Table 2). For validation accuracy and loss, the number of epochs were 125 [val. accuracy: 0.992, minimum moving average error: 1.79×10^{-3}] and 149 [val. loss: 8.73×10^{-3}, minimum moving average error: 2.34×10^{-3}], respectively. Validation accuracy had the smallest epoch range across all ten runs, suggesting validation accuracy during training improved in a similar manner despite different training runs. Notably, the range of accuracy or loss between runs suggested overfitting starting around 175 epochs. Thus, the mean epoch based on validation loss (149 epochs) was chosen as the value for training to reduce overfitting while maximizing the number of cycles for weight optimization during training (Fig. 5).

3.4. Probability threshold value selection

The ROC curve (Fig. 6A) illustrates how TPR and FPR are affected as T values decrease, from left to right. Zooming in on the knee of the curve (Fig. 6B), there are a range of T values from 0.25 to 0.90 that produce reasonable segmentations, where TPR and FPR minimally affected. Based on this defined range of T values, the highest median DC [metric: median[range]] and the lowest average HD [DC: 0.984 [0.936,0.990]; HD: 1.69 mm [1.01,6.78]] was achieved at T=0.85 (**Supplementary Table S3**). DC and HD values marginally improved between 0.60 and 0.85, suggesting robustness of the U-Net. Thus, T=0.85 was selected as the parameter value.

3.5. U-Net segmentation results of OPI and AD Rat Models

The U-Net training parameters based on the above analysis converged on a LR=2×10^{-4}, and number of epochs=149, rounded to 150 for future analyses after ensuring that there was no impact of the results. The most accurate training of the modified 2D U-Net architecture was with 100 OPI rat training scans and a threshold value of 0.85. Other parameters included steps per epoch=20, and batch size=10. The training runtime was 30 minutes for 150 epochs, while the mean U-Net segmentation computation time was 8 seconds per scan for both the OPI rat and the AD rat dataset, whereas for manual segmentation, it took approximately 30 minutes per scan. Model training accuracy quickly improved until 60 epochs (Fig. 7), where accuracy increased marginally from 98.08% to 99.37% from epochs 60–150.

Accuracy metrics of the 2D U-Net for each model (1: OPI Rat, 2: AD Rat) are organized by group and timepoint and are formatted as [metric:

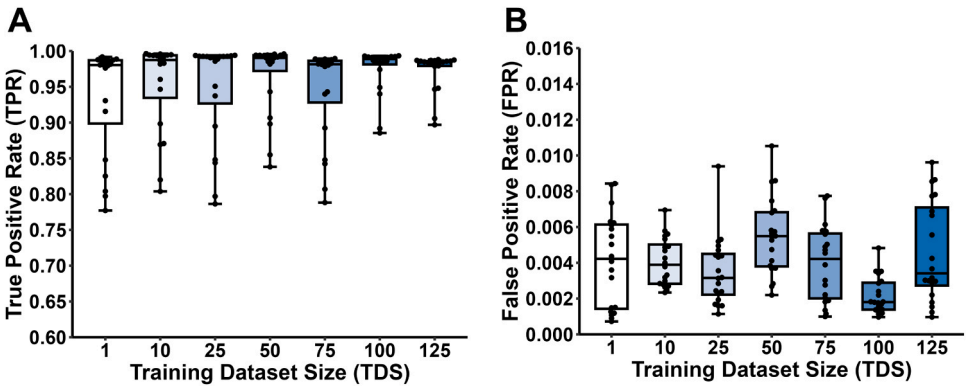


Fig. 3. : Box and whisker plots of (**A**) true positive rate (TPR) and (**B**) false positive rate (FPR) at different training dataset sizes (TDS). A TDS=100 produced the lowest median FPR without decreasing the median TPR. Median, first and third interquartile ranges are shown. Error bars indicate min-max range of data.

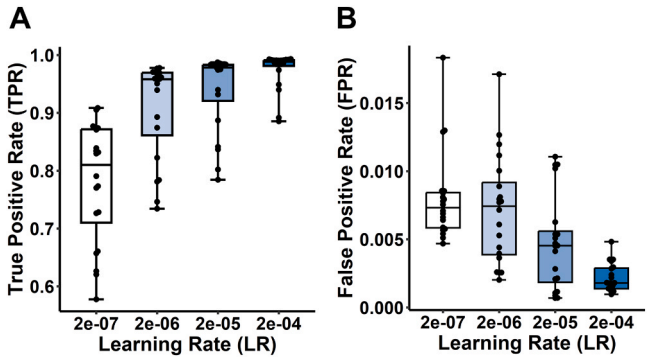


Fig. 4. : Box and whisker plots of (**A**) true positive rate (TPR) and (**B**) false positive rate (FPR) at different learning rates (LR). An LR of 2×10^{-4} produced the highest TPR and the lowest FPR. An LR of 2×10^{-3} did not produce segmentations, so no TPR or FPR values could be calculated. Boxes denote the median, first and third interquartile ranges; whiskers denote range of data.

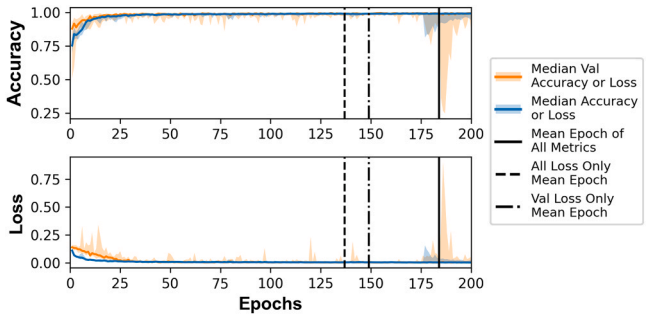


Fig. 5. : Plots of accuracy (**top**) and loss (**bottom**) during training of the neural network; In both graphs, the blue line shows each metric calculated from the training dataset (Tr) and the orange line represents each metric calculated from a single image from randomly selected scan from the training dataset, called the validation dataset (Val). The colored lines represent the median value from ten training runs, and the shaded regions represent the range of values from the ten runs. The vertical solid black lines indicate potential stopping points based on the mean minimum moving average error from all training and validation (Tr + Val) metrics (solid, mean=184), Tr + Val loss metrics (dash, mean=137), and val loss (dash-dot, mean=149). The range of accuracy or loss depicts overfitting starting at 175 epochs. The mean epoch value closest to 175 but with validation loss value less than 175 was epoch=149.

