PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND

Dissertations in
Health Sciences

RICCARDO DE FEO

CONVOLUTIONAL NEURAL NETWORKS FOR THE SEGMENTATION OF SMALL RODENT BRAIN MRI

CONVOLUTIONAL NEURAL NETWORKS FOR THE SEGMENTATION OF SMALL RODENT BRAIN MRI

Riccardo De Feo

CONVOLUTIONAL NEURAL NETWORKS FOR THE SEGMENTATION OF SMALL RODENT BRAIN MRI

To be presented by permission of the Faculty of Health Sciences, University of Eastern Finland and the Faculty of Pharmacy and Medicine, La Sapienza University of Rome, for public examination in La Sapienza, Rome on February 9th, 2023, at 2:30 p.m.

Publications of the University of Eastern Finland
Dissertations in Health Sciences
No 729

A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland, Kuopio
2023

Distributor:
University of Eastern Finland
Kuopio Campus Library
P.O.Box 1627
FI-70211 Kuopio, Finland
http://www.uef.fi/kirjasto

PunaMusta Oy
Joensuu, 2023

ISBN: 978-952-61-4785-7 (Print)
ISBN: 978-952-61-4786-4 (PDF)
ISSNL: 1798-5706
ISSN: 1798-5706
ISSN: 1798-5714 (PDF)

Author's address: A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
KUOPIO
FINLAND

Doctoral programme: Doctoral Programme in Molecular Medicine
Phd in Morphogenesis and Tissue Engineering, Cycle 34th
Coordinator: Prof. Antonio Musarò

Supervisors: Professor Jussi Tohka, Ph.D
A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
KUOPIO
FINLAND

Professor Federico Giove, Ph.D.
La Sapienza University of Rome
Rome
Italy

Reviewers: Roy Snehashis, Ph.D.
National Institute of Mental Health (NIMH)
Bethesda, MD
USA

Andrea Tangherloni, Ph.D.
University of Bergamo
Department of Human and Social Science
Bergamo
Italy

Examination Committee: Professor Massimiliano Papi, Ph.D.
Universitá Cattolica del S. Cuore
Italy
Professor Graziella Messina, Ph.D.
Milan University
Italy
Professor Emanuela Mercenaro, Ph.D.
University of Genoa
Italy
Professor Simon Fristed Eskildsen, Ph.D.
Aarhus University
Denmark

De Feo Riccardo
Convolutional neural networks for the segmentation of small rodent brain MRI.
Kuopio: University of Eastern Finland
Publications of the University of Eastern Finland
Dissertations in Health Sciences 729, 2023, 153 p.
ISBN: 978-952-61-4785-7 (Print)
ISSNL: 1798-5706
ISSN: 1798-5706
ISBN: 978-952-61-4786-4 (PDF)
ISSN: 1798-5714 (PDF)

ABSTRACT

Image segmentation is a common step in the analysis of preclinical brain MRI, often performed manually. This is a time-consuming procedure subject to inter- and intra- rater variability. A possible alternative is the use of automated, registration-based segmentation, which suffers from a bias owed to the limited capacity of registration to adapt to pathological conditions such as Traumatic Brain Injury (TBI).

In this work a novel method is developed for the segmentation of small rodent brain MRI based on Convolutional Neural Networks (CNNs). The experiments here presented show how CNNs provide a fast, robust and accurate alternative to both manual and registration-based methods. This is demonstrated by accurately segmenting three large datasets of MRI scans of healthy and Huntington disease model mice, as well as TBI rats. MU-Net and MU-Net-R, the CCNs here presented, achieve human-level accuracy while eliminating intra-rater variability, alleviating the biases of registration-based segmentation, and with an inference time of less than one second per scan.

Using these segmentation masks I designed a geometric construction to extract 39 parameters describing the position and orientation of the hippocampus, and later used them to classify epileptic *vs.* non-epileptic rats with a balanced accuracy of 0.80, five months after TBI. This clinically transferable geometric approach detects subjects at high-risk of post-traumatic epilepsy, paving the way towards subject stratification for antiepileptogenesis studies.

Medical Subject Headings: segmentation, rat, mouse, Convolutional Neural Networks, epilepsy, magnetic resonance imaging, deep learning, machine learning.

Yleinen suomalainen ontologia: Segmentaatio, rotta, hiiri, konvoluutioneuroverkot, epilepsia, magneettiresonanssikuvantaminen, syväoppiminen, koneoppiminen

ACKNOWLEDGEMENTS

Many thanks to everybody who joined me on this journey!

To Juan Miguel Valverde, Elina Hämäläinen, Eppu Manninen, Vandad Imani, Andrea Behanova, Ali Abdollahzadeh, Artem Shatillo, Alejandra Sierra; to Michele Allori for his help with 3D modeling; to Karthik Chary, Riikka Immonen, Pedro Andrade and Xavier Ekolle Ndode-Ekane for generating some of the data that made its way to this work; to professors Olli Gröhn and Asla Pitkänen; and to my supervisors Jussi Tohka and Federico Giove.

This work is a joint doctoral thesis between La Sapienza University of Rome, Italy, and the University of Eastern Finland, Finland.

Vantaa, January 2023

Riccardo De Feo

‘Thou shalt not make a machine in the likeness of a human mind’

Orange Catholic Bible

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications. The publications were adapted with the permission of the copyright owners:

- Riccardo De Feo and Federico Giove: **Towards an efficient segmentation of small rodents brain: a short critical review**. In: *Journal of neuroscience methods* 323 (2019), pp. 82–89 (De Feo & Giove 2019)
- Riccardo De Feo, Artem Shatillo, Alejandra Sierra, Juan Miguel Valverde, Olli Gröhn, Federico Giove, and Jussi Tohka: **Automated joint skull-stripping and segmentation with Multi-Task U-Net in large mouse brain MRI databases.**. In: *NeuroImage*, 2021, p. 117734. (De Feo *et al.* 2021)
- Riccardo De Feo, Elina Hämäläinen, Eppu Manninen, Riikka Immonen, Juan Miguel Valverde, Xavier Ekolle Ndode-Ekane, Olli Gröhn, Asla Pitkanen, and Jussi Tohka: **Convolutional neural networks enable robust automatic segmentation of the rat hippocampus in MRI after traumatic brain injury**. In: *Frontiers in Neurology*, 13 (2022). issn: 1664-2295. (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)
- Riccardo De Feo, Eppu Manninen, Karthik Chary, Elina Hämäläinen, Riikka Immonen, Pedro Andrade, Xavier Ekolle Ndode-Ekane, Olli Gröhn Asla Pitkänen, Jussi Tohka: **Hippocampal position and orientation as prognostic biomarkers for posttraumatic epileptogenesis: An experimental study in a rat lateral fluid percussion model**. In: *Epilepsia, 2022* (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

The code developed in this work has been released under MIT license in the following repositories:

- `https://github.com/Hierakonpolis/MU-Net`
- `https://github.com/Hierakonpolis/MU-Net-R`
- `https://github.com/Hierakonpolis/RatHippocampusGeometry`

AUTHOR'S CONTRIBUTION

CONTRIBUTION STATEMENT

For each of the co-authored publications contributing to this work, the authors contributed as follows:

Towards an efficient segmentation of small rodents brain: A short critical review

RDF made the bibliographic search and drafted the paper. FG ide- ated and coordinated the research and revised the paper.

Automated joint skull-stripping and segmentation with Multi-Task U-Net in large mouse brain MRI databases

RDF: Methodology, Software, Formal analysis, Writ- ing - original draft. ASh: Data curation. ASi: Methodology, Formal analysis. JV: Methodology. OG: Conceptualization. FG: Conceptualization. JT: Conceptualization, Software, Writing - original draft.

Convolutional Neural Networks Enable Robust Automatic Segmentation of the Rat Hippocampus in MRI After Traumatic Brain Injury

RD: methodology, software, formal analysis, investigation, and writing—original draft. EH: data curation and methodology. EM: software and investigation. RI: investigation. JV: methodology, writing—review, and editing. XN-E: data curation, formal analysis, and writing—original draft. OG: conceptualization. AP and JT: conceptualization, methodology, and writing—original draft. All authors contributed to the article and approved the submitted version.

Hippocampal position and orientation as prognostic biomarkers for posttraumatic epileptogenesis: An experimental study in a rat lateral fluid percussion model

RD: methodology, software, formal analysis, investigation, and writing—original draft. EM: software and investigation. KC: investigation, data curation. EH: data curation. RI: investigation. PA: investigation, data curation. XN-E: investigation, data curation. OG: conceptualization. AP and JT: conceptualization, methodology, and writing—original draft.

CONTENTS

Acronyms

CNN	Convolutional Neural Network
CS	Compactness Score
CSF	Cerebrospinal Flluid
CR	Charles River
DNN	Deep Neural Network
FPI	Fluid Percussion Injury
GM	Gray Matter
GPU	Graphic Processing Unit
GT	Ground Truth
HD	Hausdorff Distance
HD95	95th percentile of the Hausdorff Distance
HT	Huntington
MRI	Magnetic Resonance Imaging
PCNN	Pulse Coupled Neural Networks
PTE	Post-Trauamtic Epilepsy
ReLU	Rectified Linear Unit
ROI	Region of Interest
TBI	Traumatic Brain Injury
TPU	Tensor Processing Unit
UEF	University of Eastern Finland
VS	Volume Similarity
WM	White Matter
WT	Wild Type

1 INTRODUCTION

One of the most common operations during the analysis of preclinical brain Magnetic Resonance Imaging (MRI) is the identification of those specific areas of the scan that we are interested in analyzing. This apparently simple task can be tackled with a variety of methods, each presenting a set of strengths and shortcomings. The most used methods are manual segmentation, where one or more human raters are asked to manually label each region in the scan slice by slice, and registration-based segmentation, where one or more atlases are automatically translated and deformed to be adapted to the target scan. The results emerging from the literature review presented in the first part of this dissertation will show the reader that manual segmentation is not as straightforward as we might assume, presenting significant inter- and intra-rater variability based on the difficulty of the task, and requiring a considerable amount of time for large datasets. As an alternative to manual segmentation, chapter 2 will further introduce the state of the art on registration-based methods, discussing single- and multi- altas segmentation in detail, as well as discussing alternative methods for tissue segmentation.

Registration-based methods rely on the precise alignment of specific atlases with the target volume, and can still require a significant amount of time. Convolutional Neural Networks (CNNs) represent a promising alternative approach to registration-based segmentation. In chapter 3 I introduce the general ideas that we will need to understand the work presented in the following chapters, discussing the basic elements involved in the design and application of a CNN in MRI.

Next are outlined the materials and methods employed in the experiments presented in this work, detailing the approach developed for anatomical segmentation, the data utilized to test its validity, and the state-of-the-art registration methods used to provide a baseline comparison. Chapter 4 will detail the MRI datasets here utilized, including a very large dataset of 1,782 samples from both healthy and Huntington disease model mice, and two more datasets for the study of epileptogeneis as a results of Traumatic Brain Injury (TBI). Chapter 5 describes Multi-task U-Net (MU-Net) and MU-Net-Rat (MU-Net-R), the CNNs designed and trained for the segmentation of preclinical small animals brain MRI, detailing which architectural features were explored during their design and the full procedure for training and inference. Conversely, chapter 6 details the implementation details of the registration-based methods used as a comparison to exemplify the state of the art for automated rat and mouse brain segmentation.

To demonstrate and build over the segmentation maps generated MU-Net-

R, chapter 7 will introduce the reader to a geometric construction developed to extract a set of anatomical parameters from the segmentation masks, and use them to classify epileptogenic *vs.* non-epileptogenic animals.

We move on to the experiments exploring the performance MU-Net and MU-Net-R across the datasets described previously, detailing in chapter 8 the experiments regarding MU-Net, its development and evaluation, and in chapter 9 the training of MU-Net-R on a small labeled dataset, to later apply it on TBI and sham-surgery rats. Using these segmentation masks, Chapter 10 explores the use of the geometry of the hippocampus as a biomarker for post-traumatic epilepsy, developing a random forest model capable of discriminating epileptogenic *vs.* non-epiletogenic rats with a balanced accuracy of about 0.8, 5 months after TBI. Chapter 11 finally discusses the findings presented in the previous chapters and the existing limitations of this work.

The experiments in this work are presented by reproducing text and several images from my recent publications, reorganized and integrated with additional context and scientific background. The results suggest that CNNs would display clear advantages compared to registration-based methods, and could likely replace them in a number of applications in the near future.

2 AUTOMATED SEGMENTATION OF SMALL RODENTS' BRAIN MRI

MRI is a medical imaging technique used to acquire three-dimensional images of biological tissues in both the clinical and preclinical setting. It does so in a non-invasive way, without employing any ionizing radiation, and providing a variety of different imaging contrasts to capture different aspects of the tissues under investigation. Thanks to these characteristics a growing number of preclinical studies are based on mouse and rat MRI, both ex vivo and *in vivo*. Common techniques involving small rodents include functional magnetic resonance imaging (Jonckers *et al.* 2011), diffusion tensor imaging (Le Bihan *et al.* 2001), and structural imaging contrasts such as T_2, used for example in voxel based morphometry or cortical thickness studies (Pagani *et al.* 2016, Nie *et al.* 2014).

The segmentation of the acquired volumes in different classes is a common step in MRI pipelines. This can refer to the voxel-wise labeling of specific anatomical Regions of Interest (ROIs), lesions or anonalies, or biological tissues. An expert human rater can perform this step by labeling each MRI volume slice-by-slice, perhaps with the aid of an anatomical atlas, but this time consuming approach is often impractical. The time required to perform it increases both with resolution and dataset size. Furthermore, manual segmentations can display a large inter-rater variability, with volume overlaps usually varying between 80% and 95%, but depending on the specific regions the overlap can be as low as 70% (Ali *et al.* 2005).

For over 30 years many algorithms have been developed to accelerate and standardize the process of MRI segmentation resulting in the variety of techniques that make up the current and still evolving state of the art. Most of these algorithms focused on human MRI, and it can be less than obvious which algorithms would better transfer to small rodents: while small rodent MRI scans often offer lower contrast and less defined structures compared to human subjects, they also present less anatomical variability (Bai *et al.* 2012). Automated segmentation procedures can also be used to further enhance the registration algorithms themselves, as in the case of DARTEL from the popular SPM suite (Ashburner & Friston 2005).

The purpose of this chapter is to present an overview of the state of the art of brain MRI segmentation for the brain of small rodents. After a brief introduction to common pre-processing steps (bias field correction and skull stripping) we will discuss atlas-based and statistical classification methods, and their implementation in some of the most used and freely available toolsets for brain MRI research. While the former implements a registration based strategy from one or more labeled "atlases", the latter is based on expectation-

maximization algorithms, and some pipelines blur the line between the two. Newly introduced methods, based on CNNs (LeCun *et al.* 2015), will be discussed in greater depth in the following chapters.

2.1 QUANTITATIVE EVALUATION OF IMAGE SEGMENTATION

Before we can discuss the different approaches to brain segmentation, it is necessary to introduce the criteria by which these are evaluated. A manual segmentation generally provides the Ground Truth (GT), and the similarity between that and the algorithm's output provides a metric for the quality of the algorithm. To evaluate this similarity we can use a variety of different complementary metrics, introduced as follows.

Dice and Jaccard The most common metrics in literature are the Sørensen-Dice coefficient or Dice score D (Dice 1945)) and the Jaccard index J (Jaccard 1912)), defined as follows:

$$D = \frac{2\,|X \cap Y|}{|X| + |Y|}; \quad J = \frac{|X \cap Y|}{|X \cup Y|}$$

The two metrics can be used to quantify the overlap between any two segmentation masks X and Y, defined as sets of voxel- or pixel- wise labels associated to the labeled image. These definitions result in two dimensionless quantities varying between 0 (no overlap) and 1 (perfect overlap). While here we will only be using the Dice score, the two are entirely interchangeable and contain the same information. It is possible to calculate one from the other as follows:

$$D = \frac{2J}{1+J}; \quad J = \frac{D}{2-D}$$

Hausdorff 95 The Hausdorff Distance (HD) (Huttenlocher *et al.* 1993) refers to the magnitude of the largest segmentation error of the prediction when compared to the GT:

$$HD(Y, X) = \max(h(Y, X), h(X, Y)) \tag{1}$$

where

$$h(Y, X) = \max_{y \in Y} \min_{x \in X} |y - x|\,. \tag{2}$$

To obtain a more stable measure and reduce the impact of outliers we use the 95th percentile of the Hausdorff Distance (HD95). Unlike the Dice coefficient, this is a dimensional measure, typically indicated in millimeters.

Volume similarity We can measure the Volume Similarity (VS) between prediction and GT, following the definition provided by (Taha & Hanbury 2015). Unlike the Dice score, VS does not depend on the overlap between the two ROIs, and only depends on their volumes. This is of particular interest in exclusively volumetric studies, where the measures are directly impacted only by the volume of ROIs. VS is defined as:

$$VS = 1 - \frac{||X| - |Y||}{|X| + |Y|}. \tag{3}$$

Precision and recall Segmentation can be understood as voxel- or pixel-wise classification, and as such we can apply typical classification metrics. Precision P and recall R evaluate respectively the ratio between true positives and the total number of positive predictions, and true positives and ground truth size. Increasing the number of false positives reduces the precision, and increasing the number of false negatives reduces recall. Both metrics vary between 0 and 1, and are defined as:

$$P = \frac{|Y_t \cap Y|}{|Y|}, \quad R = \frac{|Y_t \cap Y|}{|Y_t|}. \tag{4}$$

Unlike D, VS and HD_{95}, these are not symmetric measures that can be applied to any two segmentation maps, but rely on having defined a ground truth segmentation mask Y_t and a proposed segmentation mask Y. This means for example that they are not indicated for the comparison of two different human raters.

Compactness score Compactness C is defined as the ratio between the surface area and the volume of a ROI (Bribiesca 2008):

$$C = area^{1.5}/volume. \tag{5}$$

Compactness can be used directly (Valverde *et al.* 2020), as common artifacts of machine-generated segmentation often translate to a larger C. However, smaller is not always better: the minimum theoretical value for C is always $6\sqrt{\pi}$. A direct application of compactness under a "smaller is better" assumption would be comparing the segmented ROI with a sphere. To compare the measured C with the ground truth compactness C_t calculated on the manually-

segmented ROI, let us define the Compactness Score (CS) as:

$$CS = 1 - 2\frac{C - C_t}{C + C_t}, \tag{6}$$

where $CS = 1$ indicates an identical compactness, and lower values indicate the two regions display a different ratio between surface and volume.

While the Dice score appears to be the most utilized metric in literature, the others may or may not be indicated. Furthermore, a direct comparison between the average scores across many regions as measured by different studies can be misleading. For example, it is much easier to obtain high overlaps with large, bulky regions of interest (ROIs), while smaller and elongated regions are harder to successfully co-register. The most significant results when comparing different algorithms are the ones operating the same tasks, registering the same regions, tested on the same dataset.

2.2 PRE-PROCESSING

2.2.1 Intensity correction

The performance of both automated skull stripping and segmentation algorithms can often be enhanced by an intensity non-uniformity pre-processing step (Sled *et al.* 1998). The imperfections in the uniformity of the RF excitation field and receiver coil sensitivity profile often result in an artifact consisting in a smooth variation of the signal even in homogeneous tissues, called the bias field. While these effects in practice do not have a strong enough visual impact to impair manual segmentation, they can hamper the performance of automatic skull stripping and segmentation algorithms. Several algorithms have been developed to correct this bias, such as the N3 (Sled *et al.* 1998) or N4 (Tustison *et al.* 2010) algorithms. It is recommended to implement a bias correction step before skull-stripping and segmenting the volumes.

2.2.2 Skull-stripping

Many analysis pipelines include a separate skull-stripping step, designed to discriminate brain and non-brain tissues. Several fully or semi-automated procedures have been developed for this purpose in the specific case of rodent MRI volumes, and given the specificity of some of these methods they need to be discussed separately.

Pulse Coupled Neural Networkss (PCNNs) are a biomimetic neural network based on the visual cortex of cats (Zhan *et al.* 2017). In their original implementation PCNNs operated on individual 2D slices (Murugavel &

Sullivan Jr 2009) but the algorithm has later been expanded to natively handle 3D data (Chou *et al.* 2011). 3D-PCNN remain competitive to this day, in some cases outperforming more recent methods like RATS (Oguz *et al.* 2014), in particular for skull-stripping in the presence of traumatic brain injuries (Roy, Knutsen, Korotcov, Bosomtwi, Dardzinski, Butman & Pham 2018). 3D-PCNN has been tested over the years on multiple datasets, with Dice scores generally above 0.9, up to 0,97 in ideal Signal-to-Noise Ratio conditions (Chou *et al.* 2011, Oguz *et al.* 2014, Roy, Knutsen, Korotcov, Bosomtwi, Dardzinski, Butman & Pham 2018, Li *et al.* 2013). Comparable scores have also been reported using methods based on constraint level sets (Uberti *et al.* 2009).

The RATS method, on the other hand, performs much better on T1 MRI volumes (Oguz *et al.* 2014), while most of the algorithms for rodent segmentation and brain extraction focus on T2 volumes, as it provides a better contrast for small rodents.

While deformable surface methods developed specifically for human brain extraction can be inaccurate when applied directly to rodent MRI, they can be effectively adapted. In recent years Li *et al.* (2013) adapted the BET algorithm (Smith 2002*a*) to the rodent brain, both by improving on the algorithm itself and through a more appropriate choice of the shape prior. Both this implementation and the AFNI `3dskullstrip -rat` function perform quite effectively, with Dice scores slightly above the 3D-PCNN method (Li *et al.* 2013, Roy, Knutsen, Korotcov, Bosomtwi, Dardzinski, Butman & Pham 2018).

Recently Roy, Knutsen, Korotcov, Bosomtwi, Dardzinski, Butman & Pham (2018) applied Deep Neural Networks (DNNs) (LeCun *et al.* 2015) to the skull-stripping of both human and mice subjects with remarkable results, highlighting the robustness of these algorithms in the presence of traumatic brain injuries, with Dice scores around 0.95.

Semi-automated methods for skull-stripping are less time-efficient, but they can yield improved results. A common procedure, as outlined by Delora *et al.* (2016) and Pagani *et al.* (2016), is based on registering all mouse brains to a study-specific template, to be segmented manually, and later propagate the brain mask to the individual volumes. While not fully automated this method yields excellent results, benefiting from being tailored to the specific experimental parameters of the study and the specific population, resulting in a reported Dice score of 0.96. In general, methods based on registration and single or multi-atlas segmentation are also common, implementing the same strategies that will be discussed in the following sections (Leung *et al.* 2011).

2.3 SEGMENTATION

The task of brain region segmentation aims to identify a set of predefined regions in the rodent's brain, and relies on two key components to classify the different regions: a registration algorithm and one or more atlases, with the overall quality of the segmentation depending on both. The atlas or atlases contain the prior information on the tissue classes, in the form of labeled MRI volumes or templates, while the registration algorithm adapts the atlases to the volume to be segmented. The final output of the procedure is a new volume in which the labels and the original data to be segmented are co-registered in the same space.

In this section we will discuss in turn these key aspects of brain region segmentation and their implementation. Further on we will turn our attention to clustering algorithms and the different task of tissue segmentation. An overview of the general outline of a segmentation pipelines is given by the diagram in Figure 1.

2.3.1 Single atlas segmentation

2.3.1.1 Atlases

The prior knowledge required to semantically segment different brain structures can be provided in the form of one or more anatomic atlases, composed of a minimum of two volumes: the original MR data, or a template, and an associated set of voxel by voxel labels. These can be registered to the target volumes or, more often, a study-specific template, from which the labeling can be propagated to the individual subjects. Several atlases also feature probabilistic maps, in which each voxel represents the probability of belonging to a particular class, which can be seen as an early form of multi-atlas segmentation. Over the years, many atlases have been developed both for rats (e. g. Kjonigsen *et al.* (2015), Schweinhardt *et al.* (2003), Papp *et al.* (2014), Veraart *et al.* (2011), Rumple *et al.* (2013), Johnson *et al.* (2012), Schwarz *et al.* (2006), Hjornevik *et al.* (2007), Liang *et al.* (2017)) and mouses (e. g. Hjornevik *et al.* (2007), Dorr *et al.* (2008), Ma *et al.* (2005), Aggarwal *et al.* (2009), Johnson *et al.* (2010), Kovačević *et al.* (2004), Chuang *et al.* (2011)). Atlases can differ in many ways: template building strategy, contrasts, resolution, number of subjects, breed and age of the subjects, the use of *ex vivo* or *in vivo* data, coordinate reference, and segmentation classes.

A simple template building strategy is to choose one random subject and use a deformable registration algorithm to register this volume to every other brain, compute the inverse transforms and average them, to obtain

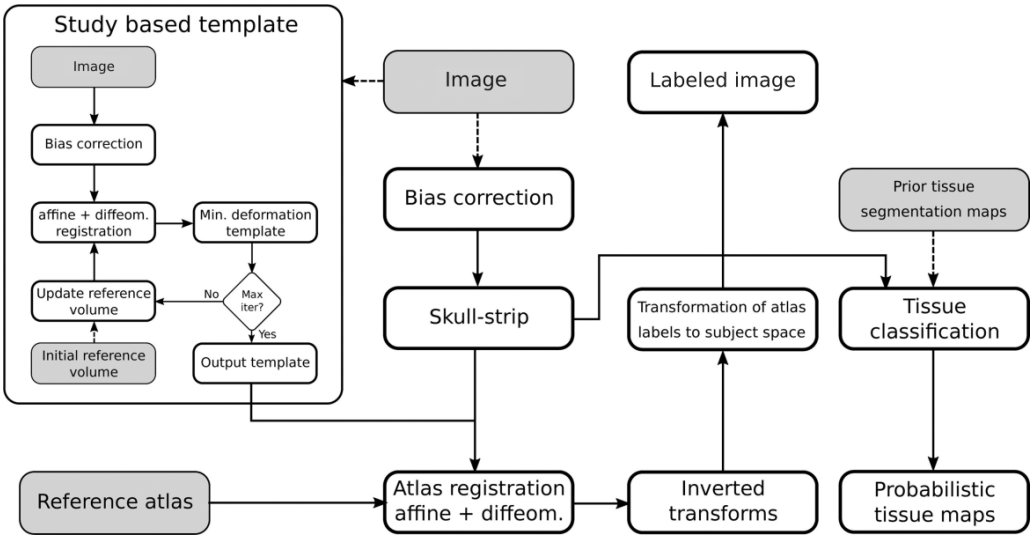

Figure 1. General outline of segmentation pipelines including multiple tasks: skull stripping, region segmentation and tissue classification. Dashed lines indicate alternative or optional paths: for example, not all tissue classification algorithms require prior maps. This general outline can be modified and improved upon: for example one could implement a multi-atlas segmentation scheme, or a more complex template building strategy. Entire branches can be omitted, e. g. the user might be only interested in tissue classification. Depending on the toolset employed, some of these steps might be automated and transparent to the user. Figure reproduced under CC license (De Feo & Giove 2019).

a first average image. This process is then reiterated several times using the average as the new registration target (Kovačević *et al.* 2004). Using a template generated from multiple subjects allows us to avoid errors due to imaging artifacts and individual variability, which might also be a consequence of excision in *ex vivo* brains. An alternative method is creating a minimum deformation template, building the average brain that minimizes the required deformation to be adapted to the entire database of individual subjects (Veraart *et al.* 2011, Johnson *et al.* 2012, Kochunov *et al.* 2001, Ma *et al.* 2005). Segmentations based on these templates can easily yield Dice scores above 0.9, with the exception of small or elongated structures, which resent from slight registration errors to a larger degree. (Ma *et al.* 2005). An effective template building strategy, based on the production of an initial reference through affine transformations and its refinement with a minimum nonlinear deformation approach, can be streamlined with the use of the `buildtemplateparallel`

script available in the ANTs toolset (Avants *et al.* 2010).

A probabilistic atlas does not emerge from a direct segmentation of the template. Every volume used to build the template is manually segmented, and its final segmentation emerges from the statistics of the labels as they are propagated to the template.

ex vivo atlases offer higher resolution and eliminate motion artifacts due to the breathing of the subject, however the brain itself is altered in the process. Aggarwal *et al.* (2009) observed a shrinkage in *ex vivo* brains from 1% up to 3.8 ± 0.6 %, depending on the axis, and highlighted the problem of different structures shrinking by a different amount, further influenced by the choice of reagents, concentrations and fixation methods. To address this problem an *ex vivo* template can be manually segmented and then non-linearly mapped to an *in vivo* population average (Aggarwal *et al.* 2009, Veraart *et al.* 2011).

The template itself can be segmented manually, or with the aid of an histological atlas. The popular Paxinos & Watson and Paxinos & Franklin atlases (Paxinos & Franklin 2004, Paxinos & Watson 1986), in their multiple versions, have been employed for this purpose by many authors, thus allowing for registration in stereotaxic coordinates.

Several attempts have been made at integrating the diverse landscape of available atlases and data for small rodents (Hawrylycz *et al.* 2011), including information such as function or gene and protein expression. The Waxholm space (Johnson *et al.* 2010), designed explicitly for MRI, CT and PET mouse brain imaging, easy to convert to a stereotaxic reference, is probably one of the most successful, but as of yet the general landscape of atlases remains quite varied.

2.3.1.2 Registration

Given one atlas, the atlas and the volumes to be segmented are registered to the same space. The quality of the labeling will be conditioned by the quality of the registration itself, and many strategies have been devised to improve this process. Image or volume registration is formulated as an optimization problem. Registration algorithms aim to find the optimal parameters for the transformation T, from the moving volume I_M to the target volume I_T, minimizing a cost function $C(I_M, I_T)$:

$$\hat{T} = \text{argmin}_{T \in S_T} C(I_M \circ T, I_T)$$

Where $I_M \circ T$ represents the I_M volume transformed by T, and S_T the space of allowed transformations T (consistently with the notation of

Table 1. CC: cross correlation; CCH: histogram-based correlation coefficient; CR: correlation ratio; GC: global correlation; GD: gradient difference; ICP: iterative closest point (Euclidean); JHCT: Jensen-Havrda-Charvet-Tsallis; LD: label difference; MS: mean squares; MI: mutual information; NMI: normalized mutual onformation; NC: normalized correlation; ND: normalized difference; PSE: point set expectation; SSC: stochastic sign change; VR: variance ratio. ICP, JHCT and PSE are designed for point-set registration.

	FLIRT	ANTs	mni_autoreg	Elastix
Linear Transformation Metrics	CR, LD, MS, MI, NC, NMI	CC, CCH, GC, GD, ICP, JHCT, MS, MI, PSE	CC, MI, ND, SSC, VR	MS, MI, NC, NMI

Jenkinson & Smith (2001). Linear transformations include translation, rigid registration, similarity, and affine, respectively allowing for translation only, then adding rotation, scaling, and shear. All of the most popular toolsets for MRI registration provide an easy way to implement these transformations out-of-the-box, including FLIRT from the FSL package (Jenkinson & Smith 2001), ANTs (Avants *et al.* 2009), SPM12 (Penny *et al.* 2011), mni_autoreg (Collins *et al.* 1994) and Elastix (Klein *et al.* 2010), offering a variety of cost functions. Intra-modal registration is compatible with a simple least squares method, while for inter-modal registration mutual information or normalized mutual information is often preferred. While these metrics are almost universally available, many toolsets feature cost functions that are not found in the others. A summary can be found in table 1. While these tools have been primarily developed for humans, the algorithms used for linear registration do not require any particular fine tuning for the rodents. Many interpolation schemes are available after the algorithm has found the optimal transformation, however when transforming the atlas volumes a nearest neighbor interpolation is used to insure the labels are preserved as integers.

State of the art registration and segmentation procedures implement affine transformations as a preliminary step, followed by non-linear, diffeomorphic mapping. A diffeomorphic transformation is a differentiable, non-linear transformation with a differentiable inverse, preserving the topological relationships of the subject's anatomy (connected or disjoint structures remain so) and ensuring that diffusion tensors remain positive definite, which is of primary importance in diffusion weighted imaging (Aggarwal *et al.* 2009). These large-deformations algorithms satisfy the inverse-consistency property, insuring that the matrices associated to the forward and reverse mappings are inverse

to each other, Invertibility is a key property of registration when applied to segmentation, to propagate the labels or probabilistic maps.

Bai *et al.* (2012) found that employing the Large Deformation Diffeomorphic Metric Mapping, LDDMM algorithm (Beg *et al.* 2005) for single atlas registration outperformed FFD and Demons (Thirion 1998), with a mean dice score of 0.81 compared to 0.72 from affine registration, although at an high computational expense. Fu *et al.* (2017) compared one linear algorithm (FLIRT) and four diffeomorphic algorithms: DARTEL, geodesic shooting (optimizations of LDDMM, Ashburner (2007), Ashburner & Friston (2011)), diffeo-demons, geodesic-SyN and greedy-SyN (from ANTs). The best performing algorithms were geodesic-SyN and greedy-SyN (Avants *et al.* 2008, 2009), with mean volume overlaps of 0.77 and 0.76 respectively, the overlap of the affine registration averaging at 0.68 across all regions.

The demons algorithm searches for a diffeomorphic transformation with a diffusion based model (**?**), whereas both LDDMM and SyN are based on the optimization of a velocity field mapping one volume to the other through an integration step. While LDDMM is symmetric in theory, the optimization problem in not formulated symmetrically. By contrast Avants *et al.* (2008) implemented an algorithm that exploits the inherent symmetry of the problem and guarantees that the path from the fixed to the moving volume remains the same when the roles are reversed, by defining an appropriate variational energy insuring that the two volumes contribute equally to the path. Geodesic-SyN allows for an unconstrained optimization within the space of diffeomorphic transformations, resulting in an higher accuracy compared to greedy-SyN. While the latter is an approximated approach, it offers a major improvement in terms of speed, at the price of a very small loss in accuracy. For the same registration task, Fu *et al.* (2017) measured a running time of 103.2 minutes when performed with geodesic-SyN, and 27.8 minutes with greedy-SyN.

Klein *et al.* (2009) also highlighted the high accuracy of geodesic-SyN for brain MRI registration in human subjects, however direct application of SyN algorithms with human optimized parameters to rodent populations is not recommended. Fu *et al.* (2016) showed that optimizing the parameters of the SyN protocol for mices results in a 18% improvement of the Dice score compared to the SyN protocol optimized for humans, and a 22% improvement over affine registration. Applying SyN with human-optimized parameters only resulted in a minor improvement compared to the overlap achieved with linear registration methods. Fu *et al.* (2017) recommend large gradient descent steps, as the anatomical variability in mouses is lower than in humans, keeping the number of time points fixed at 2, and a time integration step of 0.05, employing

Table 2. CC: cross correlation; CCH: histogram-based correlation coefficient; CR: correlation ratio; GC: global correlation; GD: gradient difference; ICP: iterative closest point (Euclidean); JHCT: Jensen-Havrda-Charvet-Tsallis; LD: label difference; MS: mean squares; MI: mutual information; NMI: normalized mutual information; NC: normalized correlation; ND: normalized difference; PSE: point set expectation; SSC: ctochastic sign change; VR: variance ratio. ICP, JHCT and PSE are designed for point-set registration.

Toolbox	**Metrics**	**Nonlinear Transformations Options**
ANTs	CC, Demons, GC, ICP, MS, JHCT, Mattes, MI, PSE	BSplineDisplacementField, BSplineExponential, BSplineSyN, Exponential, GaussianDisplacementField, SyN, TimeVaryingBSplineVelocityField, TimeVaryingVelocityField
FSL	MS	FNIRT
Elastix	MS, MI, NC, NMI	B-splines, Thin-plate splines, SplineKernelTransform, WeightedCombinationTransform, BSplineTransformWithDiffusion, BSplineStackTransform
SPM12	Multinomial model	Geodesic shooting, DARTEL

a Gaussian regularizer with $\delta^2_{gradient} = 3$ and $\delta^2_{totale} = 2$, and using cross correlation as a the similarity metric.

While they performed worse according to Fu *et al.* (2017) in high resolution MRI, the DARTEL and geodesic shooting algorithms for SPM12 are also widely used. Both tools are based on an intermediate tissue segmentation step based on a clustering method, registering simultaneously different tissue classes. To the best of our knowledge the FNIRT tool included in FSL has not been compared to the ANTs or SPM12 tools in small rodents, but it does not appear to outperform them for human MRI Klein *et al.* (2009).

FNIRT constructs a diffeomorphic transformation as a sum of diffeomorphic transformations. While this does not guarantee that the sum would be diffeomorphic, FNIRT approaches this problem by rejecting at each iteration non-diffeomorphic deformation fields and projecting them on the closest diffeomorphic field. This

allows for the selection of transformations characterized by a Jacobian within a specified range, whereas different algorithms might result in diffeomorphic transformations with a Jacobian arbitrarily close to zero. The Jacobian determinant of a diffeomorphic transformation can itself be used to characterize local contractions or expansions, allowing for the localization of voxel-level differences in the local shape of brain structures (Pagani *et al.* 2016). FNIRT allows for the direct selection of the optimal deformation withing a specified range. A more inclusive list of toolsets for nonlinear registration can be found in table 2.

2.3.2 Multi-atlas segmentation

One single atlas is unable to characterize individual variability, and its propagation can turn into systematic errors all random errors in the atlas building process. An effective alternative to single atlas segmentation is to employ a database of different atlases, computing the final segmentation using several manually segmented volumes. Each atlas is registered to the target volume, and the final segmentation is derived through a label fusion procedure. The general idea of multi-atlas segmentation resulted in a large variety of techniques for the labeling of biomedical images (Iglesias & Sabuncu 2015) and it can be considered a class of supervised learning algorithms, several of which have been employed for the segmentation of rodent brain MRI.

Lancelot *et al.* (2014) demonstrated a marked improvement of a simple majority voting strategy over both single atlas and the propagation of one probabilistic atlas for the rat brain. Bai *et al.* (2012) compared several common registration and label fusion strategies for the segmentation of *in vivo* mouse brains. They investigated the interplay of affine, FFD, Demons and LDDMM registrations with majority voting, STAPLE and Markov Random Fields (MRF) as label fusion strategies, comparing them to single atlas segmentation. The quality of the registration step remained the most important variable, resulting in the highest Dice score improvements. LDDMM registration improved the average Dice score from 0.724 (affine registration) to 0.812 for single atlas registration, while multi-atlas methods improved the final overlap scores by about 0.03~0.04. The best results were obtained by combining LDDMM registration with either majority voting or STAPLE, resulting in a dice score of 0.845.

The MRF approach (Bae *et al.* 2009) jointly models the distribution of a voxel label with its neighborhood, while the STAPLE algorithm (Warfield *et al.* 2004) estimates the performance of each generator atlas and constructs an estimate of the "true" segmentation via an expectation-maximization algorithm. Unlike a majority vote rule, which selects at each voxel the highest

occurring label, STAPLE is able to identify the correct segmentation even when there are repeated errors in a majority of the segmentations (Warfield *et al.* 2004). However the higher complexity of these algorithms did not constitute an improvement over a much simpler majority voting strategy in this case, failing to capture the subtler variation of the mouse brain.

STEPS (Cardoso *et al.* 2012, 2013) incorporates a local similarity metric in the STAPLE algorithm and combines it with a MRF model to address the problem of global vs local image matching. Ma *et al.* (2012, 2014) confronted it with STAPLE and single-atlas registration after optimizing the parameters required by STEPS with a grid search, highlighting a marked improvement over both procedures on their dataset. The overall Dice score improvement granted by multi-atlas methods is not equally distributed among brain regions. Harder to segment brain structures like the fimbria and the anterior commissure register the highest improvements, of about 0.2 (Ma *et al.* 2012, Bai *et al.* 2012, Ma *et al.* 2014), while the thalamus or the cerebellum underwent improvements smaller by one order of magnitude .

Nie & Shen (2013) proposed a weighed average approach in which the quality of the local alignment is estimated with a mutual information strategy combined with a demons registration approach, implementing a support vector machine classifier. They report an improved 0.859 Dice score over the initial 0.788 single-atlas overlap for in vivo volumes, and respectively 0.90 and 0.85 scores for in vitro volumes. Lee *et al.* (2014) also implemented a majority voting strategy weighed by intensity similarity after a b-spline deformation driven by corresponding particles, reporting a 0.05 overlap improvement over the 0.84 score for the pairwise registration.

While all of these authors provided Dice scores to evaluate their results, direct comparison of Dice scores across different studies, focusing on different atlases and different ROIs, is not recommended. Lacking a study confronting these different segmentation protocols under the same conditions, and resulting in roughly similar dice scores, their performance can hardly be compared in a meaningful way. We can note however that it approaches the inter-rater overlap between different human raters (Ali *et al.* 2005).

2.3.3 Clustering methods

While atlas-based methods can be used for tissue segmentation, an alternative approach is to frame it as a clustering problem, labeling the individual voxels as members of different tissue classes. One of the classical tasks this algorithm is applied to is a 3 classes segmentation of Gray Matter (GM), White Matter (WM) and Cerebrospinal Flluid (CSF). MRI volumes provide effective contrast between

these classes, but the problem is complicated by the bias field, noise and partial volume effects.

Earlier statistical approaches attempted to label single voxels based on probability values determined from the intensity distribution of the image, treating voxels as independent samples drawn from a population. Zhang *et al.* (2001) combined an expectation maximization approach with a Markov Random Field model (Li 1994) to take into account the spatial context of the specific voxels, articulated in a three steps expectation maximization algorithm alternating estimates of the class labels, distribution parameters and bias field, to maximize the interclass variance. This algorithm is currently implemented as the FAST tool in the FSL toolbox. As the initial estimates can suffer in the presence of strong bias fields the algorithm can also be initialized with an a-priori probability map. The number of classes can also be increased, for example to account for strong lesions, or reduced, if the WM-GM contrast is too small in the target volume.

Ashburner & Friston (2005) developed the algorithm that would be implemented in SPM, combining registration and Gaussian mixture clustering. While this expectation maximization algorithm does not explicitly model spatial dependency in the same way of a MRF, context information is derived from the deformable registration of a probabilistic map of the different tissue classes. At each step the mixture parameters, bias fields and deformation are estimated separately while keeping the others constant. As of the current implementation in SPM12, the algorithm supports segmentation in several classes: GM, WM, CSF, bone, soft tissue, background/air. Each class is described by multiple Gaussians to account for partial volume effects and for the possibility that the true distribution might not be normal. Sawiak *et al.* (2009) developed a toolbox to facilitate the extension of SPM functionality to the animal brain, including mouse specific priors out-of-the-box, called SPMMouse.

The Atropos tool (Avants, Tustison, Wu, Cook & Gee 2011) included in ANTs implements an n-tissue segmentation algorithm capable of integrating multimodal information to enhance the segmentation performance with minimal memory requirements. Combining both of the strategies described above Atropos can include either MRFs, template based priors or a weighted combinations of both, as well as bias correction. It can also be used for brain extraction and label propagation from a probabilistic atlas. Atropos supports partial volume classes, for example the class of voxels containing both WM and GM can be classified as a separate category.

Supporting different initialization and optimization strategies, likelihood models, and optimization options, Atropos is a powerful tool with a significant number of parameters the user can tweak to fine tune the tool to their specific needs, and it has been applied to very different tasks like the segmentation of

cysts in mouse kidneys tissues (Xie *et al.* 2015). However this is not always a benefit, and in some cases a more straightforward approach like the ones previously described can still yield good results with less fine-tuning.

In the case of *ex vivo* studies, the fixation procedure can severely impact the performance of a classic 3-classes segmentation. Pagani *et al.* (2016) and Li *et al.* (2009) worked around the overestimation of WM tissue at the expense of GM with Atropos and FSL respectively when implementing voxel based morphometry measures, by increasing the number of classes and reconstructing GM by merging the new classes appropriately. However the large loss of CSF as a consequence of fixation still impaired the quality of WM/CSF discrimination, as reported by Pagani *et al.* (2016).

Many of the methods discussed in this chapter reach overlap scores comparable to those obtained between segmentations from different human raters (Ali *et al.* 2005), especially for diffeomorphic registration. These results are obtained using intuitive methods based on aligning two different images, or based on classic machine learning algorithms offering a reasonable interpretation of their results, in contrast with the methods we will explore in the next chapter. Furthermore, the availability of cross-modality cost functions means that we can easily reuse the same atlases on contrasts different from the one they were originally developed for.

Conversely, these methods also suffer from significant drawbacks. While diffeomorphic registration can account for some individual variability, this may not be the case where anatomical alterations are large and heterogeneous, e. g. as a result of pathology. Appropriately selecting the correct parameters for diffeomorphic inter-modal registration may constitute a serious additional challenge. Furthermore, diffeomorphic registration also suffers from long processing times, as does the production of a minimum deformation template, hindering the application of these methods in large datasets or in a real-time setting.

3 NEURAL NETWORKS

In this chapter we will look into Deep Neural Networks (DNNs) in general, Convolutional Neural Networks (CNNs) (LeCun *et al.* 2015, Douglass 2020), and their application in medical imaging. In this way we will have all the elements we need to later discuss the experiments described in this dissertation.

Neural networks here can be seen as general function approximators. We will be assuming the existence of an unknown underlying function connecting input and output that we try to learn from data. Using this class of powerful methods we can attack a wide range of problems, from a simple regression problem to the processing and classification of natural or medical images, simply by providing examples of input/output pairings.

3.1 FULLY CONNECTED NEURAL NETWORKS

To illustrate the basic principles of DNNs, we will start from the simplest architecture, based on fully connected layers only.

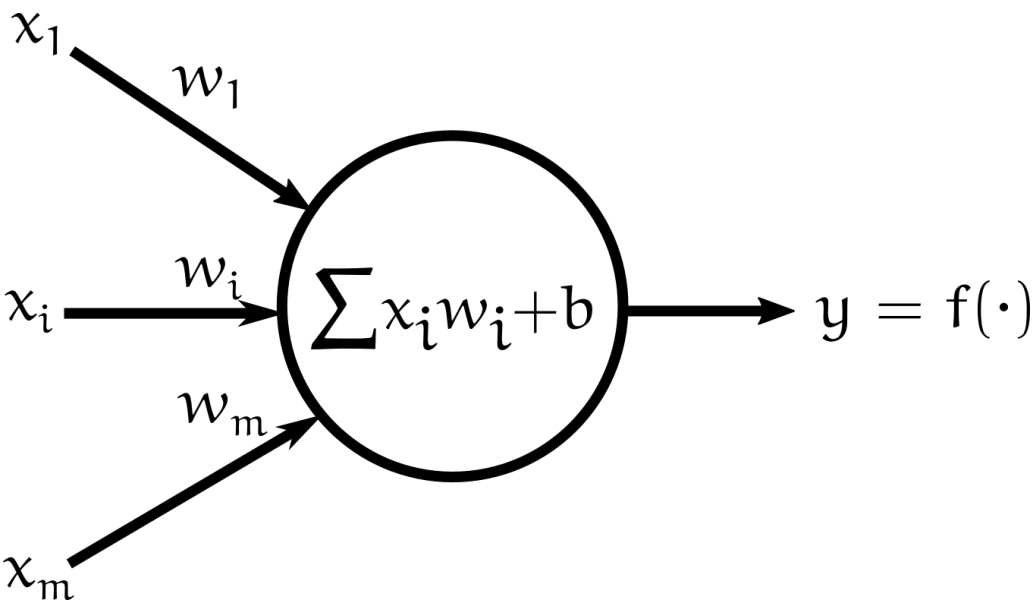

Figure 2. The architecture of a perceptron, performing a weighted sum of its inputs plus a bias term, and applying a nonlinear function.

3.1.1 The perceptron

The elementary construction block of DNNs is the perceptron (Rosenblatt 1958) (Figure 2), inspired by the biological neuron. Given m input variables, let x_i be the i-th input, w_i the associated weight parameter, and b the bias term. Then the perceptron performs the weighted sum of its inputs plus the bias term, and then applies a nonlinear function f:

$$z = \sum_i x_i w_i + b; \quad a = f(z). \tag{7}$$

The function f is called the activation function, and the most common choice is the Rectified Linear Unit (ReLU), where $f(z) = \max(0, z)$. The presence of a nonlinear function is essential to computing nontrivial problems, as we shall soon see. Equation 7 can also be expressed as the vector product:

$$z = \mathbf{x}'\mathbf{w}; \quad a = f(z) \tag{8}$$

where

$$\mathbf{x} = [x_1 \dots x_m]; \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} \tag{9}$$

3.1.2 Architecture

One layer To form one layer of the DNN, we simply need to stack several perceptrons in parallel, accepting the same inputs (Figure 3). Each perceptron is characterized by its own unique weights. For a layer composed of k perceptrons we generalize Equation 8 using a weight matrix $\mathbf{W}$:

$$\mathbf{z} = \mathbf{x}\mathbf{W}; \quad \mathbf{a} = f(\mathbf{z}) \tag{10}$$

where

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & \dots & w_{1,k} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \dots & w_{m,k} \end{bmatrix}. \tag{11}$$

DNN A DNN can include an arbitrary number of layers N, where the outputs of each layer become the inputs to the next (Figure 3). In fact, the *Deep* in Deep Neural Network (DNN) simply indicates that we are stacking two or more layers on top of each other. The first layer will be the input layer, accepting the

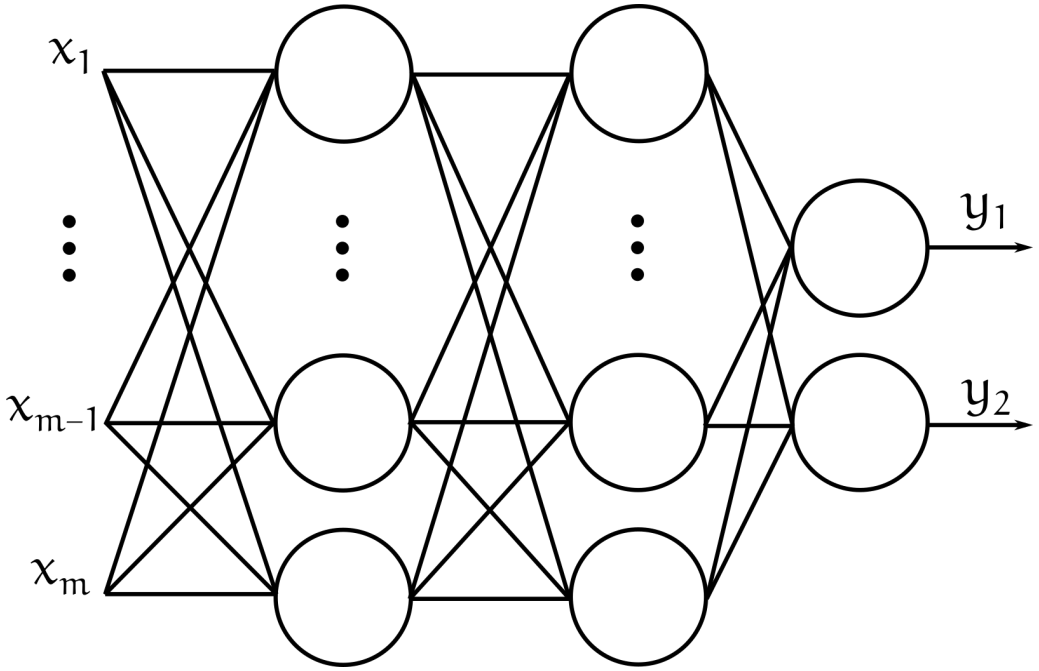

Figure 3. A three-layers DNN, accepting m inputs and returning two outputs.

$\mathbf{x}$ input vector to the DNN, and the last will be the output layer, generating a vector of one or more elements $\mathbf{y}$. Every intermediate layer is called an hidden layer. It is obvious at this point that without the nonlinear functions discussed above any DNN would be equivalent to a shallow neural network with $L = 1$ layers, as the composition of N linear operations is again a linear operation:

$$\mathbf{z} = \mathbf{x}\mathbf{W^1} \dots \mathbf{W^N} = \mathbf{x}\mathbf{W}. \tag{12}$$

The weights (or parameters) w are initialized randomly, and to find parameters that would work for our problem we need to go through a process of optimization.

Output activation The activation function utilized on the output layer is often different from the nonlinearity applied throught the network. For example, in a binary classification problem, we would generally use a sigmoid activation, a smooth function ranging between zero and one defined as $\sigma(z) = 1/(1 + \exp(-z))$. Expanding our classification task to K categories, we typically want to indicate the category characterized by the maximum prediction

using a function approximating the max operator, returning outputs between zero and one. This would be the softmax function:

$$\text{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_i)} \tag{13}$$

3.1.3 Loss function

Before we can start looking for the values of the weights of our DNN we need a criterion that would evaluate how far any generated output would be from the correct one (in supervised learning) or at least be able to recognize a desired or undesired outcome (reinforcement learning). For the rest of this work, we will always be working with supervised learning. Furthermore, we need this criterion to be formalized as a differentiable function, because we will need to take the derivative. As an example we provide here the categorical cross entropy loss, in a classification task between K output categories, for one sample:

$$CCE = -\sum_{i=1}^{K} y_i \log \hat{y}_i \tag{14}$$

where y_i indicates the ground truth value for the i-th category, with $y_t = 1$ characterizing the correct category t and $y_i = 0$ for $\neq t$; $\hat{y}_i$ indicates the output prediction for the i-th category. This means that for each data sample this expression is reduced to $CCE = -\log \hat{y}_t$. As y_t is a predicted probability or pseudo-probability of the correct class, this quantity is minimized as $0 \leq y_t \leq 1$ approaches 1. The optimal prediction minimizes the loss function.

3.2 TYPES OF MACHINE LEARNING

Neural networks can be applied in all of the different branches of machine learning. While we will only be dealing with supervised learning in this work, to understand what is it that we refer to under the name of supervised learning it could be worth introducing the different main types of machine learning: supervised, unsupervised, and reinforcement learning.

3.2.1 Supervised

In supervised learning we produce a mapping from the input features to the output labels based on existing labeled data. This can include sophisticated methods such as the ones described in this chapter, as well as random forests

(Breiman 2001) or support vector machines (Cortes & Vapnik 1995), and much simpler methods such as regression. Supervised learning is not limited to trained algorithms, including e.g. k nearest neighbors classification. Based on this definition, single and multi-atlas segmentation can also be argued to be examples of supervised learning.

3.2.2 Unsupervised

In contrast, unsupervised methods do not utilize existing labels but only leverage the patterns or structural properties of the data. These includes clustering or dimensionality reduction, as well as generative methods or anomaly detection. As an example, a generative model trained on normal data can be utilized to detect anomalies in new sample images (Wang *et al.* 2020).

At the threshold between supervised and unsupervised we find semi-supervised methods, where partial or less detailed labels are utilized (Papandreou *et al.* 2015, Kervadec, Dolz, Tang, Granger, Boykov & Ayed 2019, Zeng *et al.* 2021).

3.2.3 Reinforcement learning

Reinforcement learning is concerned with behaving agents taking actions in an environment. For example, the problem of playing a game of Go (Silver *et al.* 2017) or to control a wind turbine in response to the changing conditions of its environment (Saenz-Aguirre *et al.* 2019) can both be approached with reinforcement learning. Reinforcement learning does not require prior knowledge of what the optimal action to take would be, but depends on rewarding or punishing behaviors based on their outcome, and a balance between exploration of new possible behavior and exploitation of current knowledge.

3.3 DATA

Data is an essential aspect of supervised learning. Before we can move on to discuss the training of a DNN, we need a dataset to train and validate it. Because of the large number of parameters involved DNN are generally data-hungry and can require large datasets to avoid overfitting, that is the overspecialization of our networks to the specific data we used in training. The data is generally divided in two or three sets: training, validation, and test set. The first is used to train the network, the second ban be used to iteratively check performance during training or to test different architecture configurations. Finally, the test set is used only once, to test the final performance of the neural network on data it has never encountered.

There is no general approach to predict exactly the amount of data that will be required. In general we expect classification to require a larger amount of data, as it can be easy for a large DNN to learn to recognize each sample individually. In contrast, medical imaging segmentation can require a surprisingly small quantity of data. For example, the anatomical segmentation of the healthy mouse brain in MRI can be a very narrow domain. All segmentation maps will look very similar to each other, containing the same objects in similar relative positions, in animals acquired according to the same protocol using the same scanner. At the same time, we can understand segmentation to a voxel-wise classification, with each sample containing over 1.5 million (non-independent) samples. In this case, we can train a functional neural network (when operating on the same domain) with an handful of samples. Conversely, we will need far more data for a neural networks trained for object segmentation in an automated vehicle, which can encounter a large number of different objects, in a large number of different environments, and in different conditions of visibility.

3.4 OPTIMIZATION

DNNs are typically optimized through gradient descent, which leaves us with the problem of calculating the gradient of the loss function with respect to each parameter. While we could use the trapezoid rule or an analogous numerical method, these are too slow in practice. The automatic differentiation algorithm included in frameworks like PyTorch (Paszke *et al.* 2017) calculates the gradient one layer at a time, starting from the final output of the loss function and working its way backwards. This process called backpropagation reuses the computations generated initially by the DNN to efficiently calculate the gradient. Given the large number of simple operations that can be parallelized, backpropagation is greatly sped up by using specialized hardware such as Graphic Processing Units (GPUs) or Tensor Processing Units (TPUs).

With L as the loss function, having calculated its gradient ∇L, we can take a step in parameter space to minimize L by updating the weights $\mathbf{W}$ of our DNN:

$$\mathbf{W}_{new} = \mathbf{W}_{old} - \alpha \nabla L(\mathbf{W}) \tag{15}$$

where α is a scaling factor called the learning rate and $\mathbf{W}$ refers to the combined weights of all DNN layers. To sum up the general idea of gradient descent optimization, we iteratively repeat the following steps:

1. Select samples $\mathbf{x}$ from the training set (a batch).

2. Evaluate the output of the DNN $\hat{y} = \mathbf{x}$.

3. Calculate the loss function $L(\hat{y}, y|\mathbf{W})$.

4. Calculate the gradient $\nabla L(\mathbf{W})$.

5. Perform one gradient descent step.

6. Repeat until convergence.

Performing one pass over the entire training set defines one epoch, and normally a large number of epochs could be required to train the algorithm. The learning rate, the size of the batch, and other parameters describing the architecture (e. g. the size and number of layers) are called hyperparameters of the neural network. We could add an external loop to the one described above to explore a different combinations of hyperparameters, iteratively testing the network on the validation data. The validation data can also be used for early stopping: rather than simply train the network for a predefined length of time, we can keep training until the average L over the validation data (the validation loss) stops decreasing.

Even in those cases when it might be possible to fit the entire training set in one batch, this is typically not advised. By changing the batch the shape of L in parameter space keeps changing, adding noise that can help us avoid being stuck in local minima. Conversely, a batch size of one can be excessively noisy, and increase convergence time.