Table 2
Epoch evaluation outcomes of the 2D U-Net for OPI Rat model for 10 training runs [median[range]]. Accuracy and loss measures are calculated from the training dataset, while the validation accuracy and loss are from the validation dataset during training. The neural network training data tend to worsen performance or overfit at epoch=175 or greater. To reduce the chance of overfitting the data, the metric with an average epoch value of $\leq$ 175, was epoch:149. This shows that the metric with lowest minimum moving average error was not the best metric to use for epoch selection. For each metric, the table shows the mean, standard deviation, range, and error at mean epoch.

Training Metric	Value	Epoch	Epoch Range (Min-Max)	Minimum Moving Average Error
Accuracy	0.990 ± 0.002	179 ± 59	37–279	9.09×10^{-6} ± 7.30×10^{-6}
Val. Accuracy	0.992 ± 0.005	284 ± 8	269–290	1.44×10^{-3} ± 8.82×10^{-4}
Loss	5.72×10^{-3} ± 2.01×10^{-3}	125 ± 74	30–260	1.79×10^{-3} ± 1.88×10^{-3}
Val. Loss	8.73×10^{-3} ± 9.80×10^{-3}	149 ± 59	59–264	2.34×10^{-3} ± 2.24×10^{-3}

median[range]] for the OPI Rat dataset (Table 3) and [metric: mean ± sd] for the AD Rat dataset, based on the size of each dataset (Table 4). The most accurate segmentations results were at T=0.85 for both models (Fig. 8). Both models achieved similar DCs and HDs values across all groups (OPI Rat Model: DC:0.984 [0.936,0.990], HD: 1.69 [1.01,6.78], AD Rat Model: DC: 0.975±0.015, HD: 1.49±0.59 mm). For the OPI rat model, the best and worst generated labels (Tables 3 and 4) were an MDZ, Day 3 scan [DC: 0.990, HD: 2.02 mm] and a DUO, Day 28 scan [DC: 0.936, HD: 2.50 mm], respectively, while for the AD rat model, they were a WT, Month 8 [DC: 0.991, HD:1.24 mm] and TG, Month 10 [DC: 0.898, HD: 3.25 mm], respectively. The most consistent tissue classification error with the 2D U-Net appeared at tissue border at the ventral aspect of the brain, between the brain and the trigeminal nerves for both datasets (Fig. 8). For the images with the worst performance, the U-Net failed to find the start and end slices of the brain in the anterior-posterior direction, or did not segment the posterior portions of the cerebellum due to a drop-off in imaging signal.

The trained modified 2D U-Net and it's documentation is available at: https://github.com/ajchaudhari/OPI_Rat_NN.

4. Discussion

In this study we presented a modified 2D U-Net CNN framework for fully automated and efficient WBD that achieved high accuracy across two disparate rat models of neurological disease. Optimization of the framework with one disease model provided robust segmentations that improved the generalization of the U-Net CNN to other disease models. This framework resulted in a 225x reduction in segmentation time compared to manual segmentation (from 30 mins/scan to 8 seconds/scan) and achieved excellent volumetric overlap (DC>0.9) and reasonably well-aligned edge voxels (HD<3 mm). As the spatial resolution of

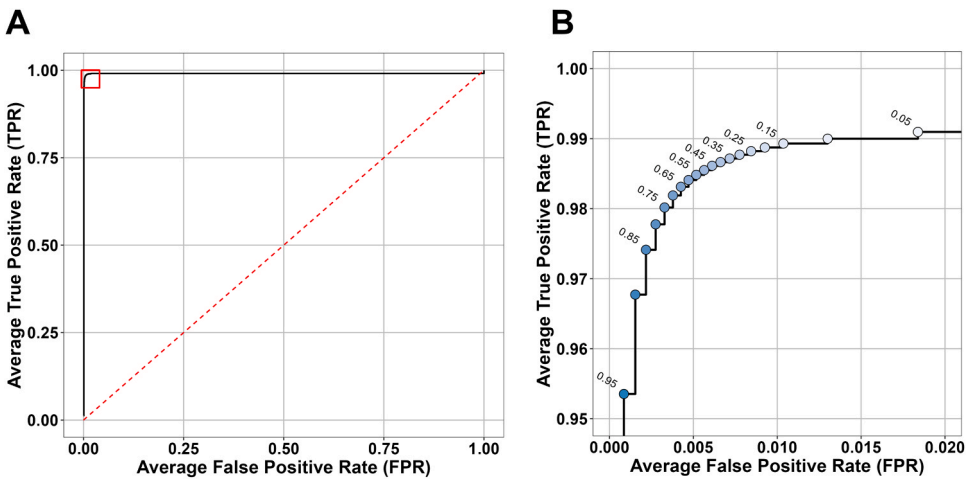


Fig. 6. : Receiver operator characteristics (ROC) analysis as a function of threshold (T) values; (**A**) ROC curve and (**B**) a zoomed-in view of the ROC curve in (**A**), indicated by the red box. The black curve represents the mean TPR and FPR data as the threshold decreases (left to right). The red dashed line indicates the random classifier cutoff. (**A**) indicates that T is a strong classifier for determining brain vs non-brain pixels. In (**B**), the zoomed in graph shows specific values of T between 0.05 and 0.95, in increments of 0.05. Values in increments of 0.10 are listed above the corresponding points. Each point is color coded with a gradient from blue [T=0.95] to white [T=0.05]. Along the knee of the ROC curve, there are a range of values from 0.25 to 0.90 that marginally affect TPR and FPR (Supplementary Table S3). Within that range of T values, DC and HD values indicate that T=0.85 produces the highest level of segmentation accuracy.