The specific choice of α can also help speed up convergence. A large α mean we will take longer steps in parameter space, however, this also means it could be easier to keep jumping over the global minimum without finding it. A solution to this problem is to use an optimizer adaptively decreasing α, or to directly schedule its decrease at regular intervals.

A sophisticated optimizers can include additional features. This is the case for the Adam optimizer (Kingma & Ba 2014), displaying an adaptive fine-tuning of the learning rate, momentum and weight decay. To increase the stability of the gradient descent, rather than using the gradient we can use a running average of the gradients to increase stability. This feature is known as momentum. Weight decay is instead a regularization strategy to reduce overfitting, adding a term to the loss function to penalize large values for the network parameters: $\lambda/2\|\mathbf{W}\|^2$. This prevents the network from overly specializing for specific features and, according to recent work by Power *et al.* (2022), it appears weight decay would allow a DNN to extract a general rule from small artificial datasets after a long training time.

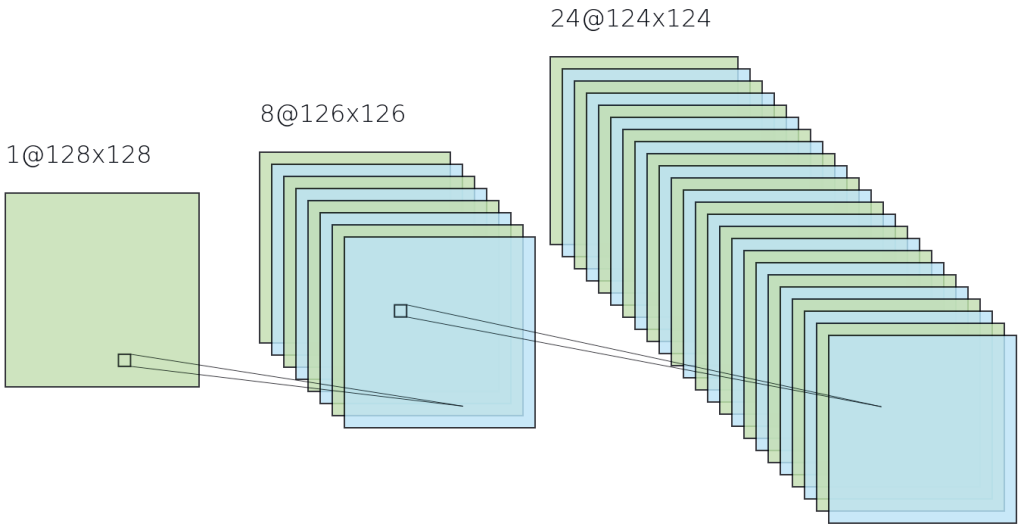

Figure 4. A basic representation of a CNN, applying 8 filters to a 1-channel input, followed by a second layer appying 24 filters. Drawn using `http://alexlenail.me/NN-SVG/LeNet.html` and reproduced under CC license.

3.5 CONVOLUTIONAL NEURAL NETWORKS

For image processing the fully-connected architecture is not optimal. This is a generic architecture that makes no assumptions on the possible relations between different inputs, however the proximity of different pixels is a fundamental element in images. We would also like to be able to use the translational symmetry we find in images: we would like to use the same parameters to recognize the same features regardless of where they appear in the image. To put it in different words, a cat is still a cat even after translating it by one or more pixels. An added benefit of this approach is that the input size is not fixed, as long as all convolutions can be performed.

Instead of connecting each neuron to the entire input, we can imagine to have a set of neurons accepting inputs from a very narrow receptive field, e. g. a 5×5 box. If the image features multiple channels e. g. a color image, it would be a $3 \times 5 \times 5$ input domain. By translating this neuron over the input image we can use its outputs to build a new feature map. In this context this operation is called a convolution, and the neuron is called a filter. A network constructed based on these operations is a CNN, and a set of convolutions applied to the same input or set of input maps constitutes one convolutional layer in a CNN. Filters in the following layers will typically (but not necessarily)

operate on all channels at the same spatial position. With reference to Figure 4, if we first apply 8 filters sized 5×5 to our input, we will have 8 resulting feature maps. The following layer would then apply $8 \times 5 \times 5$ filters on these feature maps.

Other architectural elements Beyond the basic operation of convolution there are other important elements we need to consider. As visible in Figure 4, any filter other than 1×1 will shrink the size of the feature maps. It may sometimes be desirable then to restore them to the original size, through a padding operation, typically with zeros. To instead shrink the feature maps and downsample them we would use a pooling operation, dividing the feature maps in smaller (typically 2×2) domains and taking the maximum (max pooling) or the average (average pooling). The inverse operation is called unpooling (Noh *et al.* 2015), where by saving the positions of the selected elements we can place back elements of a smaller feature map in those original positions, setting the rest to zero. Other upscaling operations can be more complex, such as transpose convolutions, which upscale feature maps with learnable parameters, or simple upscaling and interpolation. Lastly, normalization layers such as batch normalization (Ioffe & Szegedy 2015) help control the mean and variance of each layer, whereas they would normally keep "chasing" the changing statistics of the connected layers, a problem known as the internal covariate shift.

3.6 BRAIN MRI SEGMENTATION IN SMALL RODENTS

3.6.1 Preprocessing

We are finally ready to enter the subject of applications of CNNs in Magnetic Resonance Imaging (MRI). As in many domains, there are a number of MRI-specific caveats to consider. One such issue is that while MRI contrasts can be quantitative, quantitative MRI is not the norm in the clinical or preclinical setting. The contrasts we typically encounter are T_1 or T_2-weighted MRI, developed to be clear and meaningful for the human eye but carrying marked differences based on a combination of scanner manufacturer, coils, and acquisition parameters. Because of this the specific intensity values are not highly meaningful and can vary greatly even throught a dataset where each sample has been acquired according to the same procedure. MRI volumes are then normalized individually on input converting the intensity values to Z-scores, rather than using the overall dataset statistics. This simply means subtracting the mean and dividing by the standard deviation all voxel intensity values.

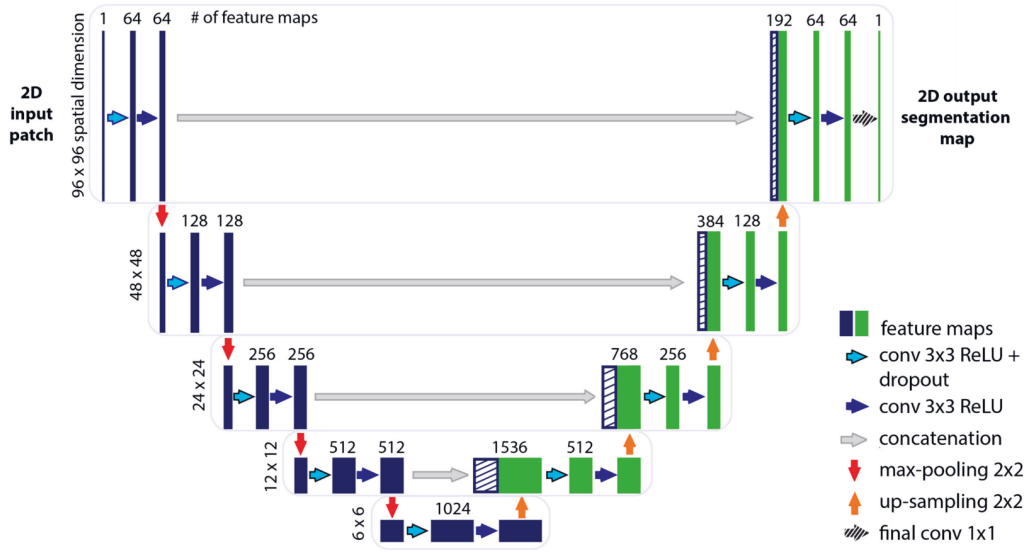

Figure 5. Illustration of the U-Net architecture for a 96×96 input patch, reproduced under license CC BY from Livne *et al.* (2019). Each box corresponds to a multi-channel feature map. Dashed boxes indicate the concatenated feature maps. Numbers above each box indicate the number of channels, numbers to the side indicate the sizes of the feature map.

A second challenge is the large size of MRI volumes. While we are often interested in processing them at full resolution, to keep the highest possible level of detail, these can often be too large to be processed using the available GPU memory. To solve this problem we would need to switch to a patch-based approach, or to preliminarily crop the scans to a smaller area of interest. Applying skull-stripping can help us identify a bounding box and discard data we are not interested in.

Data augmentation A special case of preprocessing is data augmentation. As we are loading samples while training a CNN, we might want to randomly apply a transformation that would result in another realistic training sample, exploiting the symmetries available in our dataset. For example we might want to apply random rotations, translations, or scaling to simulate new samples that are not part of our training set, but are still representative of the domain distribution we are interested in.

3.6.2 U-Net

Many of the architectures used in medical imaging are based on U-Net (Ronneberger *et al.* 2015), initially developed for the segmentation of 2D biomedical images. The architecture is displayed in Figure 5. Ignoring the concatenation arrows (in gray), this would be a typical encored-decoder architecture: an encoding branch processes the input downsampling it to a smaller feature map, distributed on a larger number of channels, compressing the initial information into a bottleneck layer at the bottom. This process is then reversed by the decoding branch, leading us back to the original size, and generating the final segmentation map. The idea behind this process is to force the network to select which features are important, requiring higher level abstractions, and only then generate the output. To downsample the inputs different levels on the encoder are connected by pooling operations, while on the decoder we may find transpose convolutions, as in the original implementation by Ronneberger *et al.* (2015), unpooling layers Noh *et al.* (2015) or simple upsampling and interpolation, ad displayed here in an implementation by Livne *et al.* (2019). The skip or concatenation connections act by concatenating the final feature maps at each level on the encoder to the input maps at each layer of the decoder, allowing it to easily integrate multi-scale information and better propagate the gradient during training.

This architecture has been shown to generalize even from a limited amount of annotated data (Xie *et al.* 2015), and as such is well suited for medical imaging, where datasets as large as the ones commonly used for CNNs are rare. Its architecture can easily be generalized to 3D data (Milletari *et al.* 2016), and has been applied in different forms for brain MRI, both by combining the outputs from different 2D views (Wachinger *et al.* 2018) and by directly processing the 3D data Roy, Conjeti, Navab & Wachinger (2018). A large body of work was inspired by U-Net, combining it with other ideas in machine learning and computer vision such as attention or squeeze-and-excitation blocks (Badrinarayanan *et al.* 2017, Oktay *et al.* 2018, Rundo *et al.* 2019). Valverde *et al.* (2019) (Valverde *et al.* 2020) recently demonstrated the effectiveness of U-Net-like architectures in preclinical research, designing the first DNN for the segmentation of ischemic lesions in rodents and achieving segmentation accuracy comparable or better to inter-rater agreement in manual segmentation.

3.6.3 Other approaches

Alternative approaches include work based on image-to-image translation (Isola *et al.* 2017) or mixed-scale architectures (Pelt & Sethian 2018). For small-animal MRI, the applications of CNNs have been limited to skull-stripping:

Roy, Knutsen, Korotcov, Bosomtwi, Dardzinski, Butman & Pham (2018) trained a CNN algorithm based on Google Inception (Szegedy *et al.* 2015) for the skull-stripping in humans and mice after traumatic brain injury, achieving better performance than other state-of-the-art methods (3D Pulse Coupled Neural Networks (3D-PCNN) (Chou *et al.* 2011) and Rapid Automatic Tissue Segmentation (RATS) (Oguz *et al.* 2014)).

A more fundamental change in approach is given by predictive registration (Gutierrez-Becker *et al.* 2017, Dalca *et al.* 2018, Yang *et al.* 2017), where instead of attempting to predict the segmentation map directly the CNN is used to estimate the velocity field associated to a diffeomorphic deformation, given two volumes. The resulting momentum can either be used directly or as a prior to initialize an optimization algorithm, thus drastically reducing the optimization time. A promising development for this approach is the recent publication of the SynthMorph algorithm (Hoffmann *et al.* 2021), to train a contrast-agnostic predictive registration network trained on synthetic data.

3.6.4 Loss functions

We already discussed categorical cross entropy, Equation 14. This is a valid loss function for image segmentation, treating each voxel as a separate classification sample. However, just as we look at the overall overlap between different regions to evaluate a segmentation algorithm using the Dice score (section 2.1, we might benefit from using a similar optimization criterion.

Dice loss Recent literature suggests that Dice-based loss functions (Milletari *et al.* 2016, Sudre *et al.* 2017, Roy, Conjeti, Navab & Wachinger 2018) would constitute an improvement over cross-entropy losses for the segmentation of medical images (Karimi & Salcudean 2019). Let $p_l(i)$ be the predicted probability of voxel i of belonging to category l, and $g_l(i)$ the ground truth for voxel i ($g_l(i) = 1$ if and only if l is the correct prediction for this voxel). Then, a Dice-based loss function can be written as:

$$L_{Dice} = -\sum_{l=1}^{K} \frac{2\sum_i p_l(i) g_l(i)}{\sum_i p_l^2(i) + \sum_i g_l^2(i)}, \tag{16}$$

where K is the number of Regions of Interest (ROIs) plus the background class.

Generalized Dice loss When the segmentation process targets rare observations, a severe class imbalance is likely to occur between candidate labels. This refers to a situation in which some classes are under represented

in the training set compared to other classes, leading to a loss of performance. This problem is especially relevant when using loss functions that attribute the same importance to all pixels, independently. While the Dice loss already partially addresses this problem, Sudre *et al.* (2017) also included weighting in its formulation, as follow:

$$L_{GDice} = 1 - 2\frac{\sum_l w_l \sum_n g_{ln} p_{ln}}{\sum_l w_l \sum_n g_{ln} + p_{ln}}, \tag{17}$$

with the weight parameters for label l defined as $w_l = (\sum g_{ln})^{-2}$, and g_{ln}, p_{ln} indicating respectively the ground truth and predicted label for voxel n under category l. These increase the weight of smaller classes over the larger ones.

3.6.5 Comparison with registration-based methods

Interpretability In a machine learning context, we could understand registration-based methods to be akin to a K nearest neighbors algorithm. The general intuition behind K nearest neighbors is to perform inference on new data using the nearest K point in feature space, having defined an appropriate distance metric. In registration-based segmentation we look for a transformation that would minimize a "distance", expressed as a cost function, to place one or more and unlabeled samples in the same space. The atlases are then used to label the target volume, with methods such as STEPS (Cardoso *et al.* 2012, 2013) providing a more sophisticated strategy to combine the different atlases and evaluate their similarity to the target. The idea behind registration-based methods is intuitive and we can more easily understand the process resulting in a prediction.

In contrast, CNNs are a black-box method. It is true that we can visualize the role of different layers by examining which features they respond to the most (Zeiler & Fergus 2014), however the actual interpretation of which process and features led to a prediction is difficult. Approaches such as DeepSHAP (Lundberg & Lee 2017) can highlight which of the input features contributed the most to a prediction, and can indeed be generalized to segmentation by understanding it as a voxel-wise classification problem. However, while technically possible, these results are again hard to understand and interpret in any meaningful way.

Robustness To start, we can point out that both CNNs and registration-based methods should be considered atlas-based, with the fundamental difference that CNNs encode the knowledge contained in multiple atlases in their parameters, while methods based on registration look for an explicit

transformation between each individual atlas and a target space. Becaue of this fundamental difference, each approach has a unique advantage over the other in terms of robustness.

By abstracting the anatomical knowledge and encoding it in its parameters, a CNN can better adapt to the presence of anatomical alterations, as we will later show with an experiment. Conversely, registration is limited by the necessity of directly adapting specific volumes, with a more limited capacity for abstraction. Furthermore, while in both cases a larger set of atlases can help tackle the problem, for CNNs this brings no increase in the inference time, while the diffeomorphic registration of a larger set of atlases will. This also results in a larger scope of possible applications, for example the direct segmentation of brain lesions, without relying on strict assumption about the distribution of the data or a specific imaging modality (Valverde *et al.* 2019, 2020).

In contrast, at the moment of writing CNNs are highly limited in their capacity of adapting to new contrasts, and like many DNNs can suffer from data drift in all practical applications, that is a decreasing accuracy over time as the distribution of real-world data keeps shifting while the model remains unchanged. Registration-based methods can on the other hand be applied cross-modality, as we saw in section 2.3.1.2. In specific applications and source-target contrast pairing this may be more difficult, result in larger inaccuracies, or fail entirely, but at the moment remains an advantage over CNNs. Generative models (Isola *et al.* 2017), transfer learning (Valverde *et al.* 2021*a*) and the production of synthetic training data (Billot *et al.* 2020) are promising approaches to overcome this challenge using CNNs.

4 DATASETS

In this chapter we will describe the datasets employed in the present work. These MRI datasets differ in their resolution, target animals (mice or rats), the presence of pathological states, acquisition parameters and licensing rights. These were acquired during several studies and have been reutilized for the present experiments, meaning that we are describing retrospective data. No new scans were acquired specifically for the experiments described in the following chapters.

Using the Charles River (CR) data we have trained a Convolutional Neural Network (CNN) and tested it on a large body of data including both healthy mice and mice modeling Huntington's disease. The MRM NeAt dataset was then used to test the same segmentation method on publicly available data. EpiBioS4Rx and EPITARGET were used to train a CNN to perform anatomical segmentation in the presence of Traumatic Brain Injury (TBI), and look for anatomical biomarkers of epileptogenesis.

4.1 CHARLES RIVER DATASETS

Animals A total of 849 mice (Charles River Laboratories, Germany) were used: 32 mice for the train and validation set and 817 mice for the test set (Table 3. All mice were housed in groups of up to 4 per cage (single sex) in a temperature (22$\pm$1°C) and humidity (30-70%) controlled environment with a normal light-dark cycle (7:00-20:00).

Train and validation set animals were scanned at four different ages (5 weeks, 12 weeks, 16 weeks, 32 weeks) resulting in 128 volumes. All train and validation set animals were Wild Type (WT) males imaged during the course of a single study.

The test set animals were part of 10 studies scanned at a single or multiple ages from 4 up to 60 weeks, and included both WT and several Huntington (HT) genotypes: R6/2, Q175, Q175DN, Q111, Q50 and Q20, for a total of

Table 3. Summary characteristics of the CR and MRM NeAt datasets. BM refers to the brain mask. The test dataset included various genotypes of both sexes.

Dataset name	# Animals	# MRIs	# ROIs	Type
Train and validation	32	128	4 + BM	WT males
Test	817	1,782	2 + BM	various
MRM NeAt	10	10	37 + BM	WT males

1,782 Magnetic Resonance Imaging (MRI) scans. The groups included both males and females. These volumes were acquired as part of ten studies of Huntington's disease, kindly provided by the CHDI 'Cure Huntington's Disease Initiative' foundation. This large test set, including samples from different studies and animal models over the course of four years, allows us to test our neural networks in a realistic setting. An ideal industry application of automated region segmentation would in fact be training a network utilizing a limited number of volumes, only to later apply it over the years to many different studies in which data is imaged using the same scanner.

MRI Mice were anesthetized using isoflurane (5% for induction, 1.5-2% maintenance) in 70%/30% mix of N_2O_2 carrying gas, fixed to a head holder and positioned in the magnet bore in a standard orientation relative to gradient coils. Respiration rate and temperature were monitored using PC-SAMS software and Model 1030 Monitoring & Gating System, Small Animal Instruments, Inc., Stony Brook, NY. The temperature was maintained at $\sim 37\,\mathrm{C}$ using Small Animal Instruments feedback water heating system.

All acquisitions were performed using a horizontal $11.7\,\mathrm{T}$ magnet with a bore size of $160\,\mathrm{mm}$, equipped with a gradient set capable of maximum gradient strength of $750\,\mathrm{mT/m}$ and interfaced to a Bruker Avance III console (Bruker Biospin GmbH, Ettlingen, Germany). A volume coil (Bruker Biospin GmbH, Ettlingen, Germany) was used for transmission and a surface phased array coil for receiving (Rapid Biomedical GmbH, Rimpar, Germany). T_2 weighted anatomical images were acquired using a TurboRARE sequence with effective TR/TE $= 2500/36\,\mathrm{ms}$, 8 echoes, $12\,\mathrm{ms}$ inter-echo distance, matrix size 256x256, FOV 20.0x20.0 mm^2, 31 $0.6\,\mathrm{mm}$ thick coronal slices, $-0.15\,\mathrm{mm}$ interslice gap, and 8 averages. Concerning the test data, MRI experimental parameters only differed in acquiring 19 $0.7\,\mathrm{mm}$ thick contiguous coronal slices.

Annotation Volumes within each study were manually segmented by an experienced rater, who had received a training and passed the qualification tests according to SOP (Standard Operating Procedure) for volumetric analysis in mice. Different studies were analyzed by different raters. Each training volume was manually segmented by a single rater drawing the brain mask and delineating 4 regions of interest: cortex, hippocampi, striati and ventricles. The brain mask did not include the olfactory bulb or the cerebellum. For the test set, only 3 regions were manually labeled: brain mask, cortex and striati.

Table 4. Summary of ROIs delineated in the MRM NeAt atlases. L and R indicate respectively the ROI belongign to the left and right hemispheres.

MRM NeAt ROIs		
Brain Mask	Fimbria R	Hippocampus R
External Capsule	Caudate Putamen R	Anterior Commissure R
Globus Pallidus R	Internal Capsule R	Thalamus R
Cerebellum R	Superior Colliculi R	Ventricles
Hypothalamus R	Inferior Colliculi R	Central Gray R
Neocortex R	Amygdala R	Olfactory bulb R
Brain Stem	Rest of Midbrain R	Basal Forebrain Septum R
Fimbria L	Hippocampus L	Caudate Putamen L
Anterior Commissure L	Globus Pallidus L	Internal Capsule L
Thalamus L	Cerebellum L	Superior Colliculi L
Hypothalamus L	Inferior Colliculi L	Central Gray L
Neocortex L	Amygdala L	Olfactory bulb L
Rest of Midbrain L	Basal Forebrain Septum L	

4.2 MRM NEAT

The MRM NeAt dataset includes atlases of 10 individual T_2*-weighted *in vivo* brain MR images of 12-14 weeks old C57BL/6J mice; each with 37 labelled anatomical structures (listed in Table 4) in addition to the brain mask (Ma *et al.* 2008). This dataset was downloaded from `https://github.com/dancebean/mouse-brain-atlas`, where an improved atlas is available (bias correction has been applied, left and right labels have been separated and 4th ventricle label added). This dataset was used to evaluate the STEPS algorithm by Ma *et al.* (2014) and is used here for the purpose of comparing our work and STEPS on a larger number of ROIs, on isotropic resolution MRI, using freely available data. As detailed in Ma *et al.* (2008), T_2-weighted MR data with a voxel-size of $0.1\,\mathrm{mm}^3$ requiring about 2.8 hours of scan time were acquired with a 3D large flip angle spin echo sequence using a super-conducting $9.4\mathrm{T}/210$ mm horizontal bore magnet (Magnex) controlled by an ADVANCE console (Bruker) and equipped with an actively shielded 11.6 cm gradient set (Bruker, Billerica, MA).

Table 5. Number of MRI scans per cohort (EpiBioS4Rx, EPITARGET) in different time points and treatment groups [TBI, sham-operated experimental controls]. In parenthesis, the first number indicates the number of volumes used for manual annotation of the hippocampus and the second indicates the number of volumes used for manual annotation of brain masks. TBI+ and TBI- refer to the presence of absence of epileptic seizures among phenotyped animals. *Abbreviations*: d, day; TBI, traumatic brain injury.

Timepoint	TBI+	TBI-	TBI	Sham
EpiBioS4Rx				
2 d	9	23	43 (7, 0)	12 (5, 1)
9 d	9	23	42 (6, 1)	13 (5, 1)
30 d	9	23	42 (6, 1)	13 (5, 0)
150 d	9	23	40 (7, 1)	13 (5, 1)
EPITARGET				
2 d	29	84	118 (4, 0)	23 (2, 2)
7 d	29	82	117 (4, 2)	23 (2, 0)
21 d	29	84	117 (4, 0)	23 (2, 2)

4.3 UNIVERSITY OF EASTERN FINLAND DATASETS

4.3.1 EpiBioS4Rx

The Epilepsy Bioinformatics Study for Antiepileptogenic Therapy (EpiBioS4Rx, `https://epibios.loni.usc.edu/`) is an international multicenter study funded by National Institutes of Health with the goal of developing therapies to prevent posttraumatic epileptogenesis. The 7-month MRI follow-up of the EpiBioS4Rx animal cohort has been described in detail previously (Ndode-Ekane *et al.* 2019, Immonen *et al.* 2019). Here, we have analyzed the data from the University of Eastern Finland (UEF) subcohort. We describe only the details that are important for the present study.

Animals Adult male Sprague-Dawley rats (Envigo Laboratories B.V., The Netherlands) were used. They were single-housed in a controlled environment (temperature 21-23°C, humidity 50-60%, lights on 7:00 am to 7:00 pm) with free access to food and water. Severe traumatic brain injury was induced in the left hemisphere by lateral fluid percussion under 4% isoflurane anesthesia

(Ndode-Ekane *et al.* 2019). Sham-operated experimental controls underwent the same anesthesia and surgical procedures without the induction of the impact.

As summarized in Table 5, the entire cohort included 56 rats (13 sham and 43 with TBI), of which the 12 (5 sham, 7 TBI) first animals to complete follow-up were selected for manual annotation of the hippocampus. Mean impact pressure was 2.87 $\pm$ 0.82 atm in the entire cohort and 2.92 $\pm$ 1.37 atm in the manual annotation subcohort. Animals phenotyped as epileptic are here indicated as TBI+, non-epileptic as TBI-, and animals that could not be phenotyped were not included in TBI+ *vs* TBI- experiments.

MRI Rats were imaged 2 days (d), 9 d, 1 month, and 5 months after TBI or sham surgery (Table 5) using a 7-Tesla Bruker PharmaScan MRI scanner (Bruker BioSpin MRI GmbH, Ettlingen, Germany). During imaging, rats were anesthetized with isoflurane. A volume coil was used as radiofrequency transmitter and a quadrature surface coil designed for the rat brain was used as receiver. Local magnetic field inhomogeneity was minimized using a three-dimensional field map-based shimming protocol. All images were acquired using a three-dimensional multi-gradient echo sequence. A train of 13 echoes was acquired, where the first echo time was 2.7 ms, the echo time separation was 3.1 ms, and the last echo time was 39.9 ms. The voxel size was 0.16x0.16x0.16 mm^3, the repetition time was 66 ms, the flip angle was 16°, the number of signal averages was 1, and the imaging time was 10 min 44 s. Images with different echo times were summed to produce a high signal-to-noise ratio image for segmentation and image registration.

4.3.2 EPITARGET

EPITARGET (`https://epitarget.eu/`) was a European Union Framework 7 -funded, large-scale, multidisciplinary research project aimed at identifying mechanisms and treatment targets for epileptogenesis after various epileptogenic brain insults. The 6-month MRI follow-up of the EPITARGET animal cohort from UEF has been described in detail previously (Lapinlampi *et al.* 2020, Manninen *et al.* 2020). We describe only the details that are important for the present study.

Animals Adult male Sprague-Dawley rats (Envigo Laboratories S.r.l., Udine, Italy) were used for the study. The housing and induction of left hemisphere TBI or sham injury were as described for the EpiBioS4Rx cohort. However, injury surgery was performed under pentobarbital-based anesthesia instead of isoflurane. The entire cohort included 144 rats, and images from the first 6

rats (2 sham, 4 TBI) were selected for manual annotation of the hippocampus. Mean impact pressure was 3.26 $\pm$ 0.08 atm in the entire cohort and 3.22 $\pm$ 0.02 atm in the manual annotation subcohort.

MRI Imaging was performed as described for the EpiBioS4Rx cohort, except that (a) imaging was performed 2 d, 7 d, and 21 d after TBI or sham surgery (Table 5) and (b) all images were acquired with a two-dimensional multislice multigradient echo sequence. A train of 12 echoes was collected, where the first echo time was 4 ms, the echo time separation was 5 ms, and the last echo time was 59 ms. In-plane image resolution was 0.15x0.15 mm^2, slice thickness was 0.5 mm, number of slices was 24, repetition time was 1.643 s, flip angle was 45°, number of signal averages was 4, and imaging time was 11 min 37 s. Images with different echo times were summed to produce a high signal-to-noise ratio image for segmentation and image registration.

4.3.3 Annotation

For outlining the ROIs, the 3D (EpiBioS4Rx) and multi-slice 2D (EPITARGET) T_2^*-weighted MRI images were imported as NIfTI files (.nii) into Aedes 1.0 - an in-house tool with graphical user interface for medical image analysis. Aedes is available at `http://aedes.uef.fi/` and runs under MATLAB (MATLAB Release 2018b, The MathWorks, Inc.).

Brain mask A trained researcher collaborator (Elina Hämäläinen) outlined the brain surface on $160\,\mathrm{mm}$-thick (EpiBioS4Rx) or $150\,\mathrm{mm}$-thick (EPITARGET) horizontal MRI slices, covering the entire dorsoventral extent of the cerebrum (excluding the olfactory bulbs and cerebellum). In addition, E.H. outlined the brain surface on $160\,\mathrm{mm}$-thick (EpiBioS4Rx) or $500\,\mathrm{mm}$-thick (EPITARGET) coronal brain slices to increase the accuracy of dorsal and ventral delineation of the brain surface (Figure 7 and 8). In the EpiBioS4Rx cohort, we drew the whole brain outline for 6 scans from 6 different rats, outlining on average 33.7 $\pm$ 1.4 (range 31 – 37) horizontal slices for each MRI scan. In the EPITARGET cohort, we prepared the whole brain mask for 6 rats and the mean number of MRI slices outlined per case was 10.8 $\pm$ 0.9 (range 10 – 12).

Brain mask completion Only six brain masks were manually labeled in the EpiBioS4Rx dataset and every second sagittal slice was annotated. To reconstruct complete brain masks, we first applied a binary closing operation with a hand-crafted kernel to reconstruct the brain mask, and then filled any remaining holes in the mask volume (Figure 6). Morphological operations were implemented using the scikit-image library (Van der Walt *et al.* 2014).

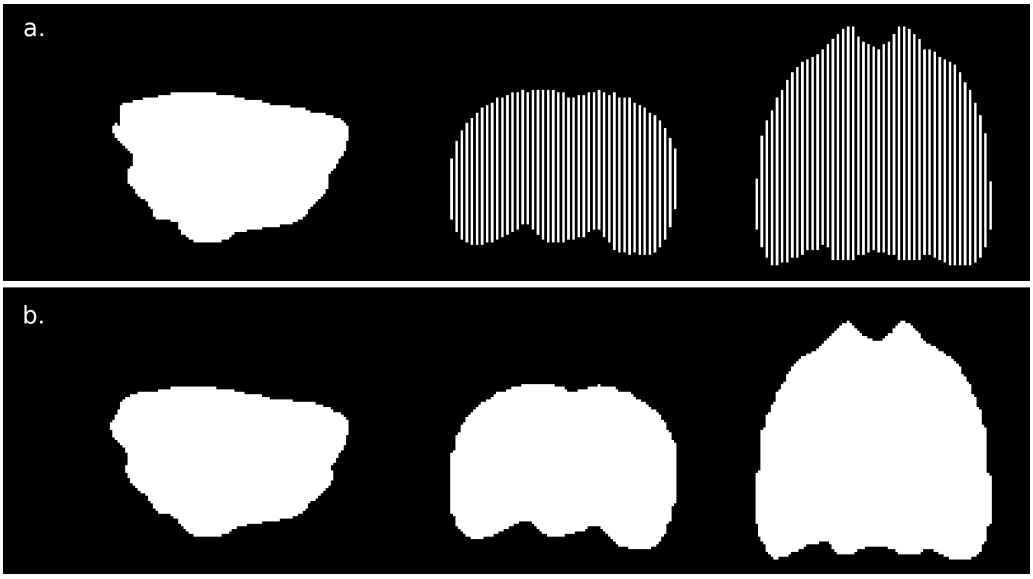

Figure 6. Brain mask completion for the EpiBioS4Rx dataset. a. First, we manually labeled the brain mask in every second sagittal slice. b. Second, we applied a binary-closing operation to obtain a brain mask for the whole brain. Reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

To generate brain masks for the complete training datasets, we trained a single 3D CNN for each dataset as described in Chapter 5.2, using the same overall structure and number of channels, but limiting the output to the brain mask. Using this network, we generated brain-mask labels for the remaining animals, so that our CNN could be trained on these data for both skull stripping and hippocampus segmentation.

Hippocampus Outlines of the ipsilateral (left) and contralateral hippocampus were drawn by E.H. on each coronal MRI slice where the hippocampus was present (slice thickness in EpiBioS4Rx 0.16 mm and in EPITARGET 0.50 mm). In addition to the hippocampus proper and the dentate gyrus, the outlines included the fimbria fornix, but excluded the subiculum (Figures 9 and 10). Manual annotation was performed with the help of thionin-stained coronal $30\,\mathrm{mm}$-thick histological sections of the same brain available at the end of follow-up, and with the Paxinos rat brain atlas (Paxinos & Watson 2006). In the EpiBioS4Rx cohort, we outlined the hippocampi of 15 rats (8 TBI, 7 sham) imaged at 2 d, 9 d, 30 d, and/or 5 months post-injury or sham surgery. In the EPITARGET cohort, we outlined the hippocampi of 6 rats (4 TBI, 2 sham)

Figure 7. Manual annotation of the brain mask in an EpiBioS4Rx rat, MRI scan acquired 150 days after TBI, displaying every annotated sagittal slice. Reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

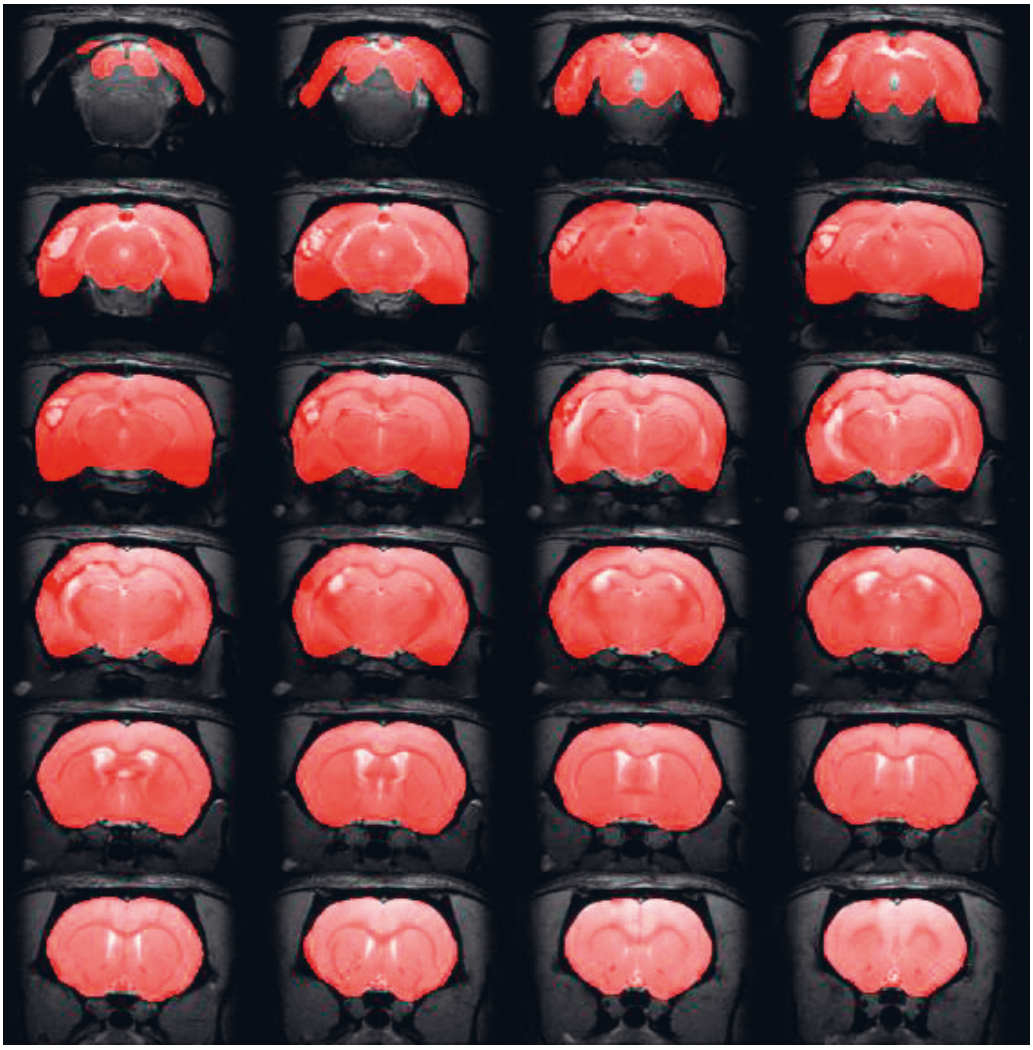

Figure 8. Manual annotation of the brain mask in an EPITARGET rat, MRI scan acquired 21 days after TBI, displaying every annotated coronal slice. Reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

imaged 2 d, 7 d, and/or 21 d after TBI or sham surgery.

4.4 VALIDATION

Each experiment on the train and validation dataset as well as the NeAt dataset was validated according to a 5-fold cross validation (CV) scheme. Experiments on EpiBioS4Rx and EPITARGET utilized a 6-fold CV scheme instead, as while the number of labeled animals was small for these datasets, it was conveniently a multiple of 6 in both cases. Volumes were distributed in each fold according to the individual identity of each animal, preventing the use of the volumes from the validation animals for training. The animals were randomly assigned to each fold once, and the same animals remained assigned to their respective folds through all experiments. To evaluate registration algorithms, to label the test fold the remaining data was utilized as atlases according to the methods outlined in chapter 6. For CNNs, each fold was used as the validation data while the remaining data was used for training.

Charles River For train and validation dataset, this resulted in a training set of 25 or 26 mice and a validation set of 6 or 7 mice in each fold. The test dataset was entirely used as an external test set to evaluate our CNN trained on the train and validation dataset.

MRM NeAt For the MRM NeAt dataset, 5-fold CV resulted in 8 volumes used for training (or as registration atlases) and 2 for testing in each fold.

EpiBioS4Rx and EPITARGET In EpiBioS4Rx each fold contained two animals, conversely in EPITARGET each fold contained only one animal.

4.4.1 Nested cross validation

For MU-Net-R (later described in Chapter 5.2) the results were evaluated by selecting each fold as the testing data of one ensemble of networks, trained using the remaining data. To train each ensemble with early stopping we applied nested 6-fold cross validation: the training set was further randomly divided into 6 folds, using one fold as validation data during the training loop. In this way, we trained an ensemble of six networks, one for each validation fold. The final prediction for the test fold was the majority voting prediction from all networks generated from the same training set. The use of nested CV was necessary to train ensambles with early stopping in the absence of an additional labeled test set. While this strategy can result in overfitting the network on the validation data, it's reasonable to expect this effect to be at least partially countered by

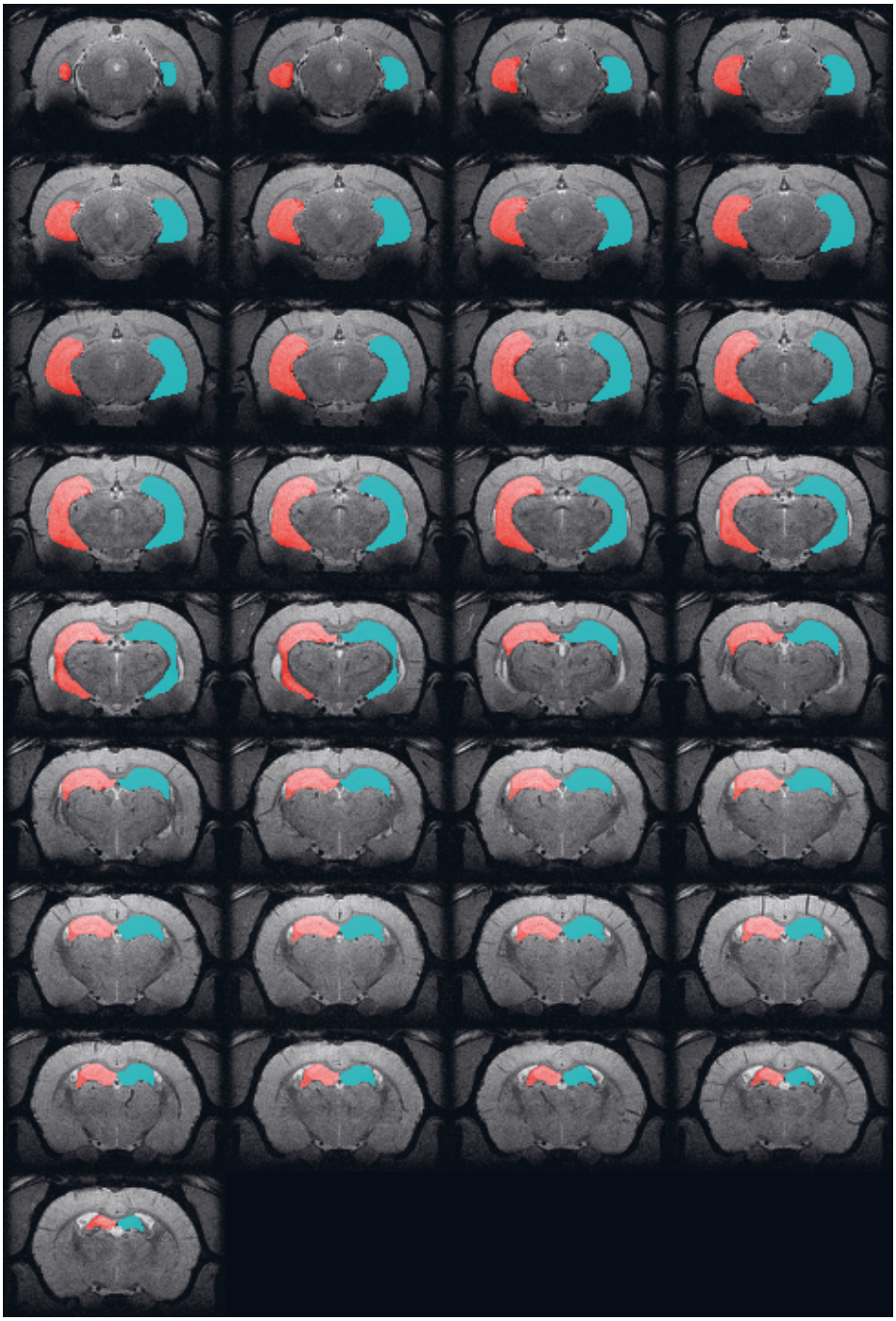

Figure 9. Manual annotation of the hippocampus ipsilateral (red) and contralateral (blue) to the lesion, in an EpiBioS4Rx rat nine days after TBI. Reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

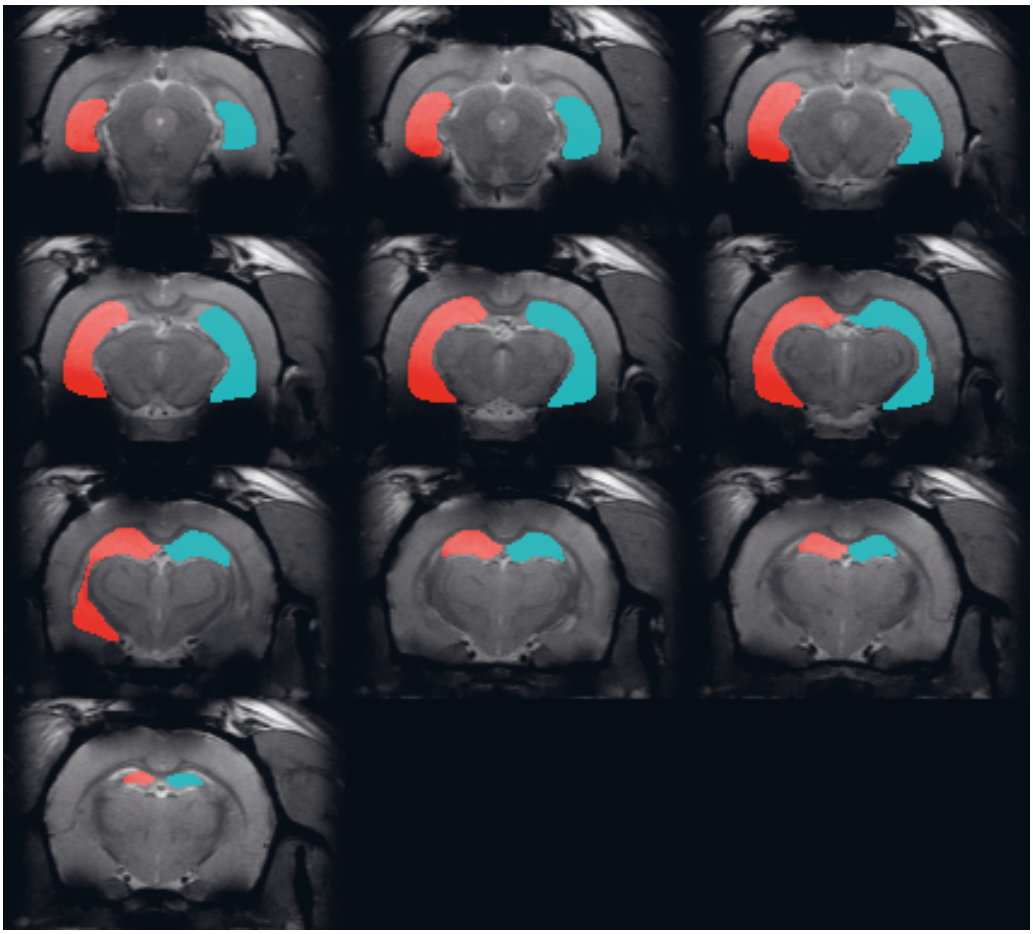

Figure 10. Manual annotation of the hippocampus ipsilateral (red) and contralateral (blue) to the lesion, in an EPITARGET rat two days after TBI. Reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

ensembling. To better evaluate the performance of the network trained on these datasets, the results on new, unlabeled data were also systematically evaluated visually (section 9.5.

An alternative strategy found in literature is to train and test on the mixed individual slices regardless of animal identity. While increasing the variance of the training data would generally improve performance, in this case it would happen at the expense of the training and testing sets not being independent, as discussed by Valverde *et al.* (2021*b*).

4.4.2 Statistical significance

Unless otherwise specified, we used a paired permutation test (Chung *et al.* 2013) to evaluate the significance of differences between the Dice scores obtained by different methods, pairing the Dice scores obtained on the same MRI volumes. The unpaired permutation test was used instead when comparing results obtained on different volumes, for example, when comparing the accuracy of a model on volumes from younger mice with that of the same model on older mice. We performed permutation tests using 100,000 iterations, and considered average differences to be significant when p was smaller than 0.05. The unpaired permutation tests of Dice coefficients between different animal groups were performed by permuting animals (not images) between the two groups.

The use of a paired permutation test is especially indicated here as the different metrics evaluated are usually compared across different segmentation maps of the same MRI volumes. This can lead us to question whether the assumption of independence for more common tests (e. g. the Mann-whitney U test) are verified. Instead, using the paired permutation test this intrinsic pairing can be taken into account.

5 CONVOLUTIONAL NEURAL NETWORKS

Having discussed Convolutional Neural Networks (CNNs) in general in chapter 3, we now have enough elements to discuss the neural networks originally developed in the present work. These are ensembled U-Net-like networks (Ronneberger *et al.* 2015), modified to perform at the same time skull-stripping and region segmentation. We will first see the Multi-task U-Net (MU-Net) architecture in detail, examining the different architectural features compared before selecting one specific architecture. Later, we will describe MU-Net-R, its adaptation to a low-data setting for rat brain segmentation.

5.1 MU-NET

5.1.1 Architectures

MU-Net (Figure 11) presents an encoder-decoder U-Net-like architecture, with each branch articulated in four convolutional blocks. Unlike U-Net, the final block of the decoder branch further bifurcates into two different output maps representing our two tasks sharing the same feature representation, skull-stripping and region segmentation. Each convolutional block on the encoding path is followed by a 2x2 max-pooling layer. The last feature map feeds into the bottleneck layer, a 64 channel 5x5 convolutional layer with batch normalization (Ioffe & Szegedy 2015) connecting the deepest layer of the encoding path with the decoding path.

The decoding path is composed of 4 more blocks alternating one un-pooling layer (Noh *et al.* 2015) and one convolutional block. Un-pooling operations effectively replace up-convolution layers in U-Net without any learnable parameters, while preserving spatial information. These layers operate by simply placing the elements of the un-pooled feature maps in the position of the respective maximum activation from the corresponding pooling operation, and setting the rest to zero. Skip connections concatenate the output of each dense layer in the encoding path with the respective un-pooled feature map of the same size before feeding it as input to the decoding convolutional block.

The output of the last decoding layer acts as the input of two different classification layers, which share the same feature representation up to this point: a 1×1 single channel convolution with a sigmoid activation function, and a 1×1, 5 channels layer followed by a softmax activation function. Respectively, these are the output layers for the skull-stripping task and the region segmentation task.

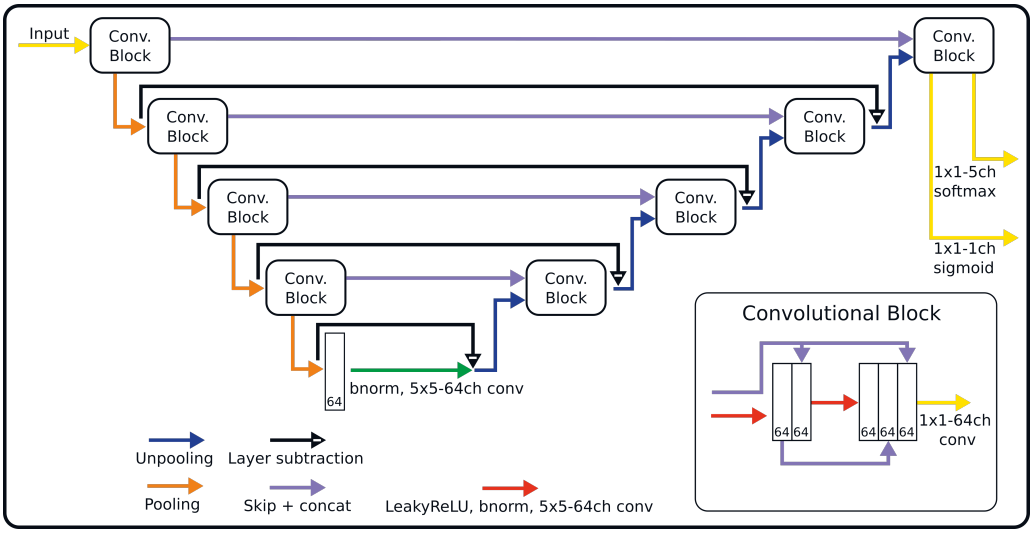

Figure 11. General outline of the architectural features implemented and compared in the networks discussed, varying according to the presence or absence of the in-block dense connections (purple arrows in the convolutional block), presence or absence of the layer subtraction connections (black), and the use of 2D or 3D filters. Image reproduced under CC license (De Feo *et al.* 2021).

Convolutional block Each convolutional block includes 3 convolutional layers preceded by leaky ReLU activation (Maas *et al.* 2013) layers and batch normalization. All 3 convolutions are padded and result in 64 output channels, in analogy with Roy, Conjeti, Navab & Wachinger (2018). The first and second convolutions employ 5×5 filters, while the third uses a 1×1 filter. This becomes especially relevant in the presence of dense connections, acting as a bottleneck for the 64×3 channels of the concatenated inputs and compressing the size of the feature maps.

5.1.2 Architectural variants

We compared several variations to the basic network architecture.

Dense connections In the models including dense connections (Huang *et al.* 2017) we modify each convolutional block by concatenating to the input of each convolution the outputs of the previous convolutions within the same block (Figure 11).

Dual Framing connections Dual framing connections refer to additional skip connections in the Dual Frame U-Net model. Han & Ye (2018) proposed this architecture for computed tomography reconstruction from sparse data based on signal processing arguments to reduce artifacts and improve recovery of high frequency edges. Dual framing connections consist in the subtraction of the input of each convolutional block on the encoding path from the output of the respective convolutional block of the same size on the decoding path, and as such the implementation of these connections does not increase the number of model parameters.

3D implementation A 3D implementation could, in principle, provide better results by taking into account the features of the adjacent slices, whereas a 2D networks evaluates each coronal slice independently. However, the larger number of parameters also increases the risk of overfitting, and the lower resolution in the anterior-posterior axis compared to the in-plane resolution might constitute confounding factors in the presence of 3D pooling operations.

For these reasons, we compared 2D and 3D implementations of our network, using $5 \times 5 \times 5$ filters and $2 \times 2 \times 2$ max-pooling layers, replacing the filters and pooling layers described above. This results in 16008076 and 10286344 parameters for the 3D networks with and without in-block skip connections, respectively. Corresponding 2D networks contain 3297676 and 2087944 parameters, respectively. Thus, opting for a 3D architecture increases the number of parameters by factors of 4.85 and 4.93 as compared to the 2D architectures. The total number of parameters was measured by using the PyTorch instruction `sum(p.numel() for p in model.parameters())`.

5.1.3 Loss function

We optimized a joint loss function L, that is the sum of two Dice loss functions corresponding to the the skull-stripping (L_{SS}) and the region classification task (L_{RS}). Using Equation 16:

$$L = L_{SS} + L_{RS}, \tag{18}$$

$$L_{SS} = -\frac{2\sum_i p(i)g(i)}{\sum_i p^2(i) + \sum_i g^2(i)}, \tag{19}$$

$$L_{RS} = -\sum_{l=1}^{K} \frac{2\sum_i p_l(i)g_l(i)}{\sum_i p_l^2(i) + \sum_i g_l^2(i)}, \tag{20}$$

5.1.4 Training

The networks were implemented using the PyTorch framework and trained with stochastic gradient descent using Adam optimizer (Kingma & Ba 2014) with the default parameters (initial learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and no weight decay) on an NVIDIA GeForce GTX 1080 Graphic Processing Unit (GPU) for up to 12 hours (train and validation) or on an NVIDIA Volta V100 GPU for up to 24 hours (MRM NeAt). Each network was trained with a batch size of one, as larger batch sized would not fit into GPU memory. An alternative to this approach would have been to utilize gradient accumulation, allowing for the combination of gradients from different batches before each gradient descent step. Taking into account that in the case of segmentation each pixel from each slice from the same volume can be considered a (non-independent) sample, I selected a batch size of one rather than incur in the additional computational cost of gradient accumulation.

The hyperparamethers of the models did not undergo any optimization. Besides a large increase in the time required for each network, this would have required either a larger amount of data or a further reduction in the number of training samples, to avoid using the same data to validate both the hyperparameters and the final performance of the network. Given these cosnideration and the increased risk of overfitting, the choice of hyperparameters was based on literature (Kingma & Ba 2014, Ronneberger *et al.* 2015, Huang *et al.* 2017, Han & Ye 2018, Roy, Conjeti, Navab & Wachinger 2018).

Qualitatively, the training pace of 2D and 3D networks was essentially the same, as evidenced in Figure 12.

We augmented the data online each time an image was loaded by scaling the volumes by a factor α randomly drawn from the interval $[0.95, 1.01]$ and rotating them around each axis by a random angle between $-5°$ and $5°$. Scaling factors smaller than one were preferred to decrease memory requirements. Each transformation was applied with 50% probability. To further contain memory requirements, a bounding box was created for each volume using the annotated brain mask as a reference. Each volume was individually normalized to 0 mean and unit variance. Hyperparameters, optimizer and data augmentation scheme were fixed before training ensuring that each architecture would fit into memory, and applied to each network with no additional fine tuning.

5.1.5 Auxiliary bounding-box network

As MU-Net was trained after cropping the volumes to a bounding box, we trained a lighter 2D network to run a first estimate for the brain mask at inference time from the complete volume. This was then used to draw a bounding box around the brain with one voxel margin. This auxiliary network

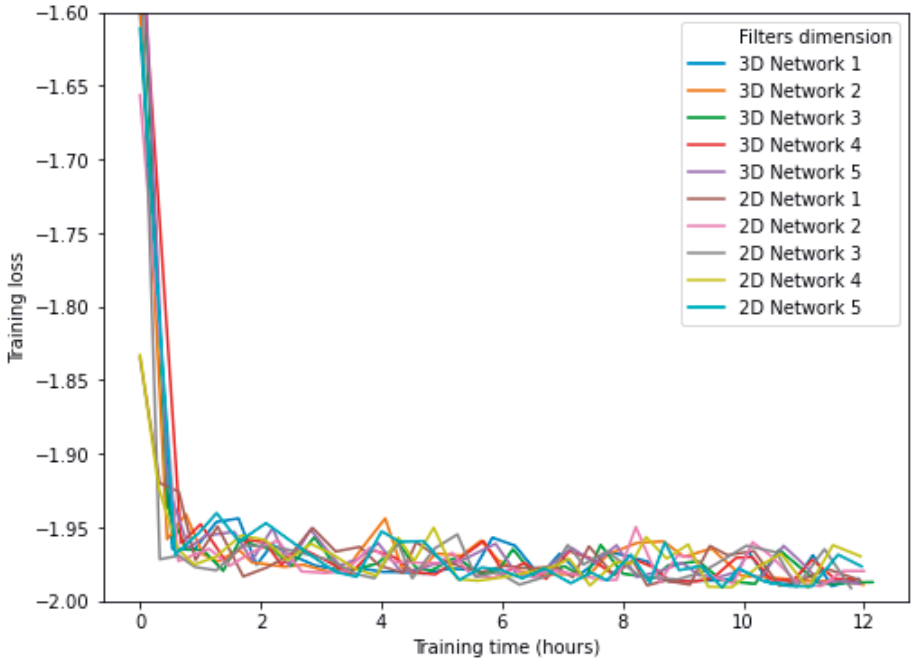

Figure 12. Training loss for 2D and 3D networks (without dense or framing connections) as a function of training time (wall clock), sampled every 30 epocs. Image reproduced under CC license (De Feo *et al.* 2021).

follows exactly the same architecture of MU-Net, omitting any framing or dense connections, and limiting the number of channels to 4, 8, 16 and 32, from the shallowest to the deepest layer. This results in a network with a total number of 122455 trained parameters.

5.1.6 Post-processing

The only post-processing steps applied on the segmentation maps were the filling of holes in the resulting 3D volume, the selection of the largest connected component as the brain mask for the skull-stripping task, and assigning all voxels predicted as non-brain to the background class.