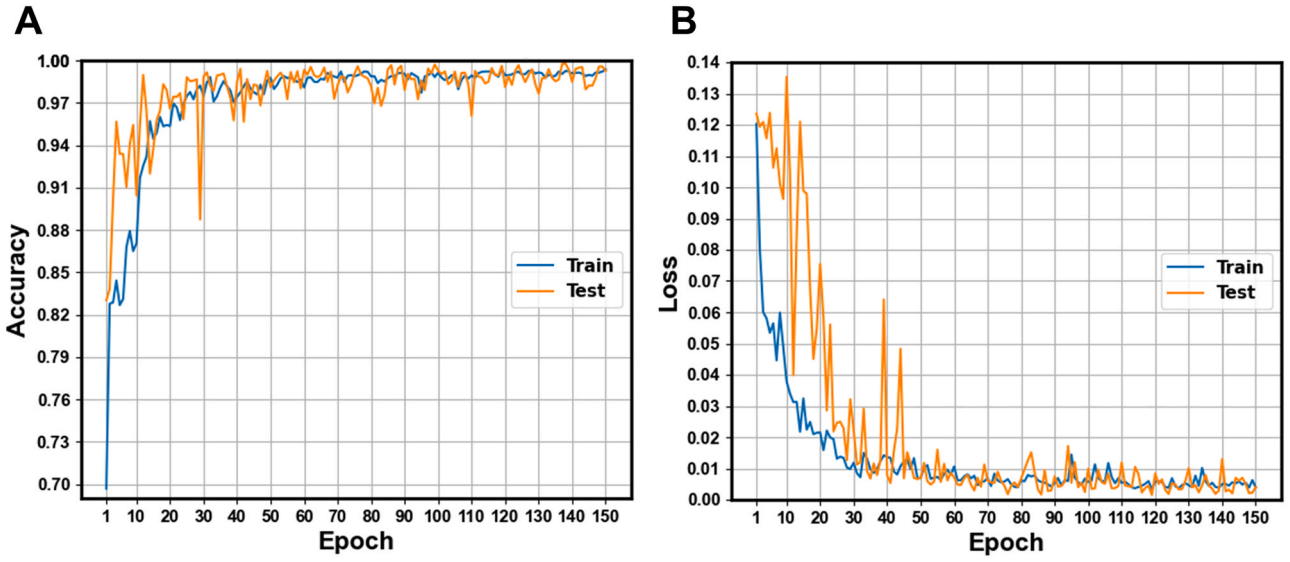


Fig. 7. : Training graphs of the optimized OPI Rat 2D U-Net CNN. Graphs show (**A**) categorical accuracy and (**B**) loss for the model training and validation over 150 epochs. Training data are in blue, and validation data are in orange. Over the last 10 epochs [mean±std]: training accuracy=[0.991±0.001] and loss=[0.0052 ±0.0007], and validation accuracy=[0.989±0.006] and loss=[0.0050±0.0033].

MRI improves, the limitations of manual segmentation are amplified as more slices need to be segmented, thereby efficient and accurate WBD is a growing challenge. The proposed method provides means to potentially address that challenge.

With a mean DC=0.98, our method demonstrated improved results compared to approaches proposed by (Hsu et al., 2020), and (Gao et al., 2021) [DC<0.97]. We achieved comparable overlap metrics with approaches proposed by (Chang et al., 2023), (Hsu et al., 2021) and (Liang et al., 2023). While the approaches proposed by (Hsu et al., 2021) and (Liang et al., 2023) utilized 3D network architectures working with 3D volumes, which result in a higher computational cost, it is encouraging to see that the performance of the proposed 2D U-Net was capable of performing segmentations that are comparable to 3D networks, but at a lower computational cost.

4.1. Generalization of the U-Net CNN

The generalizability of our framework was demonstrated using two disparate imaging datasets. Although both models were rats and utilized the same MR acquisition protocols, brain morphology and size vary significantly between strains of rats (Welniak-Kaminska et al., 2019). Fischer 344 and Sprague-Dawley rats, inbred and outbred strains, respectively, have anatomical brain differences that reduce the generalizability of atlas-based methods of segmentation methods (Goerzen et al., 2020). Nonetheless, our framework produced similar and highly accurate segmentations [metric: mean ± sd] for both OPI [DC: 0.978; HD:1.97] and AD [DC: 0.975; HD:1.49] rat models. This may help address limitations of classic skull-stripping techniques that are often tailored to images with specific strains.

We believe the MRI data from the OPI rat model contributed to the generalization of the U-Net by introducing a wide range of

Table 3
Performance outcomes of the 2D U-Net for OPI Rat test dataset, organized by group and timepoint [median [min,max]]. The U-Net achieved similar median DC and HD values across treatment groups and timepoints. MDZ group achieved the best accuracy with the highest median DC and the lowest median HD, while the DUO group achieved the lowest outcomes. DC values and HD presented were calculated between the manual segmentations and the 2D-U-Net. HD values are in millimeters (mm).

Group	# of Scans	DC	HD
DFP	4	0.985 [0.983–0.989]	1.68 [1.16–2.07]
MDZ	4	0.986 [0.983–0.990]	1.38 [1.33–2.02]
ALO	4	0.978 [0.939–0.985]	1.73 [1.61–2.53]
DUO	4	0.975 [0.936–0.987]	1.90 [1.69–2.50]
VEH	4	0.985 [0.965–0.988]	1.63 [1.01–6.78]
Timepoint	**# of Scans**	**DC**	**HD**
Day 03	5	0.983 [0.967–0.990]	1.87 [1.01–2.02]
Day 07	5	0.984 [0.965–0.989]	1.56 [1.16–1.86]
Day 28	10	0.984 [0.936–0.987]	1.75 [1.35–6.78]

Table 4
Performance outcomes of the 2D U-Net for AD rat test dataset, organized by group and timepoint [mean±sd]. Both groups achieved similar DC and HD values. Month 12 timepoint achieved the best accuracy with the highest mean DC and the lowest mean HD values, whereas Month 8 achieved the lowest. DC values and HD were calculated between the manual segmentations and the 2D-U-Net. HD values are in millimeters (mm).

Group	# of Scans	DC	HD
WT	60	0.974±0.018	1.58±0.69
TG	60	0.977±0.012	1.40±0.46
Timepoint	**# of Scans**	**DC**	**HD**
Month 07	48	0.976±0.012	1.46±0.54
Month 09	36	0.976±0.013	1.41±0.53
Month 11	24	0.968±0.024	1.72±0.81
Month 13	12	0.982±0.041	1.44±0.34

neuropathologies and anatomical variation (e.g., ventricular volume and tissue atrophy) (Hobson et al., 2017) into the training dataset of the network. As a result, we believe that the U-Net optimized its weights based on general image features that are present in all groups rather than for one specific group (e.g., healthy controls). Thus, it was capable of performing WBD on rat brain images across different neurological disease models, without additional training. Our data suggests that the composition of the training data, specifically, its diversification, is important for development of a readily generalizable neural network. While the dataset from the OPI model presented a range of morphology, more research is needed to determine how training dataset composition (e.g., treatment group or severity of disease) affects network performance.

4.2. Crucial parameters for network optimization

The performance of the U-Net showed the most improvement with the optimization of the training dataset size consistent with reports that training dataset quality and quantity are both critical for improving the performance of deep learning methods (Halevy et al., 2009). However, the exact number of images required for optimal training is highly dependent on the type of data and complexity of the task. We observed that increasing the training dataset size improved median HD and DC, and reduced inter-scan variability across accuracy metrics (DC, HD, TPR, and FPR). However, with dataset sizes larger than TDS=100, the median FPR and HD tended to increase, while the other measures remained relatively unchanged, suggesting a fundamental reduction in real-world performance. This is a key finding that contradicts the hypothesis that larger datasets for machine learning methods will create more accurate neural networks (Halevy et al., 2009). Additionally, the

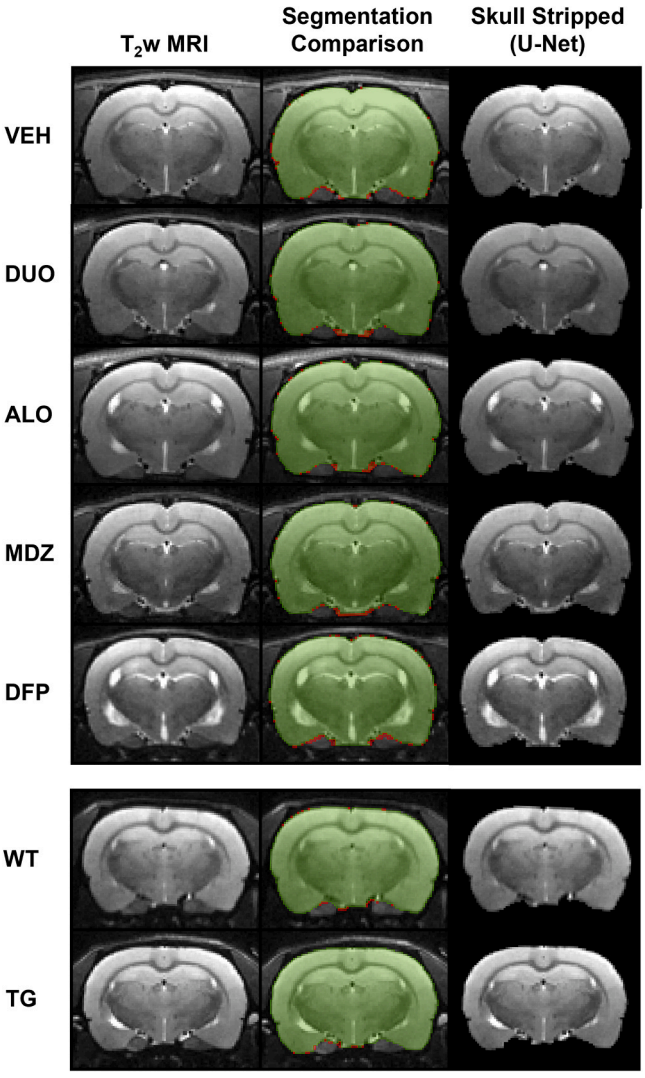


Fig. 8. : Representative images and U-Net-generated segmentations of: (**row 1**) VEH, (**row 2**) DUO, (**row 3**) ALO, (**row 4**) MDZ and (**row 5**) DFP animals from the OPI study, and (**row 6**) WT and (**row 7**) TG animals from the AD study. Columns from left to right: anatomical MR image, 2D U-Net-generated segmentation label (matched pixels in green and unmatched pixels in red) overlaid on MR image, and skull-stripped MR image created with the 2D U-Net-generated label.

threshold value (T) is a critical classifier to convert the output probability segmentation map into a binarized segmentation mask. We believe that this parameter is under-reported in the literature due to its potentially small contribution to improving accuracy measures. However, visually, expert readers were able to detect improvements in the segmentation until T=0.85, despite minimal changes in DC, HD, TPR or FPR. Thus, reporting of T and a visual inspection of results in highly recommended with CNNs or equivalent neural networks.