5.2 MU-NET-R

5.2.1 CNN Architecture

The architecture of MU-Net-R (Fig. 13) is based on MU-Net, adapted for the EpiBioS4Rx and EPITARGET datasets reducing the number of parameters for skull-stripping and hippocampus segmentation. Each block consists of three

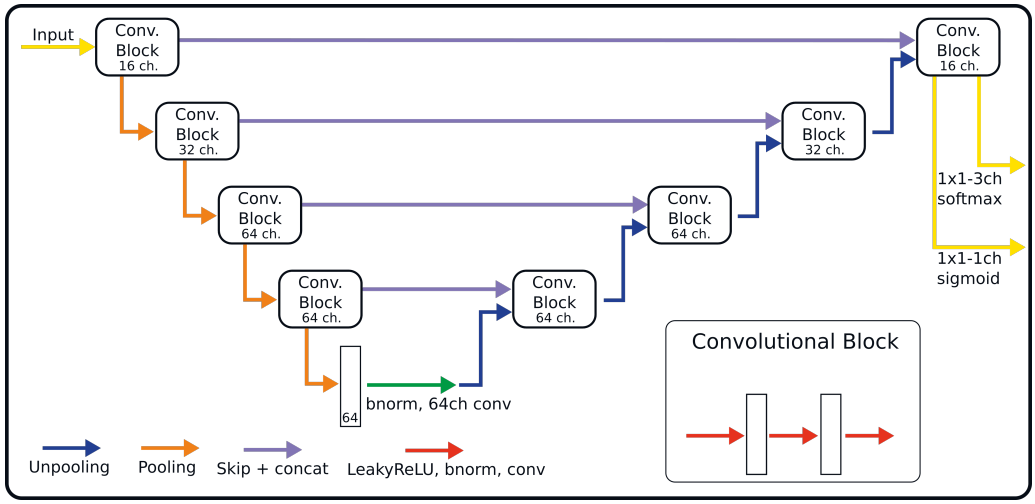

Figure 13. The architecture of MU-Net-R, indicating for each block the number of channels used in each convolution. The size of convolution kernels is $3 \times 3 \times 3$ for the EpiBioS4Rx data, and 3×3 for the EPITARGET data. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

iterations of Leaky ReLU activation (Maas *et al.* 2013), batch normalization (Ioffe & Szegedy 2015) and convolution. From the shallowest to the deepest convolution block, and in contrast with MU-Net each convolution within the block uses 16, 32, 64, and 64 channels, respectively (Fig. 13).

We opted for a different choice regarding the dimension of the filters for each dataset. Since the EPITARGET T_2^* MRI data are highly anisotropic, with higher resolution on coronal slices than in the fronto-caudal direction, we used 2D filters (3×3) in the coronal plane. This choice was based on the MU-Net experiments as well as (Isensee *et al.* 2018), which indicated that a 2D convolutions were preferable for the segmentation of images with anisotropic voxel size. Conversely, in the network trained on isotropic EpiBioS4Rx T_2^* data, we preferred 3D filters ($3 \times 3 \times 3$).

The architecture here described differs from MU-Net in the number of convolution operations, as MU-Net always employs 64-channels convolutions, and in the filter size, whereas MU-Net utilized 5x5 convolutions. These modifications reduce the total number of parameters from 2,087,944 (2D) and 10,286,344 (3D) to respectively 428,436 and 1,125,716. In this way we achieved the same segmentation quality as MU-Net with a lower number of parameters. The comparison with MU-Net and with a single network instead of ensembling is displayed in the ablation studies outlined in Table 6.

Table 6. Ablation tests comparing MU-Net-R, MU-Net, and Non-Ensembled MU-Net-R (NE). Bold numbers indicate a significant statistical difference ($p<0.05$) compared with NE. The architectural choices in MU-Net-R result in statistically equal results compared to MU-Net, with a strong reduction in the number of parameters. Compared to NE, we observe a marked increase in the network's performance as a consequence of ensembling. Networks were compared by changing the architecture only, using the RAdam optimizer, early stopping and the generalized Dice loss.

Method	ROI	Dice	HD95 (mm)	VS	CS	Precision	Recall
MU-Net-R	Ipsi	$\mathbf{0.921 \pm 0.017}$	$\mathbf{0.302 \pm 0.092}$	0.968 ± 0.019	0.974 ± 0.008	$\mathbf{0.935 \pm 0.092}$	0.909 ± 0.098
MU-Net	Ipsi	$\mathbf{0.923 \pm 0.029}$	$\mathbf{0.291 \pm 0.109}$	0.972 ± 0.016	0.980 ± 0.007	0.926 ± 0.037	0.922 ± 0.045
NE	Ipsi	0.904 ± 0.024	0.335 ± 0.100	0.966 ± 0.020	0.979 ± 0.009	0.921 ± 0.036	0.909 ± 0.049
MU-Net-R	Contra	$\mathbf{0.928 \pm 0.011}$	$\mathbf{0.259 \pm 0.072}$	0.971 ± 0.020	0.979 ± 0.007	0.936 ± 0.033	0.921 ± 0.033
MU-Net	Contra	$\mathbf{0.922 \pm 0.011}$	0.264 ± 0.079	0.973 ± 0.019	0.982 ± 0.007	0.935 ± 0.034	0.931 ± 0.030
NE	Contra	0.914 ± 0.013	0.288 ± 0.077	0.972 ± 0.023	0.981 ± 0.008	0.926 ± 0.0355	0.925 ± 0.036

5.2.2 Loss function

The loss function is composed of two terms, referring to the skull-stripping task (L_{SS}) and the hippocampus segmentation task (L_{HC}):

$$L = L_{HC} + L_{SS}. \tag{21}$$

For MU-Net-R the loss was upgraded to the generalized Dice loss (Equation 17) to achieve a better balance between the different categories. Keeping the notation consistent with section 3.6.4:

$$L_{HC} = 1 - 2\frac{\sum_{l=1}^{3} w_l \sum_n g_{ln} p_{ln}}{\sum_{l=1}^{3} w_l \sum_n g_{ln} + p_{ln}} \tag{22}$$

with $w_l = \left(\sum p_{ln}\right)^{-2}$. For the skull-stripping task we used the following Dice loss term:

$$L_{SS} = -\frac{\sum_n g_n p_n}{\sum_n g_n + p_n}. \tag{23}$$

5.2.3 Training

We minimized L with stochastic gradient descent using the RAdam optimizer (Liu *et al.* 2019). RAdam is an optimizer based on Adam (Kingma & Ba 2014) designed to better avoid local optima, obtain more generalizable neural networks, and train in fewer epochs. We utilized RAdam with default

parameters (learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and no weight decay) and a batch size of one, as constrained by the available GPU memory. Networks and training were implemented in PyTorch and ran on a workstation with a GeForce RTX 2080 Ti GPU, 64 GB RAM and an AMD Ryzen 9 3900X 12- Core Processor. MU-Net-R networks were trained for up to 250 epochs, or until the average validation loss did not improve during the last 10 epochs.

During training, data were augmented online: each time an image was loaded, we randomly applied with a 50% probability a scaling transformation by a factor of α, randomly drawn from the interval $[0.95, 1.05]$. Having gained access to a GPU with more available memory, for this experiment the scaling interval is larger than in MU-Net. Each MRI volume was independently normalized to have a mean of zero and unit variance. To avoid interpolation issues with the low-resolution data, and taking into account that while biological variability is important the animal's orientation is strictly controlled during data acquisition, we did not include rotations in our data augmentation scheme.

5.2.4 Post-processing

We applied a simple post-processing procedure to each CNN-generated segmentation. We selected the largest connected component of the MU-Net-R segmentation to represent each segmented region (brain mask and hippocampus) using the `label` function from `scikit-image` (Van der Walt *et al.* 2014). We then filled all holes in this component using `binary_fill_holes` from `scipy.ndimage` (Virtanen, Gommers, Oliphant, Haberland, Reddy, Cournapeau, Burovski, Peterson, Weckesser, Bright *et al.* 2020).

6 REGISTRATION-BASED SEGMENTATION

In this chapter we will see, for each dataset, the baseline reference for the comparison of Convolutional Neural Network (CNN) segmentation with state-of-the-art single- and multi-atlas registration. To provide a reliable and representative comparison with the proposed CNN, these have been fine-tuned for each dataset (Charles River (CR), NeAt, and the University of Eastern Finland (UEF) data, as described in chapter 4) using freely-accessible registration and segmentation tools.

6.1 REGISTRATION

6.1.1 CR and NeAt data

The registrations were performed as follows: before registration, each volume underwent non-parametric N3 bias field correction (Sled *et al.* 1998) implemented within the ANTS toolset (Avants *et al.* 2009). Taking each volume as reference, all other volumes were then registered with an affine transformation using FSL FLIRT (Jenkinson & Smith 2001) and then nonlinearly registered via FSL FNIRT (Jenkinson *et al.* 2012, Andersson *et al.* 2007) with the aid of the manually drawn brain mask.

We used correlation ratio (`corratio`) as the cost function in FLIRT and FNIRT. We used the default FLIRT and FNIRT parameters with the following exceptions. The search range of angles in FLIRT was $[-70°, 70°]$ instead of the default $[-90°, 90°]$, because the orientations of the volumes were similar. In FNIRT, we used spline interpolation instead of the default linear interpolation.

6.1.2 UEF data

For this dataset we utilized the registered volumes generated by Eppu Manninen during our collaboration, using Advanced Normalization Tools (Avants, Tustison, Song, Cook, Klein & Gee 2011) to facilitate the transfer of manual segmentations of the hippocampus to other brains with single- and multi-atlas approaches. Before image registration, the images were skull-stripped using FMRIB Software Library's Brain Extraction Tool (FSL BET (Smith 2002*b*)). We selected FSL BET for this step instead of using the masks generated by the CNN to keep the two pipelines completely independent, and ensure registration-based methods would be representative of typical registration-based results used in preclinical research. The masked images were then used for image registration. The brain masks for the EPITARGET dataset computed

using FSL BET included marked amounts of non-brain tissue associated with the experimental traumatic brain injury, which resulted in inaccurate image registrations. To improve the brain masks, we first registered the images to one of the brain images using rigid-body and affine transformations. The FSL BET brain mask of that image was then manually refined and transformed to the rest of the brain images, resulting in more accurate brain masks and registrations.

Image registration between a template brain and a target brain volume included the computation of a rigid-body transformation, an affine transformation, and a Symmetric image Normalization (SyN) transformation. We used global correlation as the similarity metric for the rigid-body and affine transformations and neighborhood cross-correlation for the SyN transformation. The computed transforms were then applied to the template brain's sum-over-echoes T_2^*-weighted image as well as its manually labeled hippocampi. All operations described in this section were performed on a 6-core AMD Ryzen 5 5600X processor.

EpiBioS4Rx data The 12 animals in the EpiBioS4Rx dataset were divided into 6 groups of 2 animals. The brain of each animal was registered to the brains of other animals that did not belong to the same group. Registrations were performed within each time-point. Thus, the brain of each of the 12 animals was registered to 10 other brains at each of the 4 time-points, which would have resulted in 480 image registrations. However, images for one brain at two time-points were missing, reducing the total to 440 image registrations.

EPITARGET data The brain masks for the EPITARGET dataset computed using FSL BET included significant amounts of non-brain tissue associated with the experimental traumatic brain injury, which resulted in inaccurate image registrations. To create better brain masks, the images were first registered to one of the brain images using rigid-body and affine transformations. The FSL BET brain mask of that image was then manually refined and transformed to the rest of the brain images, resulting in more accurate brain masks and registrations. The brain of each of the 6 animals was registered to the 5 other brains at each of the 3 time-points, resulting in 90 image registrations.

6.1.3 Segmentation

For each template registered to a target brain, we applied the same transforms to the label map of the template, using nearest neighbor interpolation. Each measure reported for single-atlas segmentation in this work was an average

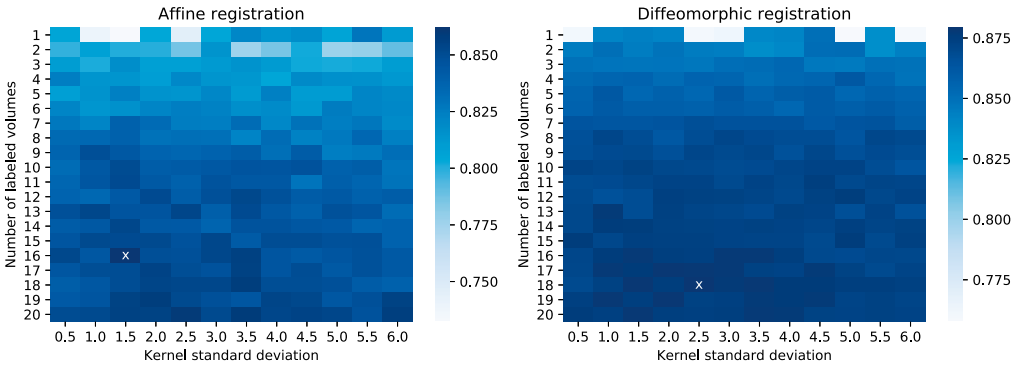

Figure 14. Grid search results for STEPS parameters for volumes aligned using affine or diffeomorphic registration. A white x marks the combination of parameters resulting in the highest average Dice overlap. Image reproduced under CC license (De Feo *et al.* 2021).

between that of each individual single-atlas segmentation map for the same target brain. Majority voting instead refers to choosing the most commonly occurring label among all registered atlases for each voxel.

6.2 STEPS

STEPS is a state of the art label fusion algorithm to combine multiple registered templates to label a target volume (Cardoso *et al.* 2013). It takes into account the local and global image matching, combining an expectation-maximization approach with Markov Random Fields to improve on the segmentation based on the quality of the registration itself. We applied the STEPS implementation distributed with NiftySeg (Cardoso *et al.* 2012, 2013).

STEPS depends on the number of templates employed and the standard deviation of its Gaussian kernel. We performed a grid search on the CR dataset to select the optimal parameters, randomly selecting 10 volumes and labeling them using STEPS. We sampled the standard deviation of the Gaussian kernels between 0.5 and 6 with a stride of 0.5, and the number of templates ranged between 1 and 20 randomly selected volumes. This same process has been performed both using diffeomorphic registration and using affine registration only (Figure 14), selecting 16 templates and kernel standard deviation of 1.5 for the diffeomorphic case, and 18 templates with kernel standard deviation of 2.5 for the affinely registered volumes. Exploring both grids required in total 287 hours.

Each volume was then segmented using these parameters, randomly

selecting an appropriate number of registered animals as atlases for the STEPS algorithm as emerged from the grid search outlined above. We repeated this procedure randomly selecting the same number of templates from mice of the same age only, and given the higher performance of this approach we kept using animals of the same age for experiments on the UEF data. We further utilized the same parameters for the UEF data without repeating the grid search, given the smaller dataset size.

When evaluating STEPS on MRM NeAt dataset, we used scripts provided by Ma *et al.* (2014) at `https://github.com/dancebean/multi-atlas-segmentation` as this implementation is optimized using this dataset.

The here described computations were executed on a workstation equipped with a 6-core, 12-thread Intel Core i7-8700K CPU running at 3.70GHz. To accelerate the computations generating several intermediate file outputs, we used a RAMdisk to reduce the number of the disk operations. For the NeAt dataset, computations were performed on a 12-core, 24-thread AMD Ryzen 9 3900X Processor.

7 HAND-CRAFTED GEOMETRIC FEATURES

The EPITARGET and EpiBioS4Rx datasets described in section 4.3 include two cohorts of rats with Traumatic Brain Injury (TBI) and rats who underwent sham surgery, with a portion of the TBI rats displaying post-traumatic epilepsy. To discriminate between epileptic and non-epiletic animals, and between TBI and sham-surgery animals, we extracted a set of hand-crafted features using an original method developed specifically for this purpose. The present chapter only deals with the technical details of the feature extraction method, and for a deeper discussion of the biological motivation behind these experiments we refer the reader to chapter 10.

To define these features we build on the segmentation masks generated by MU-Net-R. Using the brain mask we defined a system of reference $\mathcal{F}$ specific for each brain. Then we defined for each hippocampus an "interpolating plane" $\boldsymbol{P^1}$ and an "inclination plane" $\boldsymbol{P^2}$ (Figure 15). $\boldsymbol{P^1}$ was defined as the plane minimizing the sum of distances from hippocampus voxels to it and $\boldsymbol{P^2}$ as a plane characterizing the inclination of the hippocampus. The purpose of these planes is to describe the relative three-dimensional positioning of the hippocampi with respect to each other and with respect to the brain. This is encoded by 39 parameters, extracted as described in the following sections.

7.1 BRAIN-SPECIFIC FRAME OF REFERENCE

As small inconsistencies in the positioning of each rat during each scan would constitute a source of noise, we identified a frame of reference for each image I, defined on domain D, based on the segmented brain mask. This new reference $\mathcal{F}$ was built to have the unit vector $\hat{\boldsymbol{y}}$ directed in the rostral direction, $\hat{\boldsymbol{z}}$ in the dorsal direction, and defining $\hat{\boldsymbol{x}} = \hat{\boldsymbol{y}} \times \hat{\boldsymbol{z}}$.

We begin by constructing a plane $\boldsymbol{R}$ defined by a point $\boldsymbol{r_p}$ and the normal vector $\boldsymbol{r_n}$ as follows. Let M be the brain mask, defined as the set of brain voxels. We found a plane maximizing its intersection with M by solving the following optimization problem:

$$\boldsymbol{R}(\boldsymbol{r_n}, \boldsymbol{r_p}) = \arg\max_{S(\boldsymbol{s_n}, \boldsymbol{s_p})} \left| \boldsymbol{S}(\boldsymbol{s_n}, \boldsymbol{s_p}) \cap M \right| \tag{24}$$

under the constraints

$$\boldsymbol{r_n} \in (-0.5, 0.5) \times (-0.5, 0.5) \times (0.5, 1)$$
$$|\boldsymbol{r_n}| = 1$$

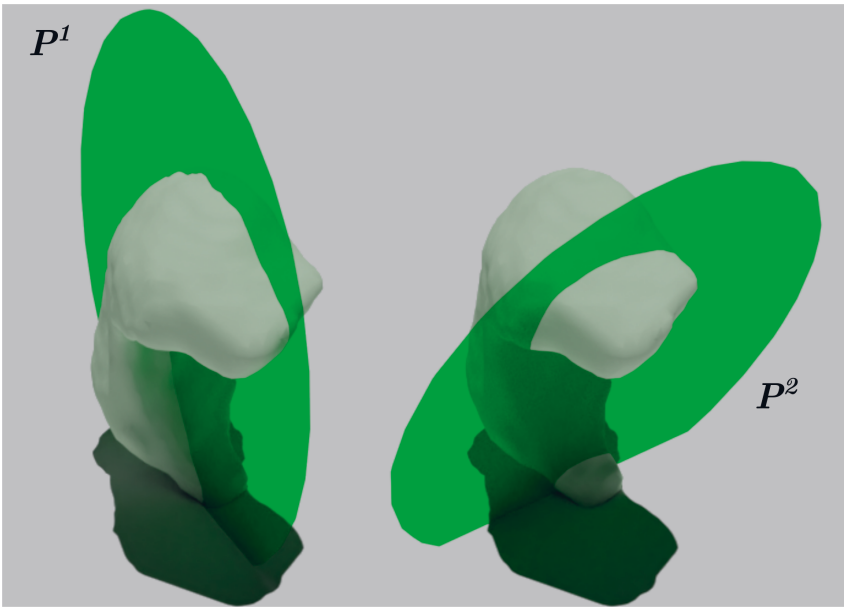

Figure 15. Visualization of the P^1 and P^2 planes in the right hippocampus of a sham-operated rat imaged at 30 days after the sham-operation (EpiBioS4Rx cohort). Image reproduced under CC license (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

$$r_p \in D.$$

For any solution to the above equation there are two equally valid representations, corresponding to the two opposite orientations of the normal vector. To avoid any ambiguity we always chose the vector r_n to be oriented towards the dorsal direction. We chose the constraints for the optimization problem by observing that while r_n would not be parallel to the vertical axis in the reference frame of the MRI scan, it would also be oriented in roughly the same direction. The initial values of r_p was set to the center of mass of brain mask, and that of r_n to the vertical unit vector.

We then defined an orthonormal basis $\mathcal{F}$ specific for each brain scan using the point r_p as the origin and three orthonormal vectors $\hat{x}$, $\hat{y}$, $\hat{z}$. To identify these vectors we first set $\hat{z} = r_n$. Next, to define a $\hat{y}$ vector, we selected a vector collinear to the line L that minimizes the distance from every point in $R \cap M$. With δ as the Euclidean distance of a point from a line:

$$L = \arg\min_{T \subset R} \int_{x \in R \cap M} \delta(x, T) \tag{25}$$

To resolve any possible ambiguity, $\hat{\boldsymbol{y}}$ was chosen to have a positive component in the frontal direction, and optimized under the constraint $|\hat{\boldsymbol{y}}| = 1$. Furthermore, $\hat{\boldsymbol{y}}$ lies on the plane $\boldsymbol{R}$, and thus $\hat{\boldsymbol{y}} \perp \hat{\boldsymbol{z}}$. Lastly, $\hat{\boldsymbol{x}}$ was obtained as the vector product $\hat{\boldsymbol{x}} = \hat{\boldsymbol{y}} \times \hat{\boldsymbol{z}}$, completing the basis.

All optimization steps were implemented using the trust-constr algorithm (Conn *et al.* 2000) in SciPy (Virtanen, Gommers, Oliphant, Haberland, Reddy, Cournapeau, Burovski, Peterson, Weckesser, Bright, van der Walt, Brett, Wilson, Millman, Mayorov, Nelson, Jones, Kern, Larson, Carey, Polat, Feng, Moore, VanderPlas, Laxalde, Perktold, Cimrman, Henriksen, Quintero, Harris, Archibald, Ribeiro, Pedregosa, van Mulbregt & SciPy 1.0 Contributors 2020).

7.2 POSITION AND ORIENTATION

Interpolating plane $\boldsymbol{P^1}$ We defined the plane $\boldsymbol{P^1}$ as an interpolating plane, minimizing the sum of distances of each voxel in the hippocampus segmentation mask H from the plane. The purpose of this plane is to describe the position and orientation in space of the hippocampus with respect to the contralateral hippocampus and the brain in general.

We fitted the best interpolating plane using singular value decomposition, implemented with scikit-spatial (`https://scikit-spatial.readthedocs.io/`), and described the plane through the point $\boldsymbol{p_p^1}$ and the normal vector $\boldsymbol{p_n^1}$. $\boldsymbol{p_p^1}$ corresponds to the center of mass of the segmentation mask, and $\boldsymbol{p_n^1}$ is selected for both hippocampi so that $p^1_{ny} < 0$.

Inclination plane $\boldsymbol{P^2}$ The plane $\boldsymbol{P^1}$ cannot capture any change in orientation parallel to the plane itself, and in fact it's trivial to observe that any rotation of H around $\boldsymbol{p_n^1}$ does not affect $\boldsymbol{P^1}$. To capture complementary information to that provided by $\boldsymbol{P^1}$ we defined $\boldsymbol{P^2}$ by the following construction, in reference to Figure 16.

We extracted the skeleton of the hippocampus segmentation mask H using (Zhang & Suen 1984). We identified two points at the upper and lower extremes of the skeleton, respectively, $\boldsymbol{a}$ and $\boldsymbol{b}$. These were extracted by identifying the most rostral points in the upped and lower halves of the skeleton. Next, we identified $\boldsymbol{c}$ as the midpoint between $\boldsymbol{a}$ and $\boldsymbol{b}$: $\boldsymbol{c} = (\boldsymbol{a} + \boldsymbol{b})/2$. The plane $\boldsymbol{P^2}$ is characterized by the point $\boldsymbol{p_p^2} = \boldsymbol{c}$ and a unit vector $\boldsymbol{p_n^2}$ orthogonal to the segment $\overline{\boldsymbol{ab}}$ and lying on the plane including $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{m}$, defined as the center of mass of H. To find $\boldsymbol{p_n^2}$ we first defined $\boldsymbol{v_1}$ as the vector pointing from $\boldsymbol{c}$ to $\boldsymbol{m}$: $\boldsymbol{v_1} = \boldsymbol{m} - \boldsymbol{c}$. We further defined $\boldsymbol{v_2}$ as the vector pointing from $\boldsymbol{c}$ to $\boldsymbol{a}$:

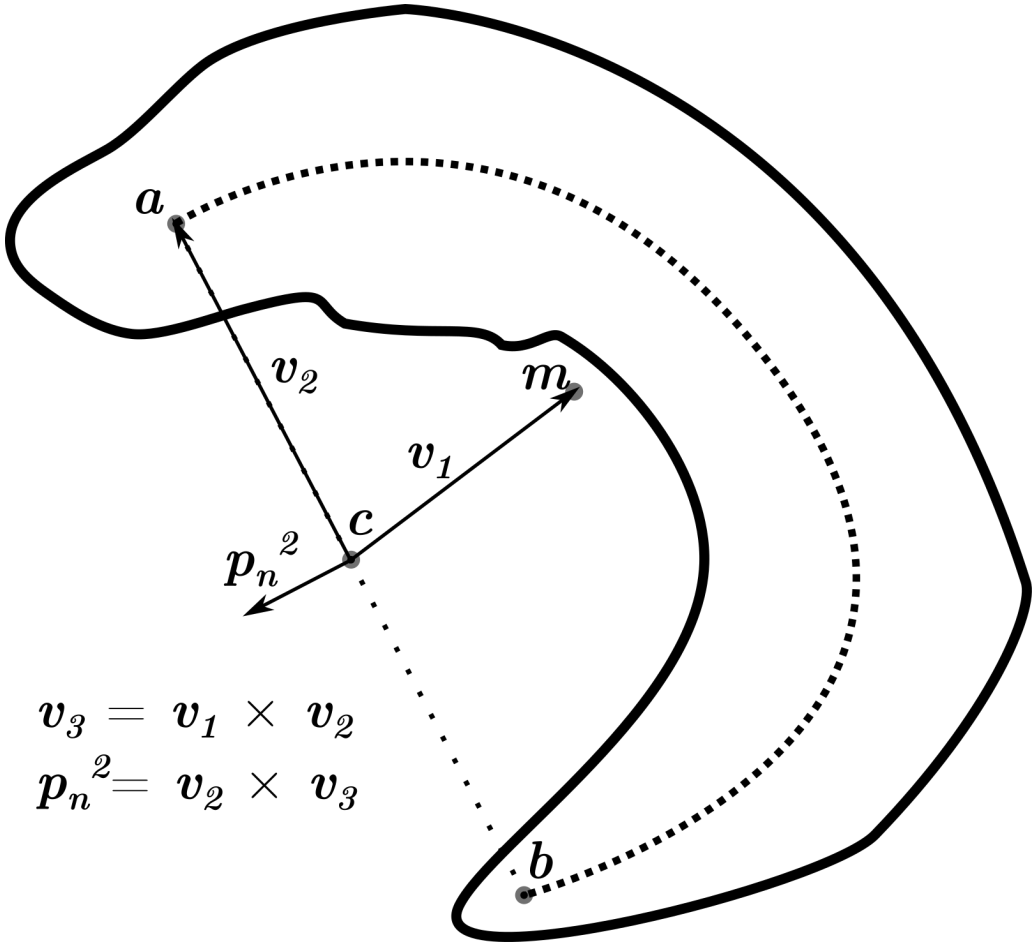

Figure 16. Geometric construction for the P^2 plane for the segmentation mask of the hippocampus (H). a and b indicate the extremes of the skeleton of H and M its center of mass. P^2 was defined by the point $p_p^2 = c$ and the vector p_n^2. Image reproduced under CC license (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

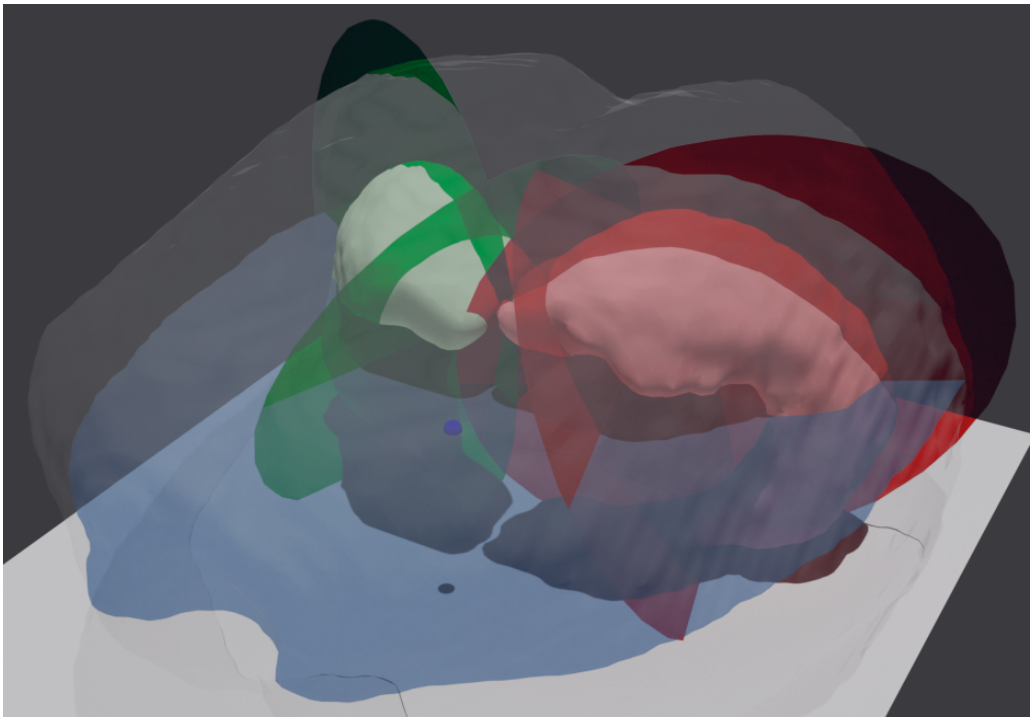

Figure 17. A rendering of the geometric construction for a sham-operated rat imaged at 30 d after TBI (EpiBioS4Rx cohort). Red indicates the left hippocampus and its $\boldsymbol{P^1}$ and $\boldsymbol{P^2}$ planes, green the right hippocampus and its planes, and blue the reference plane. A blue dot indicates the origin of the reference frame. To aid visualization, a plane horizontal in the reference system of the scan is white. The brain mask is transparent. Image reproduced under CC license (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

$\boldsymbol{v_2} = \boldsymbol{a} - \boldsymbol{c}$. Then, $\boldsymbol{v_3}$ can be expressed as the vector product $\boldsymbol{v_3} = \boldsymbol{v_1} \times \boldsymbol{v_2}$. Finally, we defined $\boldsymbol{p_n^2} = \boldsymbol{v_2} \times \boldsymbol{v_3}$.

This construction is equivalent to defining $\boldsymbol{P^2}$ as the plane containing points $\boldsymbol{a}$ and $\boldsymbol{b}$ that has a maximal sum of distances from voxels in H, and requiring that $\boldsymbol{P^2}$ is characterized by a ventrally-oriented normal vector.

7.3 PARAMETERS

Combining the elements described up to this point for both hippocampi we have the construction displayed in Figure 17. Using this construction for each scan we extracted the following parameters. From both hippocampi, we included the volume for each hippocampus relative to the brain mask v_{ipsi} and v_{contra}, the

Table 7. Parameters extracted from each scan. The hemispheric parameters were defined for each hippocampus separately. v_{ipsi} and v_{contra} refer to the ipsilateral (left) and contralateral hippocampus, respectively. Hemisphere independent parameters were defined for each scan, combining information from the two hippocampi.

Parameter	Symbol
Hemispheric parameters	
Relative volumes	v (v_{ipsi}, v_{contra})
P^1 position components	$p^1_{px}, p^1_{py}, p^1_{pz}$
P^2 position components	$p^2_{px}, p^2_{py}, p^2_{pz}$
P^1 normal components	$p^1_{nx}, p^1_{ny}, p^1_{nz}$
P^2 normal components	$p^2_{nx}, p^2_{ny}, p^2_{nz}$
P^1 distance from reference	$\lvert\boldsymbol{p^1_p}\rvert$
P^2 distance from reference	$\lvert\boldsymbol{p^2_p}\rvert$
P^1 dihedral angle with R	$\theta_{1,R}$
P^2 dihedral angle with R	$\theta_{2,R}$
P^2 dihedral angle with P^1	$\theta_{1,2}$
Hemisphere independent	
Angle between P^1 planes	$\theta_{1,1}$
Angle between P^2 planes	$\theta_{2,2}$
Interhippocampal volume ratio	v_r

components of vectors $\boldsymbol{p^1_n}$, $\boldsymbol{p^2_n}$, $\boldsymbol{p^1_p}$, $\boldsymbol{p^2_p}$, and the dihedral angles of $\boldsymbol{P^1}$ and $\boldsymbol{P^2}$ with $\boldsymbol{R}$, defined as $\theta_{1,R}$ and $\theta_{2,R}$. We further include the dihedral angle between the ipsilateral and contralateral $\boldsymbol{P^1}$ planes $\theta_{1,1}$, the angle between the ipsilateral and contralateral $\boldsymbol{P^2}$ planes $\theta_{2,2}$, and the dihedral angles between $\boldsymbol{P^1}$ and $\boldsymbol{P^2}$ for each hippocampus, indicated as $\theta_{1,2}$. Lastly, we included the volume ratio between the ipsilateral and the contralateral hippocampi $v_r = v_{ipsi}/v_{contra}$, and the vector norms $|\boldsymbol{p^1_p}|$ and $|\boldsymbol{p^2_p}|$, indicating the distances between the points characterizing each plane and the origin.

7.4 DATA ANALYSIS

Statistical analysis of the hippocampal volumes Using the trained CNNs, we labeled every MRI volume in our datasets (220 for EpiBioS4Rx and 424 for EPITARGET). As a demonstration of the applicability of the segmentation, we studied the effects of TBI on hippocampal volume through time and across both hippocampi using a repeated measures linear model, implemented using the linear mixed model function in IBM SPSS Statistics for Windows, version 26.0 (SPSS Inc., Chicago, IL, United States). Every variable was considered as a fixed effect and we assumed a diagonal covariance structure of the error term. Let t indicate the time point in days as a scalar variable, R be defined such so that $R = 1$ indicates the ipsilateral hippocampus and $R = 0$ the contralateral hippocampus. Additionally, let $B = 1$ indicate the presence of TBI, with $B = 0$ indicating sham animals, and let E be the error term. Then, our linear model for the volume V can be written as:

$$V = \alpha + \beta_t t + \beta_R R + \beta_B B + \beta_{tR} tR + \beta_{tB} tB + \beta_{RB} RB + E, \qquad (26)$$

where α, β_i are parameters of the model.

Mass univariate analysis We studied the dependency of each parameter on the timepoint and on either the presence of TBI, or according to the TBI+ (epileptic) and TBI- (non-epileptic) categories. We performed this by applying a repeated measures 2-way ANOVA, implemented in Python using the `pingouing` library (Vallat 2018), with timepoint and lesion as within-subject factors. The main effect p-values were corrected using Greenhouse-Geisser correction (Greenhouse & Geisser 1959), and all p-values were further corrected for multiple comparisons using Bonferroni's correction.

Classification We classified the scans in two binary tasks: TBI *vs.* sham, and TBI+ *vs.* TBI-. We trained random forest classifiers (Breiman 2001), implemented using `scikit-learn` (Pedregosa *et al.* 2011). To prevent overfitting, we only used trees with a depth of one (i.e. stumps), applied balanced weights to compensate for the class imbalance and did not use bootstrapping. We trained forests of 1000 trees per classifier and optimized for the Gini impurity. We did not run any hyperparameter optimization to prevent overfitting for our small training set.

8 MU-NET

MU-Net is the architecture here proposed for the segmentation and skull-stripping of small animal brain MRI. Using the train and validation dataset of WT mice, we compared the performance of different network architectures to decide which architectural features to include. Furthermore, we compared MU-Net with multi-atlas segmentation on both the CR data and the MRM NeAt dataset, and evaluated the impact of mouse age on the accuracy of our segmentation maps. The experiments reported in Sections 8.1, 8.2, and 8.3 are based on 5-fold CV on the train and validation set, and experiments in Section 8.4 on 5-fold CV on the MRM NeAt dataset. Finally, in section 8.5, we tested MU-Net trained on train and validation set on an independent test set that included 1,782 MRI volumes from 817 mice, including WT and Huntington disease mice.

8.1 ARCHITECTURE COMPARISON

We compared the performance of different networks trained with and without dense connections and dual framing connections, in both 2D and 3D implementations.

As shown in Table 8, all MU-Nets achieved Dice scores comparable to or higher than the typical inter-rater variability of manual segmentation in the mouse brain (from 0.80 to 0.90 (Ali *et al.* 2005)). The skull-stripping task achieved an excellent Dice score of 0.984. The ventricles were characterized by the lowest segmentation performance (average Dice score 0.907), while the cortex displayed the highest average overlap with the ground truth (0.966).

The network displaying the highest average Dice scores was the simplest one, including no in-block skip connections nor framing connections, and using 2D convolutions. The accuracy of this network was significantly higher than the accuracy of other all other 2D networks ($p < 0.00003$). Because of its excellent performance and simplicity this network is our choice for the MU-Net architecture, which is the architecture we used for all experiments detailed in sections 8.2-8.3.

The choice between 2D and 3D architectures was the most important factor in increasing performance, resulting in a marked increase in mean Dice scores for both tasks ($p < 0.00001$) between all 2D networks compared to the 3D ones. We further compared MU-Net with one featuring less channels per filter (49, 49, 50, 50, from the shallowest to the deepest convolutional block) to match the number of parameters to the number of parameters of the simplest 2D network. We registered a slightly (but not significantly, $p =$

Table 8. CNN and STEPS accuracies measured using Dice coefficient across different methodological choices. Cross-validation results on the train and validation dataset. Listed values are the average validation Dice scores between automatic and manual segmentation ± standard deviations of these Dice scores in 5-fold CV. ROI mean column refers to the mean Dice coefficient of the cortex, the hippocampi, the ventricles and the striati. SC and FC indicate the presence of skip connection and framing connections. MU-Net results are displayed in the first row. STEPS refers to STEPS using randomly selected templates; STEPS* refers to STEPS runs using randomly selecting mice of the same age only; affine indicates that only affine registration was used, whereas diffeo indicates this was followed by a diffeomorphic registration step; Majority voting refers to the selection of the most occuring label after diffeomorphic registration; 3DConv 2DPool: network featuring no in-block skip connections or framing connections, with 3D filtering and 2D pooling in the coronal plane; 2D SLP: 2D network with in-block skip connections and a limited number of parameters; 2D +N3: 2D network trained on data bias-corrected using the N3 algorithm. Boldface characters indicate the best performing network, achieving significantly higher Dice scores than all other networks for that ROI.

Dim	SC	FC	Brain mask	Cortex	Hippocampi	Ventricles	Striati	ROI mean
2D			**0.984±0.005**	**0.966±0.009**	**0.925±0.017**	**0.907±0.020**	**0.939±0.010**	**0.935±0.026**
2D	x	x	**0.984±0.006**	0.963±0.010	**0.924±0.016**	0.905±0.022	0.937±0.009	0.932±0.026
2D		x	**0.984±0.006**	0.963±0.011	**0.924±0.017**	0.905±0.022	0.938±0.009	0.932±0.026
2D	x		**0.984±0.005**	0.964±0.011	0.923±0.018	0.905±0.024	0.937±0.010	0.932±0.027
3D	x	x	0.982±0.007	0.956±0.016	0.914±0.033	0.900±0.025	0.926±0.045	0.924±0.038
3D		x	0.982±0.007	0.958±0.016	0.916±0.032	0.900±0.025	0.928±0.029	0.925±0.034
3D	x		0.982±0.006	0.957±0.016	0.913±0.041	0.899±0.028	0.926±0.042	0.924±0.040
3D			0.982±0.007	0.957±0.013	0.916±0.033	0.899±0.026	0.926±0.039	0.924±0.036
3DConv 2DPool			0.983±0.006	0.961±0.010	0.919±0.026	0.902±0.026	0.934±0.014	0.929±0.030
2D SLP			**0.984±0.005**	**0.965±0.009**	**0.924±0.016**	**0.907±0.021**	**0.939±0.010**	**0.934±0.026**
2D + N3			**0.984±0.005**	**0.965±0.009**	**0.924±0.020**	**0.907±0.020**	**0.939±0.009**	**0.934±0.026**
STEPS (affine)			\	0.920±0.058	0.827±0.079	0.761±0.090	0.873±0.062	0.845±0.093
STEPS (diffeo)			\	0.948±0.036	0.844±0.048	0.812±0.090	0.871±0.045	0.869±0.070
STEPS* (affine)			\	0.936 ± 0.013	0.831 ± 0.029	0.781 ± 0.049	0.887 ± 0.019	0.859 ± 0.066
STEPS* (diffeo)			\	0.954 ± 0.009	0.848 ± 0.025	0.826 ± 0.039	0.885 ± 0.016	0.879 ± 0.055
Majority Voting			\	0.889±0.179	0.780±0.232	0.677±0.208	0.816±0.245	0.791±0.230

0.077) lower accuracy compared to MU-Net, indicated as 2D SLP in Table 8.

To test whether the increased performance of 2D architectures compared to the 3D implementation depended on the reduced number of parameters or on an excessive loss of information when pooling in the anterior-posterior direction, we trained a network using 3D filters while limiting pooling operations to the coronal plane. This network achieved a segmentation accuracy in between the 3D and 2D implementations (Table 8), suggesting that both above mentioned aspects were relevant in increasing the algorithm's performance.

We studied the effect of bias field correction to the performance of MU-Net training it on images without bias-correction, and separately, on N3 bias-corrected MR images (Sled *et al.* 1998). When implemented, the N3 bias correction was always performed as the first step, before the rest of the pipeline. The validation accuracy achieved with bias correction was indistinguishable from the accuracy of MU-Net trained without bias correction (see Table 8).

8.2 AGE STRATIFIED TRAINING SETS

We evaluated the performance of MU-Net when restricting the training set to mice of a specific age. Networks trained on data from mice of 12, 16 and 32 weeks achieved higher accuracy, both on their respective validation set and the overall ground truth, compared to the networks trained on 5 weeks mice ($p < 0.00001$). As shown in Fig. 18, all networks trained on one specific age displayed a statistically significant ($p < 0.05$, unpaired) decrease in mean accuracy when validated on animals of a different age. While the difference between young and old mice may not be strongly pronounced to the human eye (Fig. 19) this difference was highest between the 5 weeks data and the other datasets.

Limiting the training data to one specific age implies that these networks were trained only on a quarter of the data used to train the networks in section 8.1. Irrespective of that, these networks still achieved average Dice score on the mixed-age validation dataset comparable with the accuracy of manual segmentation. The worst performing CNN was the network trained on 5 weeks old mice. Training on the 12, 16 and 32 weeks data and validating on mice of the same age, we observed Dice scores comparable with the overall performance of MU-Net trained on the entire dataset ($p > 0.15$, unpaired). However, we measured a lower overall performance when including mice of all ages in the validation data ($p < 0.00001$), slightly overfitting for each specific age.

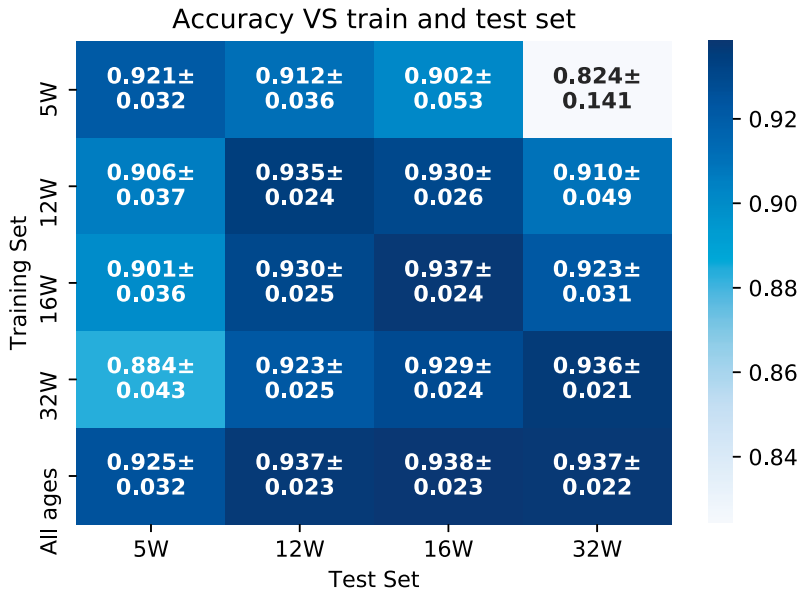

Figure 18. Mean accuracy ± standard deviation for the average accuracy of MU-Net across all ROIs, trained and evaluated on different datasets according to mouse age. Networks exclusively trained on older animals achieved lower accuracy when attempting to generalize to the youngest animals, and vice-versa. Image reproduced under CC license (De Feo *et al.* 2021).

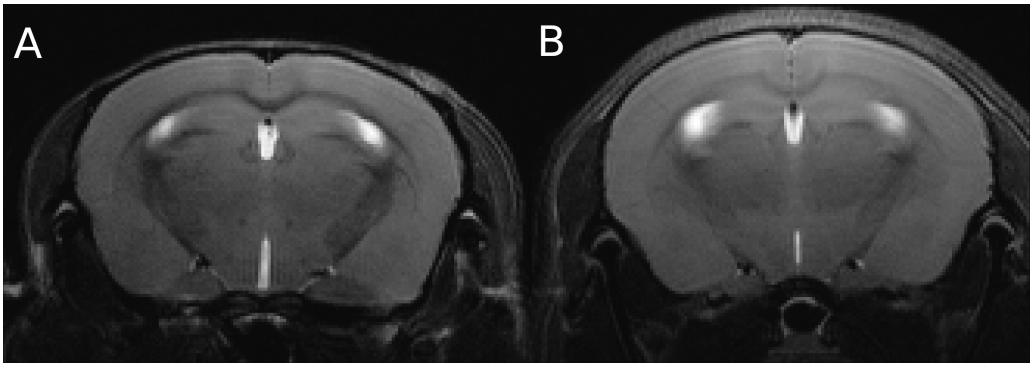

Figure 19. Comparison between two mouse brain slices from mice of A. 4 weeks and B. 60 weeks extracted from the test dataset. While the difference between the two timepoints may not be very significant to the human eye, age differences between the training and testing data resulted in a significant reduction in Dice scores.

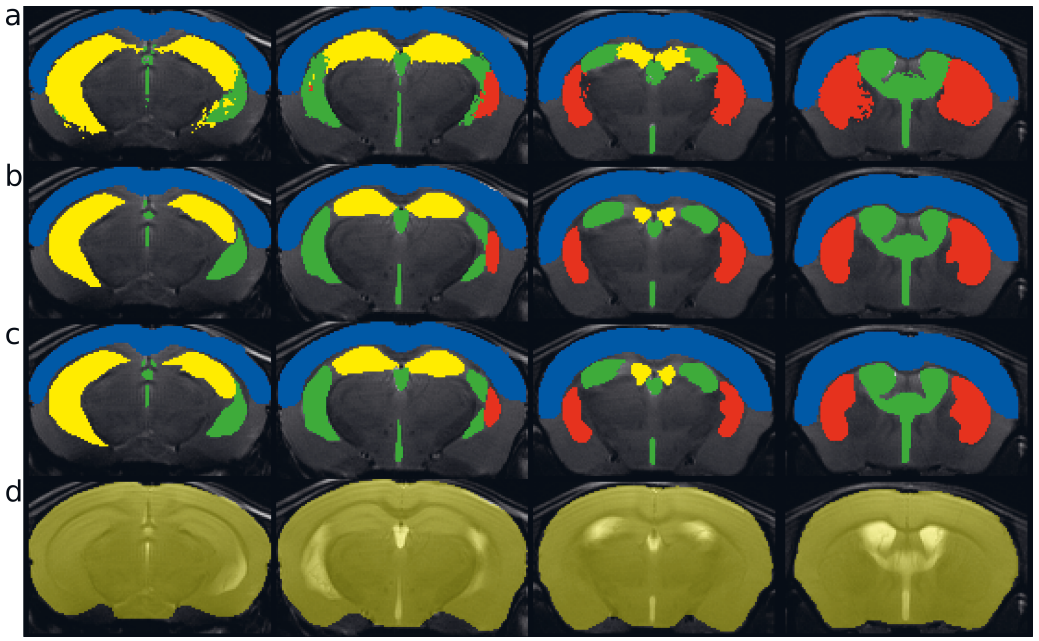

Figure 20. Segmentation comparison in four slices from a single animal: (a) STEPS, (b) MU-Net, and (c) manual annotation. In (a)-(c), the regions highlighted are the cortex (blue), ventricles (green), striati (red), and hippocampi (yellow). Panel (d) shows the inferred brain mask by MU-Net. Image reproduced under CC license (De Feo *et al.* 2021).

8.3 COMPARISON WITH MULTI-ATLAS SEGMENTATION

We compared MU-Net with multi-atlas segmentation, applying the state-of-the-art STEPS (Cardoso *et al.* 2012, 2013) label fusion method to combine the labels obtained from the registration of multiple labeled volumes. We repeated this procedure using both diffeomorphic and affine registration methods, with randomly-selected templates restricted to same-age mice. The brain mask segmentation was not evaluated as the manually drawn mask was used during the diffeomorphic registration procedure.

MU-Net achieved higher Dice coefficients than all STEPS implementations ($p < 0.00001$, Cohen's d: 4.39, see Table 8). Also, there was a marked qualitative difference between STEPS segmentation and MU-Net (Fig. 20), the latter achieving results visually indistinguishable from manual segmentation. We computed HD95 distances further confirmed this difference, with an average of $0.084 \pm 0.019\,\mathrm{mm}$ for MU-Net against $0.251 \pm 0.064\,\mathrm{mm}$ for STEPS ($p < 0.00001$). We measured a mean precision of 0.962 ± 0.008 (MU-Net) vs 0.820 ± 0.025 (STEPS) ($p < 0.00001$) and a mean recall of 0.951 ± 0.011

Table 9. Mean and standard deviation of average Dice scores evaluating the accuracy of MU-Net trained on volumes segmented via STEPS.

Training Set	Cortex	Hippocampus	Ventricles	Striatum	ROI mean
STEPS*	0.954±0.011	0.867±0.027	0.866±0.035	0.898±0.017	0.896±0.043
STEPS	0.953±0.009	0.872±0.022	0.849±0.041	0.885±0.016	0.890±0.046

(MU-Net) vs 0.952 ± 0.013 (STEPS) ($p = 0.65$).

MU-Net had an inference time of about $0.35\,\mathrm{s}$ and a training time of 12 hours. STEPS segmentation procedure required total inference time of 117 minutes for each labeled volume (on average $440\,\mathrm{s}$ for each pairwise diffeomorphic registration and $7.85\,\mathrm{s}$ for label fusion). Implementing STEPS segmentation using only templates of the same age led to a small but significant improvement in Dice coefficients over randomly choosing templates of any age ($p < 0.0007$, Cohen's d: 0.296). The employment of diffeomorphic registration was the most important factor affecting the performance of STEPS, as displayed in Table 8. A simple majority voting strategy led to significantly lower performance in all ROIs compared to all other label fusion strategies ($p < 0.003$).

Furthermore, we trained MU-Net on the outputs of the implemented STEPS procedures featuring diffeomorphic registration, and measured the Dice scores of each network's output with the ground truth (Table 9). As evidenced in Tables 8 and 9, and Figure 21, MU-Net trained on STEPS segmentations achieved higher Dice score with the ground truth than the same STEPS segmentations constituting the training sets of MU-Net ($p < 0.00001$). With the exception of the network trained on 5 weeks old mice, these hybrid networks were still under-performing compared to training on manually segmented data ($p < 0.00001$).

8.4 EVALUATION ON A LARGE NUMBER OF ROIS

We trained and evaluated MU-Net on the MRM NeAt datasets that includes atlases of 10 individual T_2*-weighted *in vivo* brain MR images of 12-14 weeks old C57BL/6J mice; each with 37 manually labeled anatomical structures (Ma *et al.* 2008). This same database was selected by Ma *et al.* (2014) to evaluate the STEPS multi-atlas segmentation algorithm on mouse brain MRI. To compare MU-Net with STEPS, we followed the STEPS implementation by Ma *et al.* (2014) as released by the authors.

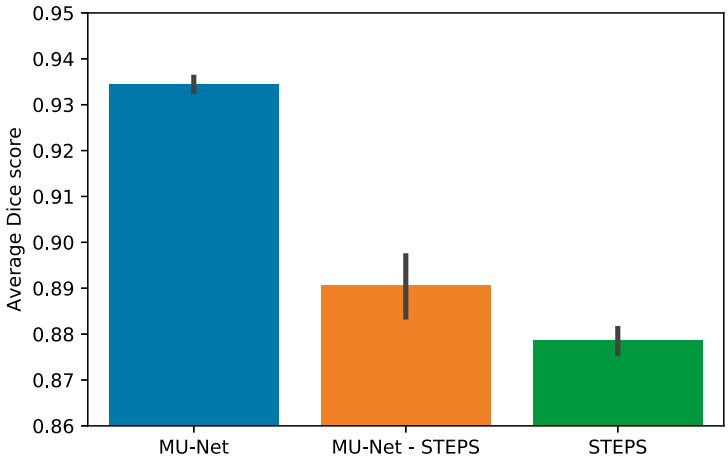

Figure 21. Average Dice score comparison between different segmentation methods, across all ROIs. MU-Net: MU-Net trained on the manually segmented data; MU-Net - STEPS: MU-Net trained on volumes segmented employing same-age diffeomorphic STEPS; STEPS: same-age diffeomorphic STEPS segmentation. The error bar represents standard deviation. Image reproduced under CC license (De Feo *et al.* 2021).

We used a 5-fold cross validation scheme for evaluation (8 templates for training and 2 templates for testing in each fold). The only adaptation required to train MU-Net on MRM NeAT dataset was to expand the number of output channels to 37 (plus one for the brain mask) to equal that of the number of ROIs. As displayed in Fig. 22, Dice coefficient of MU-Net was greater or comparable to STEPS: while in a majority of regions MU-Net's accuracy was higher than the accuracy of STEPS, this was statistically significant only for the brain mask, external capsule, hypothalamus and brain stem. In the left inferior colliculi, STEPS achieved significantly higher Dice coefficient than MU-Net. Averaging the Dice coefficients across all ROIs, we measured an average Dice score of 0.820 ± 0.031 for MU-Net and 0.814 ± 0.023 for STEPS. While this average Dice coefficient for MU-Net was higher, the difference was not statistically significant ($p = 0.170$, Cohen's d: 0.134). Similarly, we measured an higher (but not statistically significant, $p = 0.07$) average HD95 distance for MU-Net ($0.360 \pm 0.252\,\mathrm{mm}$ vs $0.240 \pm 0.038\,\mathrm{mm}$). In contrast, we measured a significantly higher average precision with MU-Net (0.823 $\pm$ 0.033 vs 0.786 $\pm$ 0.024, $p = 0.0009$) and a significantly lower recall (0.815 $\pm$ 0.032 vs 0.853 $\pm$ 0.023, $p = 0.001$). The computation time required by STEPS to segment a

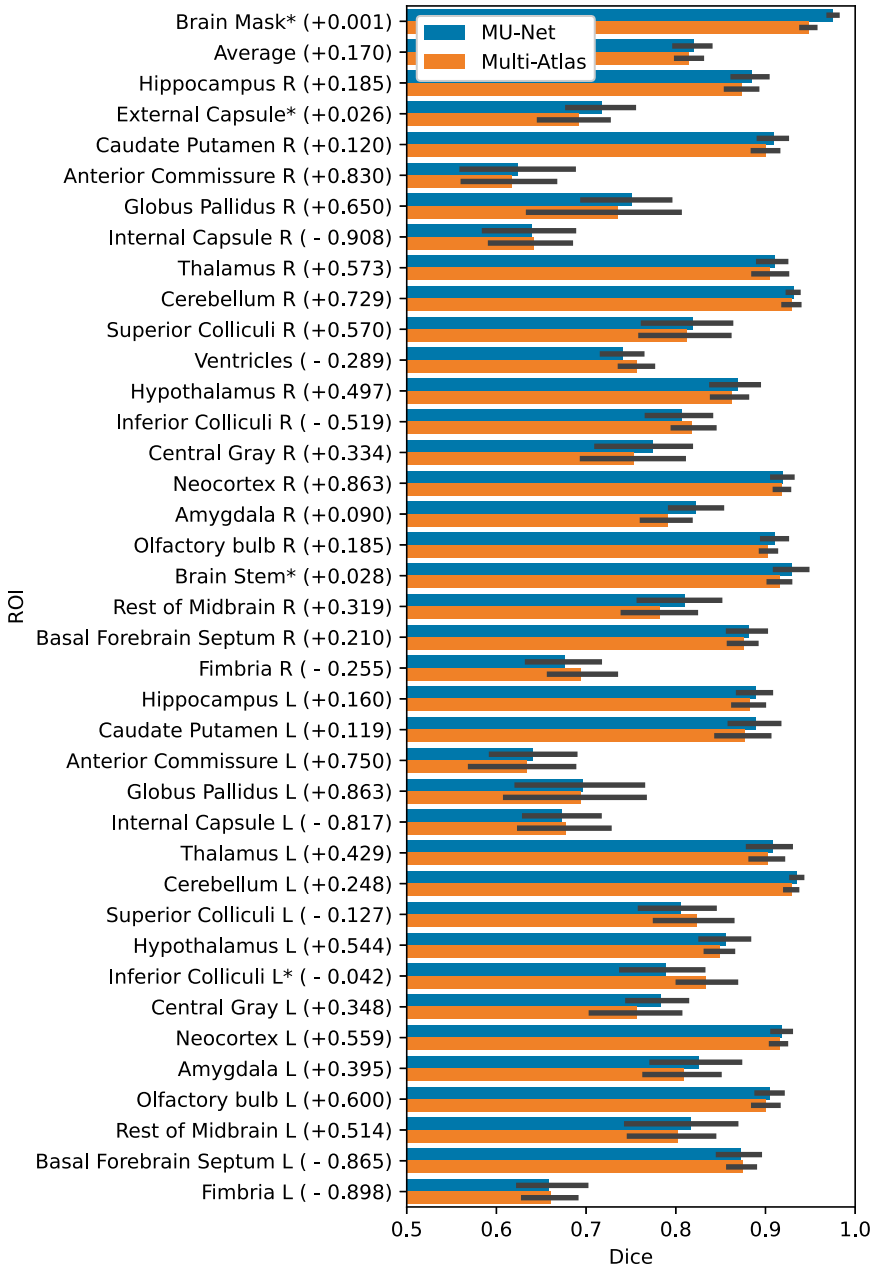

Figure 22. Comparison between the average Dice coefficients of MU-Net and STEPS multi-atlas algorithm by Ma et al. Error bars correspond to standard deviation for the average accuracy. Permutation-test based p-values for each comparison are provided in parentheses after the ROI name, + indicates that the average Dice coefficient for MU-Net was higher and − indicates that the average Dice coefficient for STEPS was higher, * indicates a statistically significant difference. Image reproduced under CC license (De Feo *et al.* 2021).