4.3. Optimization of training parameters for improved efficiency

Optimization of the LR and the number of epochs were important for improving training efficiency of the 2D U-Net. Specifically, we observed increasing LR improved speed of weight optimization in the network, which was indicated by the stabilization of the training and validation accuracy and loss (Fig. 7). We observed that as learning rates increased, the fluctuations in validation accuracy and loss became larger. At LR values greater than 2×10^{-4}, the neural network updated its weights in larger increments (Konar et al., 2020), and this appeared to cause its

validation accuracy and loss to not converge to solutions, and therefore, not generate acceptable WBD segmentations. Thus, the optimal LR for this framework was the fastest value that allows validation accuracy and loss to converge and stabilize.

Optimization of the number of epochs was also shown to be important for training time efficiency and to prevent overfitting of the data. Across ten training sessions, multiple fluctuations in accuracy and loss values, an indication of overfitting, occurred beyond 175 epochs. The metric that resulted in the best epoch optimization was the validation loss metric. A potential limitation of this method is that a constant epoch value may not produce the best results for any specific run, however, determining sensitivity to this value for each network remains critical.

4.4. Relative importance of accuracy metrics

DC, HD, TPR and FPR: Illustrated by the large area under ROC curve for the threshold value (Fig. 6), the threshold parameter itself is a robust classifier for distinguishing between brain tissue and non-brain tissue as it performs well across a wide range of values. While evaluation by ROC is important for an initial estimation of the threshold value, this only provided a range of threshold values that would be optimal due to the minimal change in mean TPR and FPR for threshold values between 0.60 and 0.85. Similarly, for TDS, the mean TPR did not vary greatly between different values of dataset size, whereas the DC and HD improved until TDS=100. Collectively, the DC and HD were found to be the most useful metrics for determining optimal network parameter values. Additionally, now that most segmentation neural networks are achieving DC>0.97, evaluating other accuracy metrics, such as the HD, would be beneficial to ensure that segmentation shape, the boundaries of the segmentations, are also accurate. From the more recently published networks (Hsu et al., 2021, 2020; Liang et al., 2023), there is heterogeneity in reporting HD, for example in terms of # of voxels as opposed to mm. Thus, standardized reporting of HD (in mm) in future published literature may allow for improved comparisons between models.

4.5. Limitations of the modified 2D U-Net method

Several limitations of our study must be noted. First, training of our network benefitted from access to heterogenous, well-curated data. Both the quality and composition of the data can impact the performance of the neural network (Halevy et al., 2009), and such data may not necessarily be accessible. Additionally, as a supervised machine learning method, our framework requires the creation of manual segmentations for each image in the dataset for training, which can be both resource and time intensive. Therefore, creation of open databases with such data, as planned, will benefit the community. Second, because our network was developed with T_2-weighted images, additional training and optimization may be required to perform WBD with the modified 2D U-Net on images with other types of MRI contrast (e.g., T_1weighting). Furthermore, while the network parameters were assessed, data augmentation strategies were not tested extensively. These augmentations were selected based on translations, rotations, and brightness intensity shifts that are commonly utilized for machine learning methods (Mikołajczyk and Grochowski, 2018) and were represented in our data. More research is needed to determine if more specific data augmentation strategies would be beneficial for improving generalization of the U-Net. Third, our method did not utilize an adaptive learning rate that has been shown to improve the performance of neural network training (Konar et al., 2020). In order to reduce the barrier for entry for optimizing machine learning methods, however, we simplified the learning rate to a single value. Fourth, our method downscaled MR images to 128×128, which could introduce segmentation inaccuracy when upscaled to original MR resolution. A similar issue would arise if source image data are smaller than 128×128 in-plane as the image would need to be interpolated to a higher resolution, which could affect the accuracy of the generated segmentation. Ultimately, the in-plane resolution is inversely-related with computational time, and by down-sampling, the process benefits by a large reduction in computation time. Fifth, there is still an active debate about whether 2D or 3D CNNs would produce more robust segmentations. A benefit of the 3D approach is that it utilizes voxel relationship information, which may provide more useful information for segmentations. However, it is limited by high computation requirements (Woo and Lee, 2021) and a reduction in the number of training data, due to utilizing volumetric input rather than multiple in-plane images from a single volume scan. Our future research will include rigorous comparisons between the 2D and 3D architecture, and examining the applicability of this WBD CNN on other preclinical disease and animal models, MR contrasts, and examining the impact of transfer learning to improve the generalizability of the network to other models. Lastly, for the worst performing segmentations, the U-Net incorrectly classified part of the trigeminal nerves along the ventral aspect of the brain as brain tissue, or the last posterior slices of the cerebellum as non-brain tissue. These findings are consistent with Hsu et al.'s method (Hsu et al., 2020), where the worst performing segmentations were in areas of low signal intensity (e.g., due to signal drop-off relative to head coil position) (Fig. 9). We similarly observed poorer performance independent of ventral distance, per se, and specifically associated with deviation in coil placement. The results of this study indicate that fine-tuning of network training parameters is beneficial for improving its robustness and generalization and suggest that such optimization may be necessary for new and modified architectures, which has been similarly reported for other machine learning algorithms (Nakayama et al., 2012).