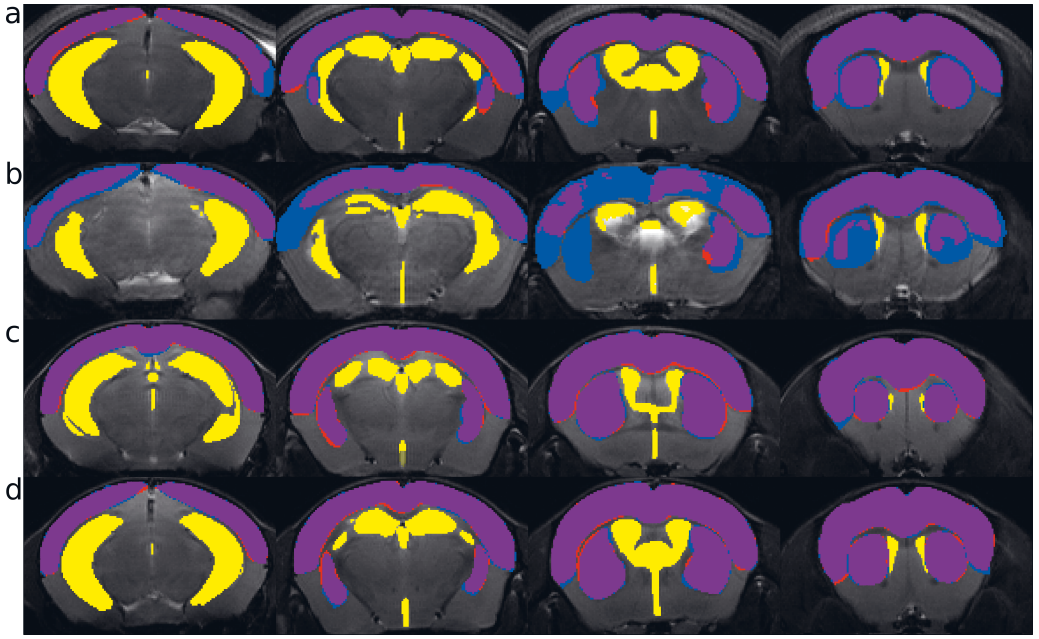

Figure 23. MU-Net segmentation compared to the manual segmentation in four slices of four volumes of the test set. Blue and red indicate, respectively, ground truth and inferred segmentation, purple their overlap (striati and cortex); yellow ROIs (ventricles and hippocampi) are inferred ROIs for which manual annotations were not available. Rows indicate (a) the highest performing volume (mean Dice 0.964, 8 weeks old R6/2 mouse); (b) the lowest performing volume (mean Dice 0.685, 12 weeks old R6/2 mouse); (c) the volume displaying performance closest to the mean performance on the entire test set (Dice 0.923, 12 weeks old Q175DN mouse); (d) one randomly selected volume (Dice 0.919, 8 weeks old Q175DN mouse). Image reproduced under CC license (De Feo *et al.* 2021).

single volume was of approximately 20 minutes while MU-Net required less than one second per volume.

8.5 EVALUATION WITH A LARGE TEST DATASET

We optimized the MU-Net model on the train and validation dataset and tested on a large test set of 1,782 MRI volumes, acquired from 817 mice with ages ranging from 4 to 60 weeks, and including both WT and HT mice. As the 5-fold cross-validation experiment produced five different MU-Net models, the segmentation maps for the test set were obtained by averaging the five prediction maps produced by the five models. To outline the brain mask, we

Table 10. Average test set metrics.

Metric	Brain Mask	Cortex	Striati
Dice	0.978 ± 0.012	0.937 ± 0.035	0.906 ± 0.041
HD95 (mm)	0.345 ± 0.303	0.223 ± 0.231	0.180 ± 0.167
Precision	0.989 ± 0.006	0.939 ± 0.050	0.929 ± 0.045
Recall	0.969 ± 0.022	0.939 ± 0.054	0.888 ± 0.062

averaged sigmoid-activated predictions from five networks and thresholded them at 0.5. For region segmentation, we averaged the softmax-activated output maps, and for each voxel, we selected the class yielding the maximal averaged value as our predicted label.

The validation performance of the individual networks provides a reasonable baseline evaluation for the ensemble, which is expected to perform equally or better than the individual networks, also taking into account the limited standard deviation of the validation Dice scores. Furthermore, ensembling compensates for the uniqueness of each training replica and possible overfitting. Because of these reasons, ensembling was preferred to simply training one final neural network on the entire dataset.

Out of the entire test set, segmentation failed completely on two volumes, where no brain mask was detected. The remaining 1780 volumes were successfully segmented with an average Dice score of 0.978 ± 0.012 for the brain mask, 0.906 ± 0.041 for the striati, and 0.937 ± 0.035 for the cortex, distributed as illustrated in Fig. 24. There was no significant difference between the segmentation accuracy of male and female animals ($p > 0.1$, unpaired). However, there was a significant difference in accuracy between HT and WT mice ($p < 0.00001$, unpaired) for all ROIs. Dice scores of WT animals were 0.4% higher for the brain mask, 1.7% higher for the cortex, and 1.9% higher for the striati. Applying N3 bias correction on all volumes before segmentation did not result in a significant Dice score difference.

A visual inspection of the segmentation maps (Fig. 23) revealed that ROIs were qualitatively similar to those obtained on the validation set and displayed in Fig. 20. We observed, however, a visible decrease in performance in the presence of strong ringing artifacts (Fig. 23.b) This is further reflected in the higher average HD95 distances in the test dataset than in the validation dataset (Table 10).

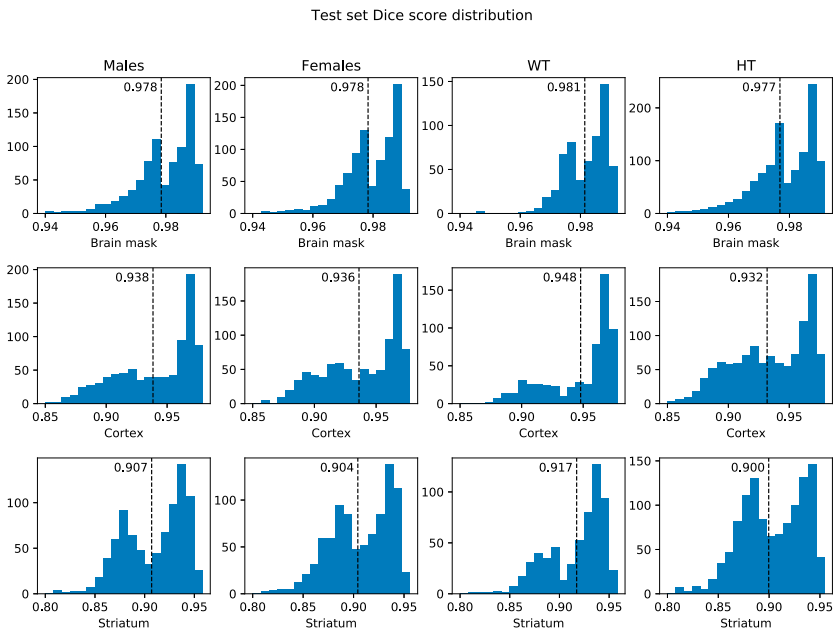

Figure 24. Test set Dice score distribution for the brain mask, cortex and striati ROIs. Males and Females include all mice of each gender, both WT and TG. Likewise, WT and TG include both males and females. Image reproduced under CC license (De Feo *et al.* 2021).

9 SEGMENTATION IN THE PRESENCE OF LESIONS

We automatically annotated every manually-labeled image in the EpiBioS4Rx and EPITARGET datasets using multi-atlas segmentation (STEPS and majority voting), single atlas segmentation, and MU-Net-R. On a qualitative level, both multi-atlas methods and MU-Net-R showed visually convincing segmentation maps, while single-atlas segmentation resulted in the lowest-quality results (Figures 25 and 26). Where the hippocampus was markedly displaced by the injury, we noticed that registration-based methods could mislabel the lesioned area as hippocampus, as displayed in Figure 26.

9.1 EPIBIOS4RX

MU-Net-R obtained excellent segmentation evaluation scores in both hemispheres as illustrated in Fig. 27. For the ipsilateral and contralateral hippocampus MU-Net achieved, respectively, average Dice scores of 0.921 and 0.928, HD95 distances of $0.30\,\mathrm{mm}$ and $0.26\,\mathrm{mm}$, precision of 0.935 and 0.936, recall of 0.909 and 0.921, VS of 0.968 and 0.971, and CS of 0.974 and 0.979.

Quantitatively, we observed a marked difference in performance between single-atlas segmentation and all other methods in terms of Dice score, Precision, HD95, and CS ($p < 0.001$ for all tests), with single-atlas segmentation obtaining the worst scores. The performance measures of all other methods were excellent. In terms of HD95 distance, the best performing method was MU-Net on the ipsilateral hemisphere, and majority voting in the contralateral one ($p < 0.05$ for both tests). The same pattern held for the Dice score. MU-Net achieved the highest precision in the ipsilateral hippocampus ($p < 0.05$). We found no significant difference between the precision of MU-Net and that of majority voting in the contralateral one ($p > 0.7$). The precision of both methods was markedly higher than STEPS and single atlas ($p < 0.05$). As an exception to the general trend, single-atlas segmentation showed the highest value of the recall metric in the contralateral hippocampus ($p < 0.001$), while STEPS outperformed it in the ipsilateral one ($p < 0.02$). We found again no significant difference ($p > 0.8$) in the VS scores for majority voting and MU-Net-R in the ipsilateral hippocampus, with these two methods achieving the highest VS scores ($p < 0.05$). In contrast, majority voting achieved higher VS in the contralateral hemisphere ($p < 0.0005$). Majority voting also better preserved the compactness properties of the hippocampal shape, achieving the highest CS among all the methods ($p < 0.0005$).

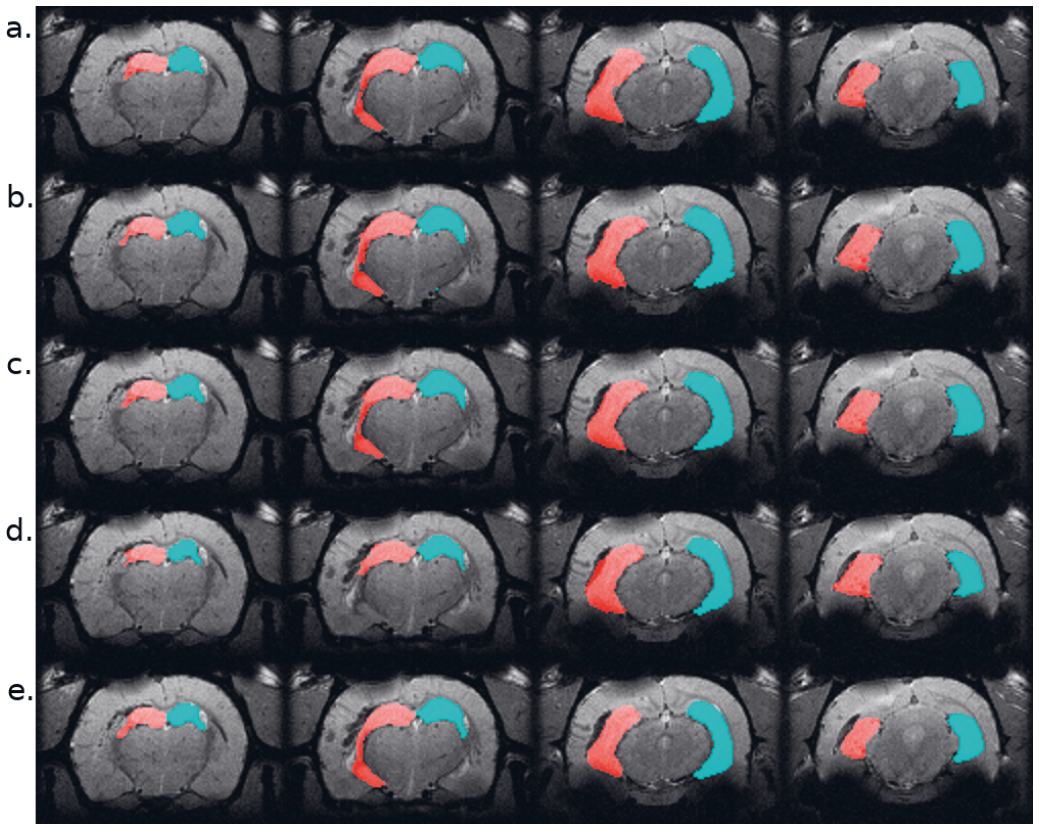

Figure 25. Segmentation maps in 4 representative slices from a randomly selected animal from the EpiBioS4Rx dataset, at the 2-days timepoint. Segmentation maps were obtained with: a. MU-Net-R; b. STEPS, c. Majority voting, d. Single-atlas segmentation. e. Displays the ground-truth segmentation. From left to right, slices are located at approximately -2.2, -3.3, -5.0, $-6.2\,\mathrm{mm}$ from bregma. Red: hippocampus ipsilateral to the lesion; blue: hippocampus contralateral to the lesion. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

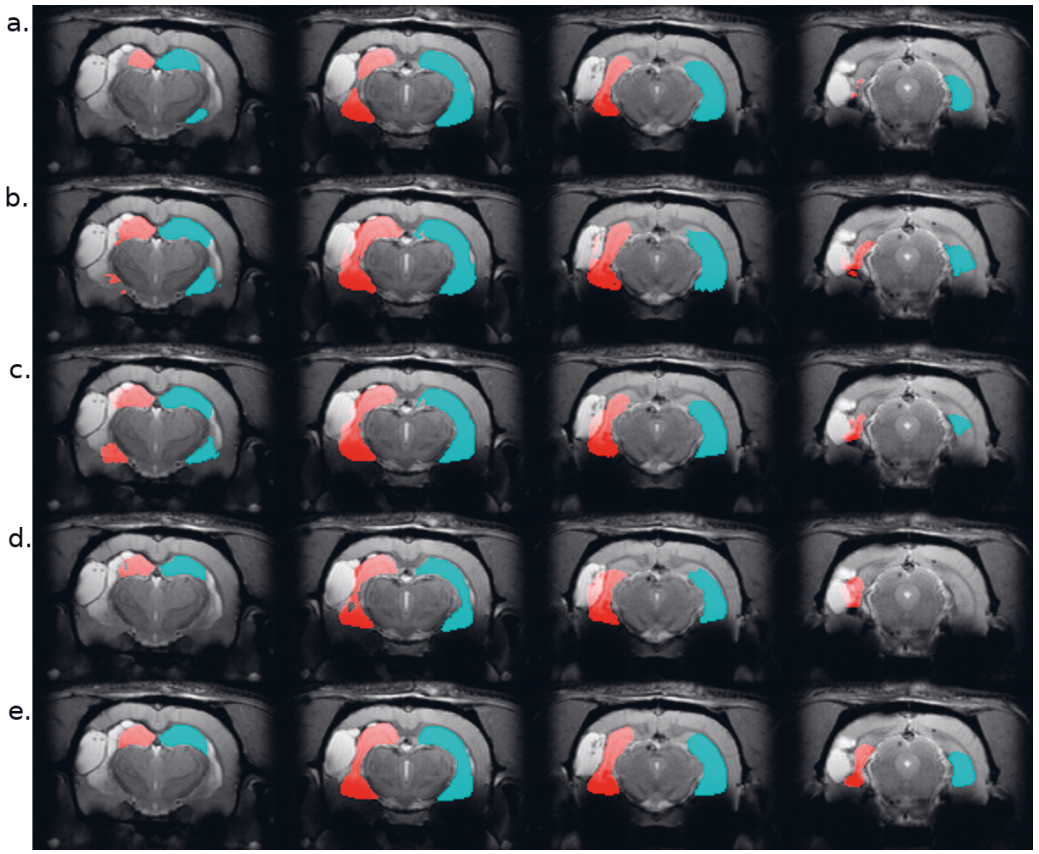

Figure 26. Segmentation maps in 4 representative slices from a randomly selected animal from the EPITARGET dataset, at the 21-days timepoint. Maps are obtained with: a. MU-Net; b. STEPS, c. Majority voting. d. Single-atlas segmentation. e. Displays the ground-truth segmentation. From left to right, slices are located at approximately -3.5, -4.5, -5.5, $-6.5\,\mathrm{mm}$ from bregma. Red: hippocampus ipsilateral to the lesion; blue: hippocampus contralateral to the lesion. Note how in this case all methods except MU-Net-R mislabel a portion of the lesion as hippocampus. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

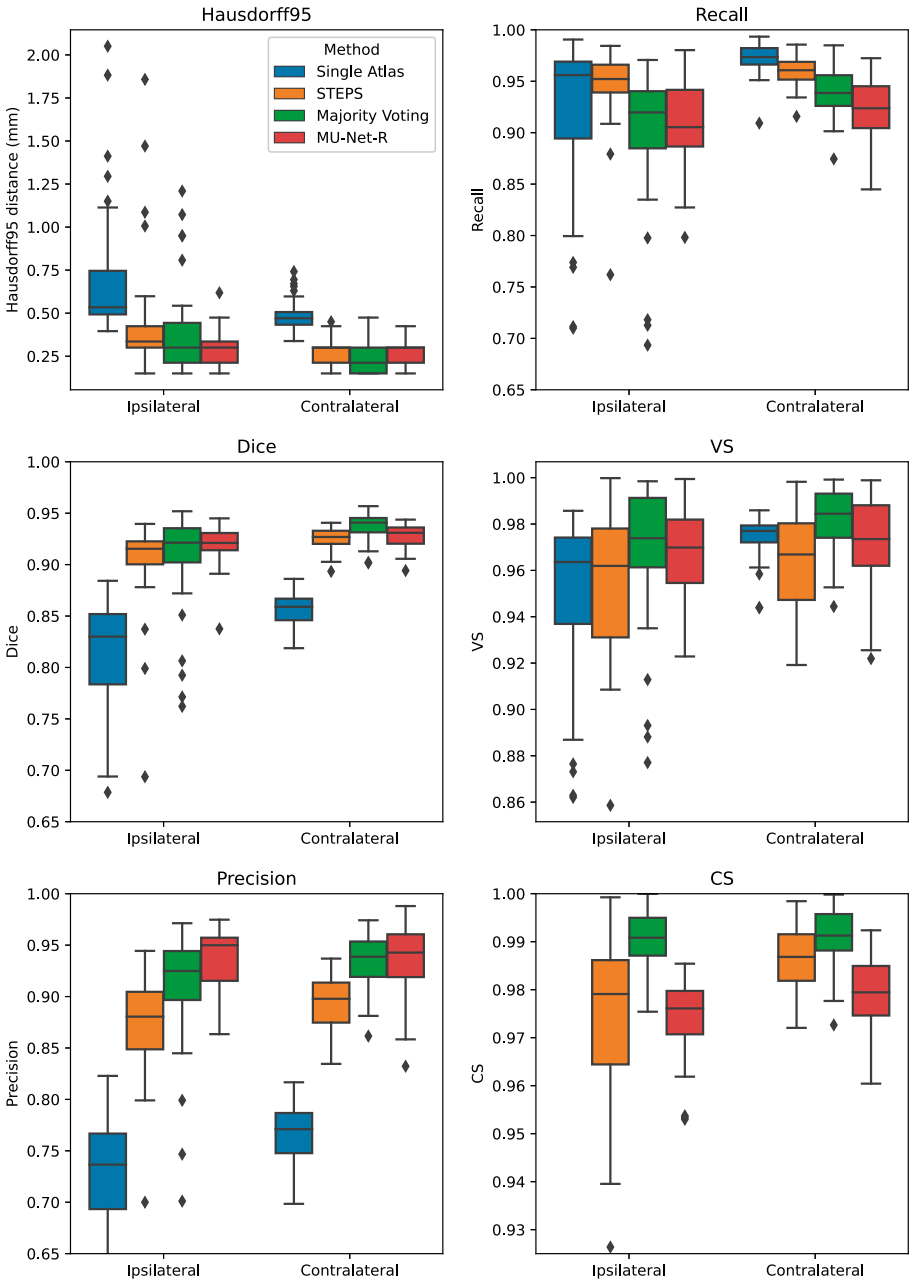

Figure 27. Box plots of all measured quality metrics for the contralateral and ipsilateral hippocampus in the EpiBioS4Rx dataset. Single atlas CS measures (not displayed) average at 0.60 for both hippocampi with a standard deviation of 0.01. Contrary to single-atlas segmentation, MU-Net-R and multi-atlas methods achieve human-level performance. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

9.2 EPITARGET

For the EPITARGET dataset, we observed similar pattern of the segmentation evaluation metrics to the ones of EpiBioS4Rx, bilaterally recording good performance metrics for MU-Net-R (see Fig. 28). We measured, respectively, for the ipsilateral and contralateral hippocampus, Dice scores of 0.836 and 0.838, HD95 distances of $0.46\,\mathrm{mm}$ and $0.43\,\mathrm{mm}$, precision of 0.897 and 0.881, recall of 0.787 and 0.804, VS of 0.928 and 0.946, and CS of 0.992 and 0.991.

In terms of HD95 distance, on the contralateral hemisphere, we found a significant difference between the under-performing single-atlas method and all other methods ($p < 0.0005$), all the other methods performing similarly according to HD95 metric. For the ipsilateral hemisphere, we additionally observed a small advantage for MU-Net-R over STEPS ($p < 0.05$). Similarly, there was no significant Dice score difference in the ipsilateral hippocampus between MU-Net-R and multi-atlas methods. Interestingly, for the contralateral hippocampus, both STEPS and majority voting achieved higher Dice scores than in the lesioned hemisphere ($p < 0.001$), obtaining also a higher Dice score than MU-Net-R ($p < 0.03$). In contrast, MU-Net-R produced similar Dice scores in the two hemispheres ($p > 0.7$). We measured higher precision for majority voting and MU-Net-R compared to all other methods ($p < 0.001$) and no significant difference between the two in both the contralateral and ipsilateral hippocampus ($p > 0.1$). We observed higher recall bilaterally for single atlas segmentation compared to all other methods ($p < 0.01$), with the exception of STEPS on the ipsilateral hemisphere, where the difference was not significant ($p > 0.2$). No significant difference was also detected for VS between the different methods ($p > 0.1$) and for CS on the contralateral hippocampus. Conversely, MU-Net-R performed better than STEPS on the ipsilateral hippocampus ($p < 0.05$).

9.3 INTER-HEMISPHERIC DIFFERENCES

We compared the quality of the automatic segmentations, by comparing all of the evaluation metrics described in Section 2.1 between the ipsi and contralateral hemispheres, finding that all segmentation methods obtained better results on the contralateral hippocampus. However, the inter-hemispheric differences for each metric (defined as the average difference in each metric for each brain between the segmentation of the ipsilateral and the contralateral hippocampus) were the smallest for MU-Net-R (Fig. 29). MU-Net-R achieved significantly smaller inter-hemispheric differences than other methods in all metrics (maximal $p < 0.02$) with the exception of recall, where both MU-Net-R

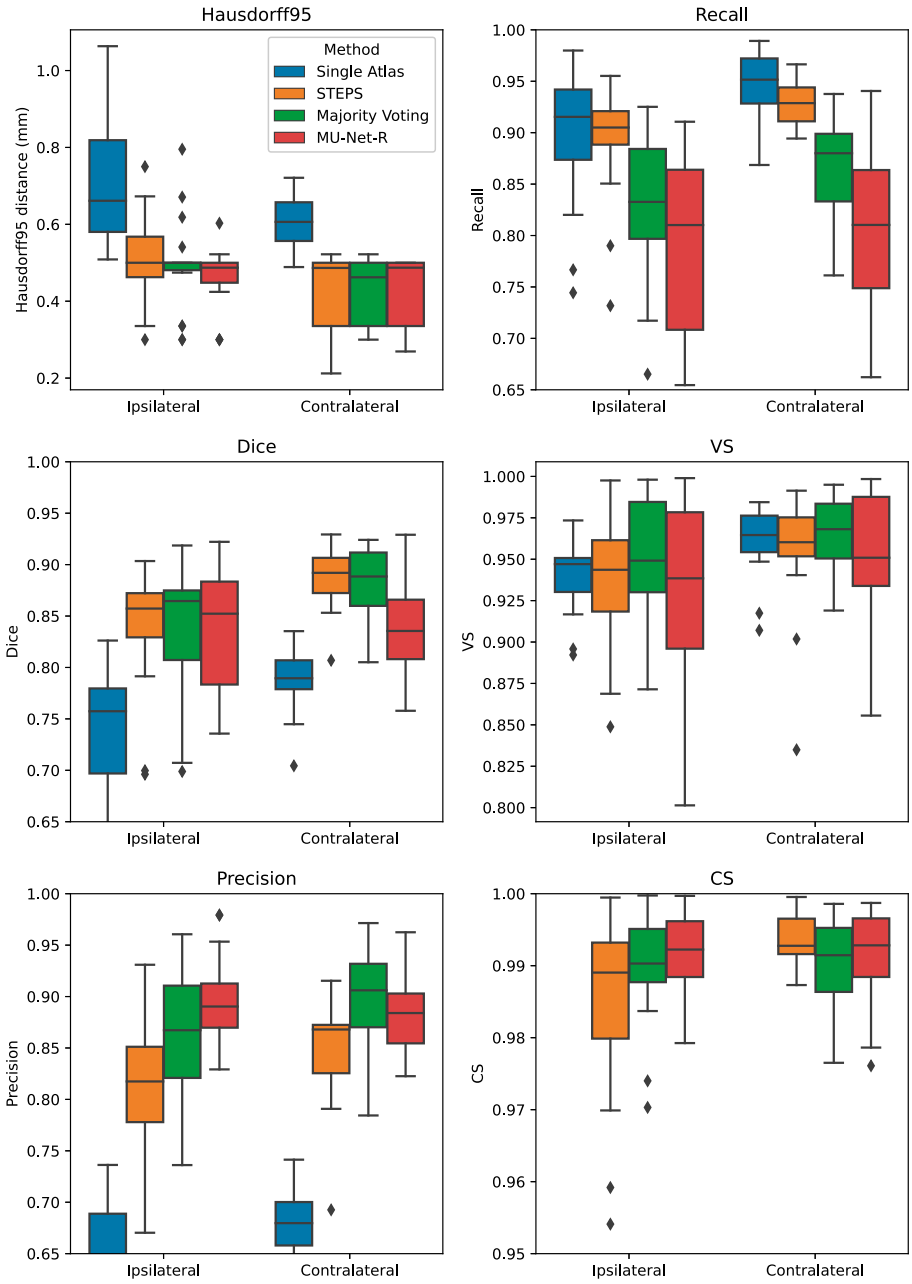

Figure 28. Box plots of all measured quality metrics for the contralateral and ipsilateral hippocampus in the EPITARGET dataset. Single atlas CS measures (not displayed) average at 0.67 for both hippocampi with a standard deviation of 0.02. While MU-Net-R and multi-atlas methods still outperform single-atlas segmentation, in this anisotropic dataset, training MU-Net-R with a smaller dataset, we register a lower performance compared to the EpiBioS4Rx results. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

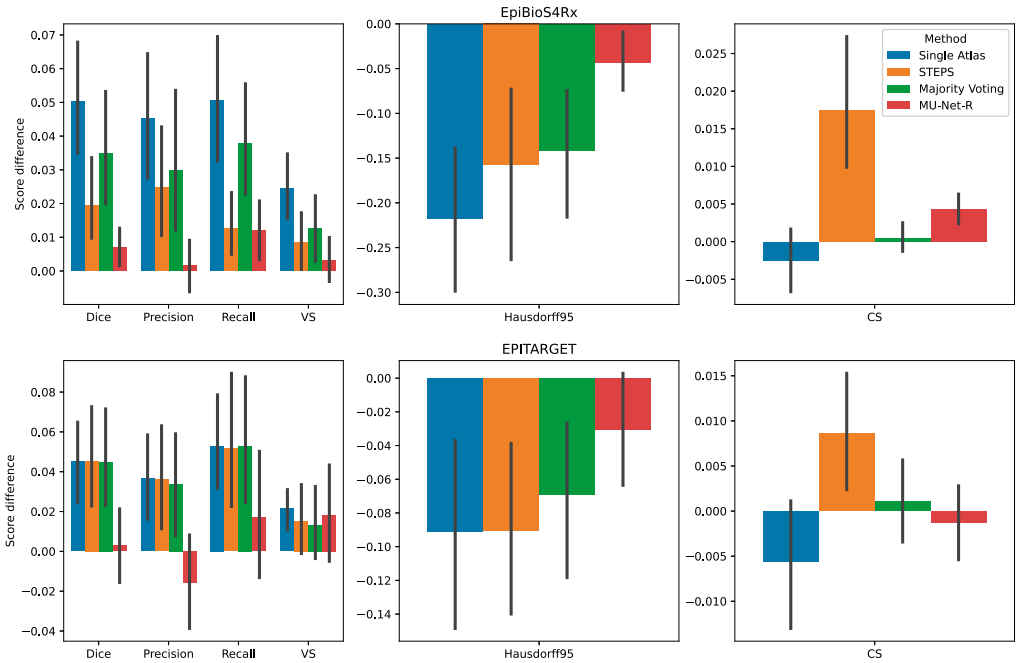

Figure 29. Mean inter-hemispheric differences for all evaluation metrics in single-atlas, STEPS, Majority Voting and MU-Net-R segmentation. Error bars correspond to 95% bootstrapped confidence intervals for the mean. MU-Net-R minimizes inter-hemispheric performance differences across all metrics and on both the EpiBioS4Rx and EPITARGET datasets, with the exception of CS on the EpiBioS4Rx dataset. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

and STEPS performed better than all other methods ($p < 0.03$), and CS, where majority voting compares favorably to both STEPS and MU-Net-R (maximal $p < 0.02$).