5. Conclusion

We assessed our modified 2D U-Net CNN framework as a fully-automated method for WBD for two disparate rodent models of neurological diseases. Critically, this work showed that our framework can enable accommodating significant variability in imaging data resulting from neuropathology and strain differences and provide both generalizability and high accuracy comparable to manual segmentation. We believe that our analysis could be useful as a template for optimizing neural networks for preclinical brain image segmentation, as well as

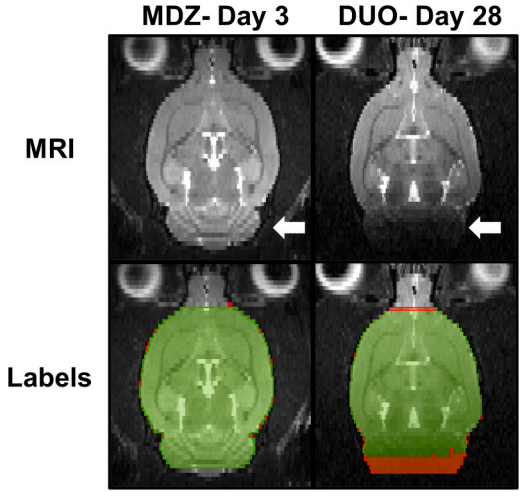


Fig. 9. : Representative images and U-Net generated segmentations of: (**col 1**) MDZ Day 3 scan and (**col 2**) DUO Day 3 scan. Rows top to bottom: anatomical MR image and 2D U-Net-generated segmentation label (matched pixels in green and unmatched pixels in red) overlaid on MR image. The MDZ Day 3 [DC: 0.9900] is the best segmented scan with the U-Net and DUO Day 28 [DC: 0.9356] is the worst segmentation. The white arrows indicate the difference in signal intensity in the cerebellum between an excellent segmentation versus a poor segmentation, where the U-Net performed suboptimally in the low signal region of the image.

provide a robust and accurate whole brain delineation tool applicable to rat neuroimaging data.

CRediT authorship contribution statement

Valerie Porter: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Brad A Hobson:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Brent Foster:** Writing – review & editing, Methodology. **Pamela J Lein:** Writing – review & editing, Validation, Funding acquisition. **Abhijit J Chaudhari:** Writing – review & editing, Validation, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

None of the authors have a conflict of interest with the work presented.

Data availability

Data will be made available on request.

Acknowledgements

The authors gratefully thank Dr. Douglas Rowland, Charles Smith, and Sarah Tam (UC Davis Center for Molecular and Genomic Imaging), and Donald Bruun, Jason Loxterkamp, Eduardo Gonzalez, Jonas Calsbeek, Joan Vu, Thomas Blackmon, and Yi-Hau Tsai (UC Davis School of Veterinary Medicine) for their assistance with animal handling and MR data acquisition. This work was funded in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1 TR001860 and linked award TL1 TR001861, the NIH CounterACT program, through grants U54 NS127758 and U54 NS079202, and the National Institute of Aging, through grants R21 AG064599 and R21 ES026515.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jneumeth.2024.110078.

References

Ashburner, J., 2012. SPM: A history. Neuroimage 62–248, 791. https://doi.org/10.1016/J.NEUROIMAGE.2011.10.025.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044. https://doi.org/10.1016/J.NEUROIMAGE.2010.09.025.

Almeida, A.J.D, Hobson, B.A., Saito, N., Harvey, D., Bruun, D.A., Porter, V.A., Garbow, J. R., Chaudhari, A.J., Lein, P.J., Quantitative T2 mapping-based longitudinal assessment of brain injury and therapeutic rescue in the rat following acute organophosphate intoxication. Manuscript submitted for publication.

Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D., 2022. Medical Image Segmentation Review: The success of U-Net.

Beare, R., Lowekamp, B., Yaniv, Z., 2018. Image segmentation, registration and characterization in R with simpleITK. J. Stat. Softw. 86, 1–35. https://doi.org/10.18637/JSS.V086.I08.

Chang, H.H., Yeh, S.J., Chiang, M.C., Hsieh, S.T., 2023. RU-Net: skull stripping in rat brain MR images after ischemic stroke with rat U-Net. BMC Med. Imaging 23, 1–14. https://doi.org/10.1186/S12880-023-00994-8/FIGURES/11.

Clarke, E., Sherrill-Mix, S., Dawson, C., 2023. ggbeeswarm: Categorical Scatter (Violin Point) Plots.

Cohen, R.M., Rezai-Zadeh, K., Weitz, T.M., Rentsendorj, A., Gate, D., Spivak, I., Bholat, Y., Vasilevko, V., Glabe, C.G., Breunig, J.J., Rakic, P., Davtyan, H., Agadjanyan, M.G., Kepe, V., Barrio, J.R., Bannykh, S., Szekely, C.A., Pechnick, R.N., Town, T., 2013. A transgenic Alzheimer rat with plaques, tau pathology, behavioral impairment, oligomeric Aβ, and frank neuronal loss. J. Neurosci. 33, 6245–6256. https://doi.org/10.1523/JNEUROSCI.3672-12.2013.

Crum, W.R., Camara, O., Hill, D.L.G., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging 25, 1451–1461. https://doi.org/10.1109/TMI.2006.880587.

Cunha, L., Horvath, I., Ferreira, S., Lemos, J., Costa, P., Vieira, D., Veres, D.S., Szigeti, K., Summavielle, T., Máthé, D., Metello, L.F., 2014. Preclinical imaging: an essential ally in modern biosciences. Mol. Diagn. Ther. 18, 153–173. https://doi.org/10.1007/S40291-013-0062-3/TABLES/1.

Davatzikos, C., 2019. Machine learning in neuroimaging: progress and challenges. Neuroimage 197, 652. https://doi.org/10.1016/J.NEUROIMAGE.2018.10.003.

Denic, A., Macura, S.I., Mishra, P., Gamez, J.D., Rodriguez, M., Pirko, I., 2011. MRI in rodent models of brain disorders, 2011 8:1 Neurotherapeutics 8, 3–18. https://doi.org/10.1007/S13311-010-0002-4.

Dhir, A., Bruun, D.A., Guignet, M., Tsai, Y.H., González, E., Calsbeek, J., Vu, J., Saito, N., Tancredi, D.J., Harvey, D.J., Lein, P.J., Rogawski, M.A., 2020. Allopregnanolone and perampanel as adjuncts to midazolam for treating diisopropylfluorophosphate-induced status epilepticus in rats. Ann. N. Y Acad. Sci. 1480, 183–206. https://doi.org/10.1111/NYAS.14479.

Eed, A., Cerdán Cerdá, A., Lerma, J., De Santis, S., 2020. Diffusion-weighted MRI in neurodegenerative and psychiatric animal models: experimental strategies and main outcomes. J. Neurosci. Methods 343, 108814. https://doi.org/10.1016/J.JNEUMETH.2020.108814.

Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B., 2020. Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. IEEE Trans. Med. Imaging 39, 3679–3690. https://doi.org/10.1109/TMI.2020.3002417.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R., 2012. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. Magn. Reson Imaging 30, 1323. https://doi.org/10.1016/J.MRI.2012.05.001.

Feo, R., Giove, F., 2019. Towards an efficient segmentation of small rodents brain: a short critical review. J. Neurosci. Methods. https://doi.org/10.1016/j.jneumeth.2019.05.003.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774. https://doi.org/10.1016/J.NEUROIMAGE.2012.01.021.

Fowler, C.F., Goerzen, D., Devenyi, G.A., Madularu, D., Chakravarty, M.M., Near, J., 2022. Neurochemical and cognitive changes precede structural abnormalities in the TgF344-AD rat model. Brain Commun. 4 https://doi.org/10.1093/BRAINCOMMS/FCAC072.

Gao, Y., Li, Z., Song, C., Li, L., Li, M., Schmall, J., Liu, H., Yuan, J., Wang, Z., Zeng, T., Hu, L., Chen, Q., Zhang, Y., 2021. Automatic rat brain image segmentation using triple cascaded convolutional neural networks in a clinical PET/MR. Phys. Med Biol. 66, 04NT01. https://doi.org/10.1088/1361-6560/ABD2C5.

Goerzen, D., Fowler, C., Devenyi, G.A., Germann, J., Madularu, D., Chakravarty, M.M., Near, J., 2020. An MRI-Derived Neuroanatomical Atlas of the Fischer 344 Rat Brain. Sci. Rep. 10 https://doi.org/10.1038/S41598-020-63965-X.

Halevy, A., Norvig, P., Pereira, F., 2009. The unreasonable effectiveness of data. IEEE Intell. Syst. 24, 8–12. https://doi.org/10.1109/MIS.2009.36.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy, 2020 585:7825 Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hobson, B.A., Sisó, S., Rowland, D.J., Harvey, D.J., Bruun, D.A., Garbow, J.R., Lein, P.J., 2017. From the cover: magnetic resonance imaging reveals progressive brain injury in rats acutely intoxicated with diisopropylfluorophosphate. Toxicol. Sci. 157, 342. https://doi.org/10.1093/TOXSCI/KFX049.

Hsu, L.M., Wang, S., Ranadive, P., Ban, W., Chao, T.H.H., Song, S., Cerri, D.H., Walton, L. R., Broadwater, M.A., Lee, S.H., Shen, D., Shih, Y.Y.I., 2020. Automatic skull stripping of rat and mouse brain mri data using U-Net. Front Neurosci. 14, 935. https://doi.org/10.3389/FNINS.2020.568614/BIBTEX.

Hsu, L.M., Wang, S., Walton, L., Wang, T.W.W., Lee, S.H., Shih, Y.Y.I., 2021. 3D U-Net Improves automatic brain extraction for isotropic rat brain magnetic resonance imaging data. Front Neurosci. 15, 801008 https://doi.org/10.3389/FNINS.2021.801008/BIBTEX.

Hutchins, G.D., Miller, M.A., Soon, V.C., Receveur, T., 2008. Small animal PET imaging. ILAR J. 49, 54–65. https://doi.org/10.1093/ILAR.49.1.54.

Jack, C.R., Bernstein, M.A., Borowski, B.J., Gunter, J.L., Fox, N.C., Thompson, P.M., Schuff, N., Krueger, G., Killiany, R.J., Decarli, C.S., Dale, A.M., Carmichael, O.W., Tosun, D., Weiner, M.W., 2010. Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative. Alzheimer'S. Dement. 6, 212–220. https://doi.org/10.1016/J.JALZ.2010.03.004.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62, 782–790. https://doi.org/10.1016/J.NEUROIMAGE.2011.09.015.

Judenhofer, M.S., Cherry, S.R., 2013. Applications for preclinical PET/MRI. Semin Nucl. Med 43, 19–29. https://doi.org/10.1053/J.SEMNUCLMED.2012.08.004.

Karimi, D., Salcudean, S.E., 2020. Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Trans. Med. Imaging 39, 499–513. https://doi.org/10.1109/TMI.2019.2930068.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med Imaging 29, 196–205. https://doi.org/10.1109/TMI.2009.2035616.

Konar, J., Khandelwal, P., Tripathi, R., 2020. Comparison of Various Learning Rate Scheduling Techniques on Convolutional Neural Network. 2020 IEEE International

Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2020. https://doi.org/10.1109/SCEECS48394.2020.94.
Lenchik, L., Heacock, L., Weaver, A.A., Boutin, R.D., Cook, T.S., Itri, J., Filippi, C.G., Gullapalli, R.P., Lee, J., Zagurovskaya, M., Retson, T., Godwin, K., Nicholson, J., Narayana, P.A., 2019. Automated segmentation of tissues using CT and MRI: a systematic review. Acad. Radio. 26, 1695–1706. https://doi.org/10.1016/j.acra.2019.07.006.
Liang, S., Yin, X., Huang, L., Huang, J., Yang, J., Wang, X., Peng, L., Zhang, Y., Li, Z., Nie, B., Tao, J., 2023. Automatic brain extraction for rat magnetic resonance imaging data using U2-Net. Phys. Med. Biol. 68, 205006 https://doi.org/10.1088/1361-6560/ACF641.
Liu, Y., Unsal, H.S., Tao, Y., Zhang, N., 2020. Automatic brain extraction for rodent MRI images. Neuroinformatics 18, 395–406. https://doi.org/10.1007/S12021-020-09453-Z/FIGURES/6.
Mikołajczyk, A., Grochowski, M., 2018. Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop, IIPhDW 2018 117–122. https://doi.org/10.1109/IIPHDW.2018.8388338.
Müller, D., Soto-Rey, I., Kramer, F., 2022. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res. Notes 15, 1–8. https://doi.org/10.1186/S13104-022-06096-Y/FIGURES/2.
Nakayama, H., Yun, Y., Uno, Y., 2012. Parameter tuning of large scale support vector machines using ensemble learning with applications to imbalanced data sets. Conf Proc IEEE Int Conf Syst Man Cybern 2815–2820. https://doi.org/10.1109/ICSMC.2012.6378175.
Ni, R., Cereda, C., Gagliardi, S., 2021. Magnetic resonance imaging in animal models of Alzheimer's disease amyloidosis, 2021, Vol. 22, 12768 Int. J. Mol. Sci. 22, 12768. https://doi.org/10.3390/IJMS222312768.
Paxinos, George, Watson, Charles, 2007. The Rat Brain in Stereotaxic Coordinates, Elsevier Inc. Elsevier Science.
Pontes-Filho, S., Dahl, A.G., Nichele, S., Mello, G.B.M. e, 2022. A deep learning-based tool for automatic brain extraction from functional magnetic resonance images of rodents. Lect. Notes Netw. Syst. 296, 549–558. https://doi.org/10.1007/978-3-030-82199-9_36/FIGURES/2.
Prescott, J.W., 2013. Quantitative imaging biomarkers: the application of advanced image processing and analysis to clinical and preclinical decision making. J. Digit Imaging 26, 97–108. https://doi.org/10.1007/S10278-012-9465-7/TABLES/4.
R Core Team, 2022. A Language and Environment for Statistical Computing.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. 9351, 234–241 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
Saré, R.M., Cooke, S.K., Krych, L., Zerfas, P.M., Cohen, R.M., Smith, C.B., 2020. Behavioral phenotype in the TgF344-AD Rat Model of Alzheimer's disease. Front Neurosci. 14, 601. https://doi.org/10.3389/FNINS.2020.00601/BIBTEX.
Shattuck, D.W., Leahy, R.M., 2002. Brainsuite: An automated cortical surface identification tool. Med. Image Anal. 6, 129–142. https://doi.org/10.1016/S1361-8415(02)00054-3.
Siso, S., Hobson, B.A., Harvey, D.J., Bruun, D.A., Rowland, D.J., Garbow, J.R., Lein, P.J., 2017. Spatiotemporal progression and remission of lesions in the rat brain following acute intoxication with diisopropylfluorophosphate. Toxicol. Sci. 157, 330–341. https://doi.org/10.1093/toxsci/kfx048.
Slicer Wiki contributors, 2020. Documentation/Nightly/Modules/SegmentComparison [WWW Document]. Slicer Wiki. URL https://www.slicer.org/w/index.php?title=Documentation/Nightly/Modules/SegmentComparison&oldid=63091 (accessed 6.14.23).
Supasai, S., González, E.A., Rowland, D.J., Hobson, B., Bruun, D.A., Guignet, M.A., Soares, S., Singh, V., Wulff, H., Saito, N., Harvey, D.J., Lein, P.J., 2020. Acute administration of diazepam or midazolam minimally alters long-term neuropathological effects in the rat brain following acute intoxication with diisopropylfluorophosphate. Eur. J. Pharm. 886, 173538 https://doi.org/10.1016/J.EJPHAR.2020.173538.
Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15, 1–28. https://doi.org/10.1186/S12880-015-0068-X/TABLES/5.
Takase, T., Oyama, S., Kurihara, M., 2018. Effective neural network training with adaptive learning rate based on training loss. Neural Netw. 101, 68–78. https://doi.org/10.1016/J.NEUNET.2018.01.016.
The MONAI Consortium, 2020. https://doi.org/10.5281/zenodo.4323059.
van Oostveen, W.M., de Lange, E.C.M., 2021. Imaging Techniques in Alzheimer's disease: a review of applications in early diagnosis and longitudinal monitoring, 2021, Vol. 22, 2110 Int. J. Mol. Sci. 22, 2110. https://doi.org/10.3390/IJMS22042110.
Weiss, J., Hoffmann, U., Aerts, H.J.W.L., 2020. Artificial intelligence-derived imaging biomarkers to improve population health. Lancet Digit Health 2, e154–e155. https://doi.org/10.1016/S2589-7500(20)30061-3.
Welniak-Kaminska, M., Fiedorowicz, M., Orzel, J., Bogorodzki, P., Modlinska, K., Stryjek, R., Chrzanowska, A., Pisula, W., Grieb, P., 2019. Volumes of brain structures in captive wild-type and laboratory rats: 7T magnetic resonance in vivo automatic atlas-based study. PLoS One 14, e0215348. https://doi.org/10.1371/JOURNAL.PONE.0215348.
Wickham, Hadley, 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
Wolf, G., Abolmaali, N., 2013. Preclinical molecular imaging using PET and MRI. Recent Results Cancer Res. 187, 257–310. https://doi.org/10.1007/978-3-642-10853-2_9/FIGURES/9.
Woo, B., Lee, M., 2021. Comparison of tissue segmentation performance between 2D U-Net and 3D U-Net on brain MR Images. 2021 International Conference on Electronics, Information, and Communication, ICEIC 2021. https://doi.org/10.1109/ICEIC51217.2021.9369797.
Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Kaus, M.R., Haker, S.J., Wells, W.M., Jolesz, F.A., Kikinis, R., 2004. Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports. Acad. Radio. 11, 178. https://doi.org/10.1016/S1076-6332(03)00671-8.