For the EPITARGET dataset and for all evaluation metrics, we observed a smaller average amplitude of the inter-hemispheric differences for MU-Net-R than for other segmentation methods (Fig. 29), with a single exception of CS for majority voting. However, differences were statistically significant only for the Dice score (maximal $p < 0.02$), where MU-Net-R demonstrated a stable performance between the two hemispheres. In this case, the average Dice score difference of MU-Net-R was 0.003 with a standard deviation of 0.040, while all other methods displayed an average difference of 0.045 and standard deviations of at least 0.044.

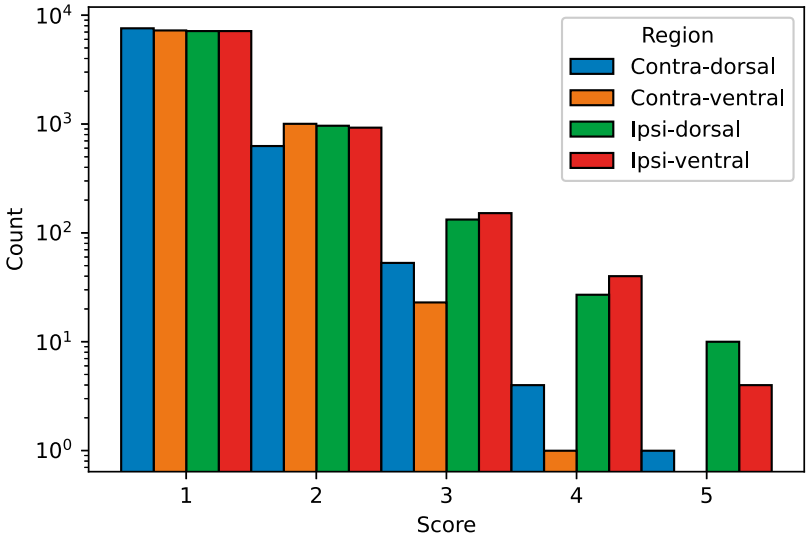

Figure 30. Qualitative score distribution for both hippocampi, divided for the dorsal and ventral aspects of the contralateral and ipsilateral hippocampus. Counts are reported on a logarithmic scale. The overwhelming majority of hippocampal regions were labeled as requiring no corrections (1), followed by regions requiring minor corrections only (2). Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

9.4 SEGMENTATION TIME

The inference time of MU-Net-R was lower than one second per volume. The training of one ensemble of MU-Net-Rs with early stopping required on average 124 minutes for EpiBioS4Rx and 64 minutes for EPITARGET. Registering a single volume pair required approximately 40 minutes for EpiBioS4Rx volumes and 6 minutes for EPITARGET volumes. After all volumes were registered, applying majority voting and STEPS label fusion required approximately 10 seconds per target volume. Thus, multi-atlas segmentation with 10 atlases required 400 minutes for EpiBioS4Rx and multi-atlas segmentation with 5 atlases required 30 minutes for EPITARGET.

9.5 VISUAL EVALUATION

We visually evaluated 33136 slices from the 220 volumes in the EpiBioS4Rx dataset. For each volume, one coronal slice at a time was selected, starting caudally and proceeding in the rostral direction. We simultaneously displayed the unlabeled MRI slice side-by-side with the same MRI slice overlaid with the ipsilateral hippocampus highlighted in red, and the contralateral one in blue. The volumes were presented to the annotator in a randomized order. For each slice, the annotator was asked to input four numbers, corresponding to an evaluation for the dorsal and ventral parts of the left and right hippocampus. The evaluation scale, inspired by Greenham *et al.* (2014), was as follows:

1. Acceptable 'as is'
2. Minor differences. Minor edits necessary. A small number of voxels, or less than 20% of the area
3. Moderate edits required, 20–50% of the area would need to be changed
4. Major edits required, $>$50% of the area would need to be manually edited
5. Gross error, no resemblance to the anatomical structure

The visual evalaution was performed by the same trained researcher who labeled the ground truth volumes, Elina Hämäläinen. In addition to every labeled slice, we evaluated two additional slices in each direction, rostrally and caudally, to allow for the detection of hippocampal regions erroneously labeled as background. As the number of misclassified voxels was small in these cases we classified errors in these slices with a score of 2, to avoid introducing a bias because of the choice of performing this evaluation on coronal slices.

In the vast majority of cases, the reported score was 1 (Acceptable 'as is'), with a small reduction in accuracy for the ipsilateral hippocampus (Fig. 30). Overall, we found that 88.00% hippocampal regions were labeled as 1, 10.64% labeled as 2, 1.09% labeled as 3, 0.22% labeled as 4, and 0.05% labeled as 5. As illustrated in Fig. 31, the accuracy of the segmentation was the lowest in the most rostral and most caudal coronal slices. Figure 32 provides examples of the segmented slices for each score.

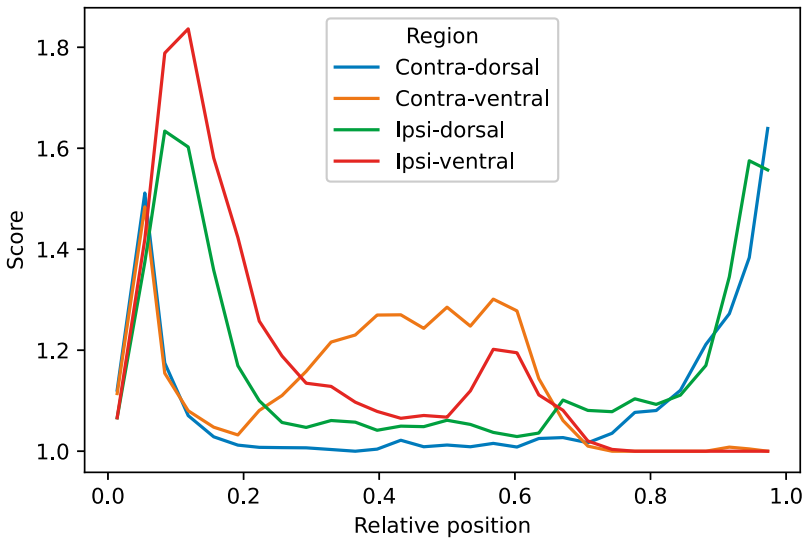

Figure 31. Average qualitative scores as a function of the relative position of the slice across the hippocampus, with 0 indicating the most caudal and 1 the most rostral coronal slice. Averages were obtained by dividing the interval in 30 bins. The vast majority of inaccuracies are located in the most rostral and caudal slices. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

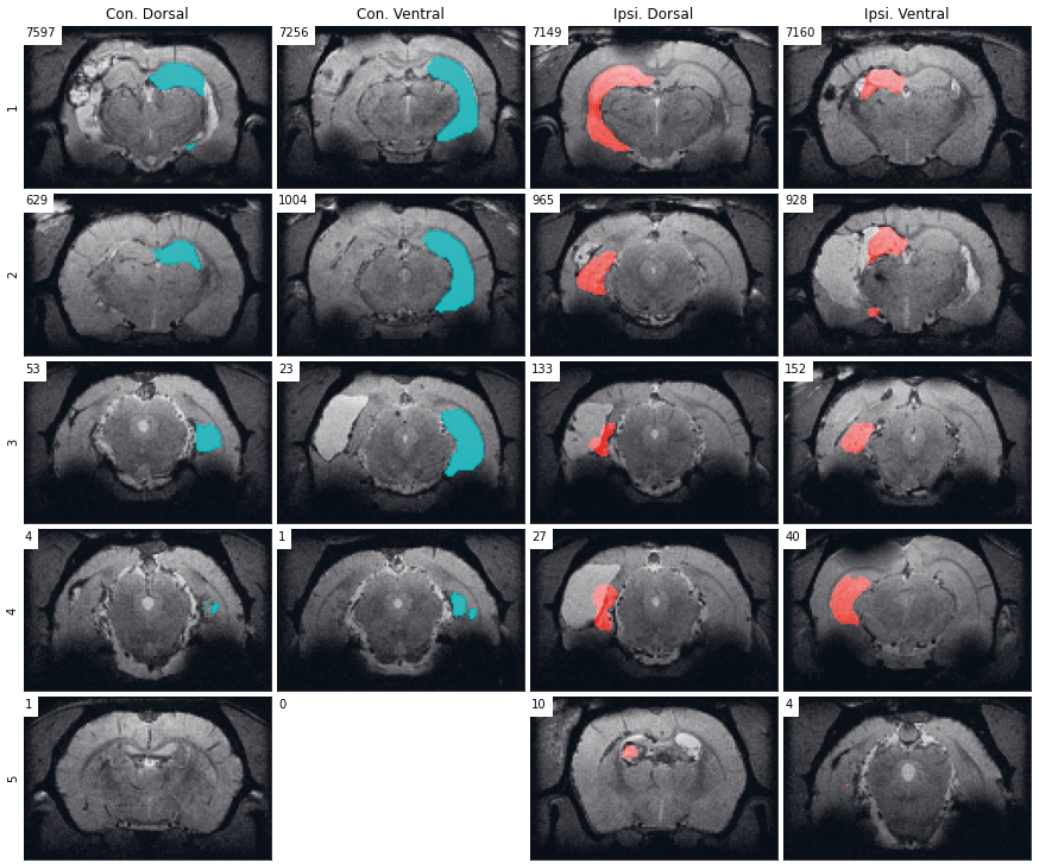

Figure 32. Examples of MU-Net-R hippocampus segmentations and their visual evaluation scores, for the dorsal and ventral regions of the ipsilateral and the contralateral hippocampus, respectively outlined in red and blue. Numbers indicate the total number of slices annotated with each score in each region. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

10 HIPPOCAMPAL GEOMETRY AS A BIOMARKER

10.1 EPILEPTOGENESIS

Globally, an estimated 2.4 million people are diagnosed with epilepsy each year (World Health Organization 2019). In 60% of those affected, epileptogenesis is initiated by structural causes such as Traumatic Brain Injury (TBI) (Scheffer *et al.* 2017). Approximately 20 hypothesis-driven intervention approaches have demonstrated some disease-modifying effect in animal models of Post-Trauamtic Epilepsy (PTE) (Dulla & Pitkänen 2021). However, no clinical treatments are available to stop or alleviate epileptogenesis in at-risk TBI patients or to alleviate the course of PTE after its diagnosis. One major reason for the stalled progression of interventions to clinical anti-epileptogenesis trials is the lack of prognostic biomarkers that could be used to stratify patient populations for antiepileptogenesis trials and reduce study costs, making sufficiently powered clinical trials affordable (Engel Jr *et al.* 2013).

Epidemiological studies have shown that PTE develops in 16-20% of patients with severe TBI (Annegers *et al.* 1998, Christensen 2012). Retrospective follow-up studies indicated that in about 80% PTE was diagnosed withing 2 years and in 60% within one year after TBI (Annegers *et al.* 1998, Haltiner *et al.* 1997, Christensen 2012). The seizure onset zone was in the temporal lobe in 57% and frontal lobe in 35% of patients (Gupta *et al.* 2014). Video-EEG monitoring follow-ups in several laboratories have also demonstrated post-traumatic epileptogenesis in rodents with TBI (Dulla & Pitkänen 2021). So far, the most often used animal model of PTE is induced with lateral Fluid Percussion Injury (FPI), triggering epileptogenesis in approximately 25% of rats with severe TBI by 6 months post-injury (Kharatishvili *et al.* 2006, Shultz *et al.* 2013, Campbell *et al.* 2014, Reid *et al.* 2016, Wang *et al.* 2016, Nissinen *et al.* 2017). Like in humans, the seizure onset zone in rats with PTE develops in the lesioned cortex (Reid *et al.* 2016, Bragin *et al.* 2016, Huttunen *et al.* 2018). In addition to neocortical damage, TBI and PTE patients as well as rodents show hippocampal histopathology (Swartz *et al.* 2006). Furthermore, 44% of patients with temporal lobe seizure onset had mesial temporal sclerosis (Gupta *et al.* 2014). In animal models, slice electrophysiology studies in lateral FPI model have demonstrated hippocampal hyperexcitability as well as histopathology comparable to hippocampal sclerosis (Santhakumar *et al.* 2000, Kharatishvili *et al.* 2007). Moreover, Kharatishvili *et al.* (2007) found that the severity of hippocampal histopathology and MRI diffusivity were associated with hyperexcitability. Hayward *et al.* (2010) reported that enhanced seizure susceptibility was associated with reduced cerebral blood flow in the ipsilateral

hippocampus at 9 months post-TBI. Shultz *et al.* (2013) indicated that changes in hippocampal surface morphometry at 6 months post-TBI differentiated rats with or without epilepsy after lateral FPI. Lastly, Pitkänen & Immonen (2014) showed that increased diffusivity in the ipsilateral hippocampus at 9 d predicted increased seizure susceptibility at 12 months post-TBI. Taken together, these data suggest that hippocampal changes are part of the epileptogenic network after TBI. However, the number of studied animals, so far, has been relatively low, too limited to predict epileptogenesis after TBI with a high accuracy.

The experiments presented in this chapter were designed to test the hypothesis that parameters indicating post-TBI structural changes in the hippocampus could present prognostic biomarkers for post-traumatic epileptogenesis. We utilized two large MRI datasets of rats with lateral FPI, EpiBioS4Rx and EPITARGET and automatically segmented all samples using MU-Net-R. Using the features extracted as described in chapter 7, we extracted 39 anatomical parameters for each animal at each timepoint, and analyzed their effectiveness in discriminating between (a) Sham-operated experimental control and TBI rats and (b) TBI rats with (TBI+) and without (TBI-) epilepsy. We evaluated each parameter separately in a mass univariate analysis, and by combining them using random forest classifiers. Our data show that both approaches were highly effective in discriminating between Sham and TBI rats. Moreover, the anatomical biomarkers described here can effectively discriminate between TBI+ and TBI- rats at 5 months after TBI. While machine learning approaches for the detection of epilepsy have found several application both in imaging and nonimaging diagnostics (Abbasi & Goldenholz 2019), our novel approach provides interpretable results outlining some of the changes in hippocampal geometry as a result of epileptogenesis.

10.2 HIPPOCAMPAL VOLUMES

Using MU-Net-R, we annotated every scan in both the EpiBioS4Rx and EPITARGET dataset and modeled hippocampus volumes as outlined in Section 7.4. As displayed in Fig. 33, when comparing sham and TBI rats in EpiBioS4Rx we found that all included factors were statistically significant in explaining the volume: lesion status (sham or TBI), timepoint, and ROI, as well as their pairwise interaction terms. With the exception of the presence of lesions ($p = 0.029$), all other p-values were smaller or equal to 0.001, for both single factors and interaction terms. The same was true for the EPITARGET dataset, where all factors were highly significant ($p < 0.002$).

Table 11. Random forest performance for each imaging time point in Sham *vs.* TBI and TBI+ *vs.* TBI- (non epileptic) classification in EpiBioS4Rx and EPITARGET cohorts. '9*' refers to the classifier we trained on the 9 d EpiBioS4Rx dataset, using the top 10 parameters in order of importance from the 7 d EPITARGET dataset. Abbreviations: Acc, accuracy; BAcc, balanced accuracy; NPV, negative predictive value; PPV, positive predictive value; Prec, precision; Sen, sensitivity; Spec, specificity; TBI, traumatic brain injury.

MRI (d)	Acc	p-val	BAcc	p-val	Sen	p-val	Spec	p-val	PPV	p-val	NPV	p-val	Prec	p-val	Recall	p-val	F1	p-val
EpiBioS4Rx cohort, Sham *vs.* TBI																		
2	0.9767	0.0001	0.9844	0.0001	0.9688	0.0008	1.0000	0.0001	1.0000	0.0001	0.9167	0.0001	1.0000	0.0001	0.9688	0.0008	0.9841	0.0001
9	0.7048	0.1655	0.6270	0.1086	0.7935	0.1722	0.4545	0.1999	0.8034	0.2726	0.4488	0.1109	0.8034	0.2726	0.7935	0.1722	0.7979	0.1917
9*	0.7667	0.0434	0.7246	0.0261	0.8129	0.0701	0.6364	0.0675	0.8627	0.0989	0.5528	0.0301	0.8627	0.0989	0.8129	0.0701	0.8366	0.0590
30	0.9395	0.0001	0.9339	0.0001	0.9406	0.0040	0.9364	0.0001	0.9776	0.0001	0.8463	0.0002	0.9776	0.0001	0.9406	0.0040	0.9586	0.0004
150	0.9023	0.0003	0.8943	0.0001	0.9219	0.0035	0.8455	0.0010	0.9458	0.0014	0.7953	0.0002	0.9458	0.0012	0.9219	0.0035	0.9333	0.0009
EPITARGET cohort, Sham *vs.* TBI																		
2	0.9563	0.0001	0.9132	0.0001	0.9777	0.0001	0.8522	0.0006	0.9700	0.0002	0.8874	0.0001	0.9700	0.0002	0.9777	0.0001	0.9738	0.0001
7	0.9403	0.0001	0.919	0.0001	0.9468	0.0001	0.9087	0.0001	0.9804	0.0001	0.7815	0.0001	0.9804	0.0001	0.9468	0.0001	0.9633	0.0001
21	0.9912	0.0001	0.9891	0.0001	0.9929	0.0001	0.9826	0.0001	0.9965	0.0001	0.9663	0.0001	0.9965	0.0001	0.9929	0.0001	0.9947	0.0001
EpiBioS4Rx cohort, TBI+ *vs.* TBI-																		
2	0.5281	0.7441	0.3961	0.7988	0.1556	0.7607	0.6739	0.6054	0.1613	0.7320	0.6698	0.7408	0.1613	0.7320	0.1556	0.7607	0.1569	0.7520
9	0.5677	0.5933	0.4704	0.5552	0.2500	0.5400	0.6783	0.5781	0.2196	0.5708	0.7206	0.5327	0.2196	0.5708	0.2500	0.5400	0.2319	0.5633
30	0.5844	0.6138	0.4804	0.5315	0.3222	0.4904	0.6870	0.6015	0.2948	0.5157	0.7202	0.5993	0.2948	0.5157	0.3222	0.4904	0.3045	0.5168
150	0.8094	0.0296	0.7998	0.0054	0.7667	0.0089	0.8261	0.1236	0.6361	0.0223	0.9004	0.0247	0.6361	0.0223	0.7667	0.0089	0.6944	0.0078
EPITARGET cohort, TBI+ *vs.* TBI-																		
2	0.5536	0.5190	0.5279	0.3262	0.4966	0.2428	0.5735	0.6637	0.2894	0.3805	0.7653	0.3922	0.2894	0.3805	0.4966	0.2428	0.3653	0.3166
7	0.5982	0.2338	0.5913	0.0839	0.5966	0.0922	0.5988	0.4639	0.3447	0.1718	0.8076	0.1645	0.3447	0.1718	0.5966	0.0922	0.4368	0.1144
21	0.5867	0.2902	0.5142	0.4122	0.3586	0.6050	0.6655	0.2093	0.2711	0.4293	0.7501	0.4475	0.2711	0.4293	0.3586	0.6050	0.3082	0.5081

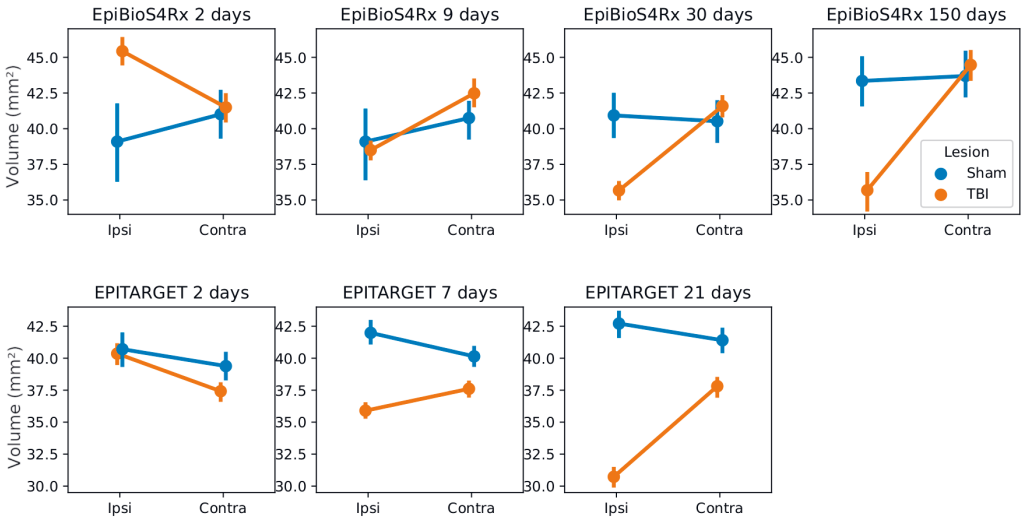

Figure 33. Comparison between the hippocampal volumes (mm^3) for each dataset and time point between sham and TBI animals. Note the different dynamics in ipsilateral hippocampal volume changes between the EpiBioS4Rx and EPITARGET animals, particularly at 2 days post-injury. The rats in the EpiBioS4Rx cohort were anesthetized with 4% isoflurane at the time of injury. The rats in the EpiTARGET cohort were anesthetized with pentobarbital-based anesthesia cocktail, which apparently reduced the acute post-impact seizure-related swelling better than isoflurane. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

10.3 TBI *VS.* SHAM CLASSIFICATION

Mass univariate analysis The repeated measure ANOVA results for each parameter displayed a widespread significant dependence on the timepoint across several parameters. In EPITARGET, a total of 14 parameters were significantly sensitive to the timepoint variable ($p < 0.05$), while in EpiBioS4Rx there were 25 parameters significantly sensitive to the timepoint variable. Fewer parameters were also significantly sensitive to TBI ($p < 0.05$). In EPITARGET , these were v_{ipsi}, v_r, the ipsilateral p^2_{pz} and $|\boldsymbol{p^1_p}|$, and the contralateral $|\boldsymbol{p^2_p}|$. In EpiBioS4Rx these were the ipsilateral p^1_{py}, p^2_{py} and $|\boldsymbol{p^1_p}|$. All of these parameters were sensitive to TBI also in combination with timepoint, as a second order effect. No parameter was significantly affected by TBI independently of the timepoint.

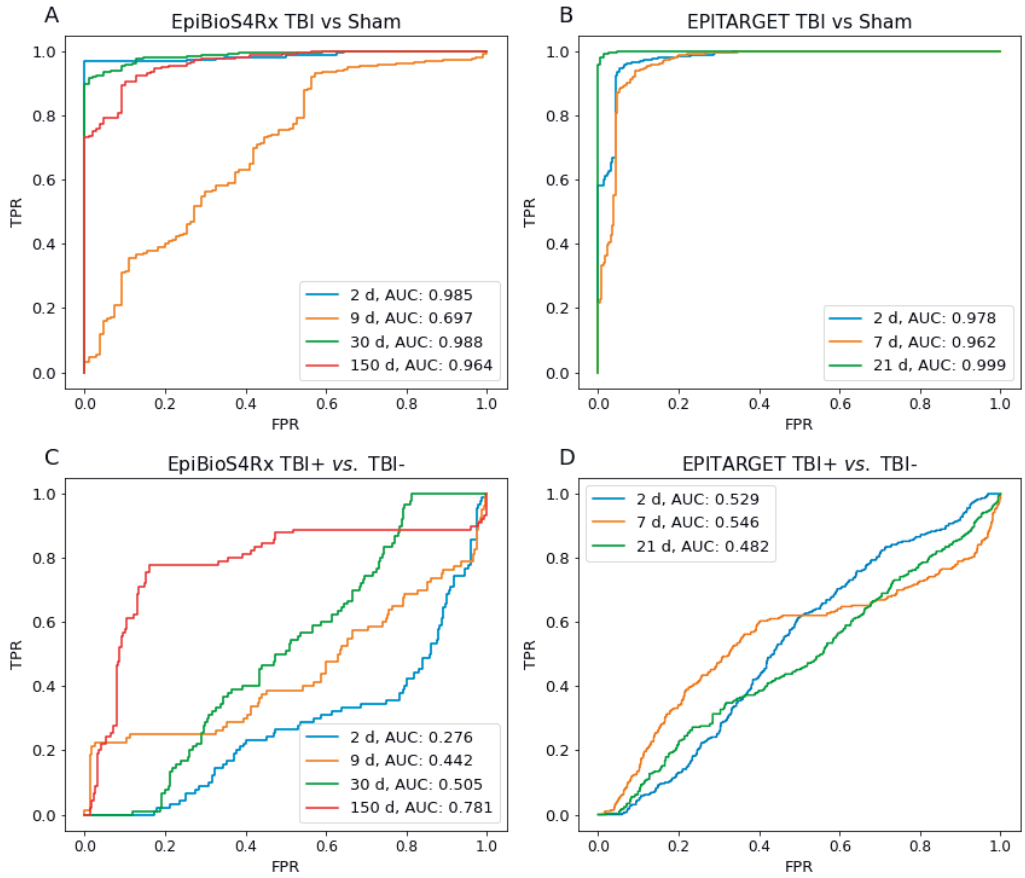

Figure 34. **(A-B)** Receiver operating curves (ROC) for all time points in TBI vs. Sham classification in the EpiBioS4Rx (left) and EPITARGET (right) animal cohorts. The random- forest-estimated probability of TBI, based on the hippocampal parameters, reached a high AUC at each time points except the 9 d EpiBioS4Rx time point. **(C-D)** ROC curves for rats with epilepsy (TBI+) *vs.* no epilepsy (TBI-) classifiers, across all time points in the EpiBioS4Rx and EPITARGET datasets. At earlier time points, the classifiers did not discriminate between the TBI+ and TBI- animals. However, in the EpiBioS4Rx cohort imaged at 150 d time point, the hippocampal parameters effectively discriminated the TBI+ and TBI- animals. Image reproduced under CC license (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

Classification We trained random forests to classify TBI vs. sham rats separately for each timepoint. Random forests of stumps were preferred after a series of preliminary experiments. More complex models such as support vector machines or random forests of larger decision trees seemed to easily overfit for the small training set.

Table 11 reports the mean and standard deviation across the folds of the stratified 10-fold CV procedure. The random forest was highly effective in discriminating between these two classes for all timepoints, achieving balanced accuracy of 90% or higher. The only exception was the 9 d timepoint in the EpiBioS4Rx dataset, with a balanced accuracy of 0.627. As a general trend, the classification quality metrics in the EPITARGET dataset were higher than in the EpiBioS4Rx dataset, with larger scores when averaged across all timepoints. This result still held true when ignoring the 9 d timepoint.For both datasets, respectively EpiBioS4Rx and EPITARGET, we observed higher average sensitivity (0.9062, 0.9724) and PPV (0.9317, 0.9823) than specificity (0.8091, 0.9145) and NPV (0.7517, 0.8784). Likewise, we measured higher average precision (0.9317, 0.9823) and lower recall (0.9062, 0.9725), with average F1 scores of 0.9185 and 0.9773. ROC curves generated for both datasets are displayed in Figure 34, with all classifiers achieving high AUC (> 0.96) with the exception of the 9 d timepoint in EpiBioS4Rx .

We report the parameter importance plot for the 2 d and the 150 d timepoints in Figures 35 and 36. The parameter importance plots for all time points in EpiBioS4Rx and EPITARGET can be found in the appendix. Interestingly, v_{ipsi} did not contribute meaningfully to the classification for the 2 d timepoint. Instead, in the EpiBioS4Rx dataset, the most important parameters were the ipsilateral $\theta_{1,R}$, p^1_{nz} (indicating the z component of $\boldsymbol{p^1_n}$) and p^1_{px} , with a minor contribution from v_r. In the EPITARGET dataset the most important parameter for the 2 d timepoint was the ipsilateral $|\boldsymbol{p^1_p}|$, followed by smaller contributions from p^1_{py}, the contralateral $|\boldsymbol{p^2_p}|$, $\theta_{1,1}$, and $\theta_{2,2}$. Volumetric parameters only had a minor contribution to classifiers at 2 d (Fig. 35).

In contrast, for the later timepoints of 7, 21, 30, and 150 days, the most important parameters were v_r followed by v_{ipsi} . With respect to 9 d time point in the EpiBioS4Rx dataset, where the classifier did not perform well, we found ipsilateral $\theta_{1,R}$, the contralateral v_{contra} and p^1_{px}, and the ipsilateral p^1_{py} to be the most relevant parameters. These were different to the important parameters in the other timepoints equal or greater than 7 days. However, by retraining this classifier only selecting the top 10 parameters by importance on the 7 d EPITARGET classifier, we obtained improved performance (balanced accuracy of 0.7246, $p = 0.0261$), indicated as 9* in Table 11.

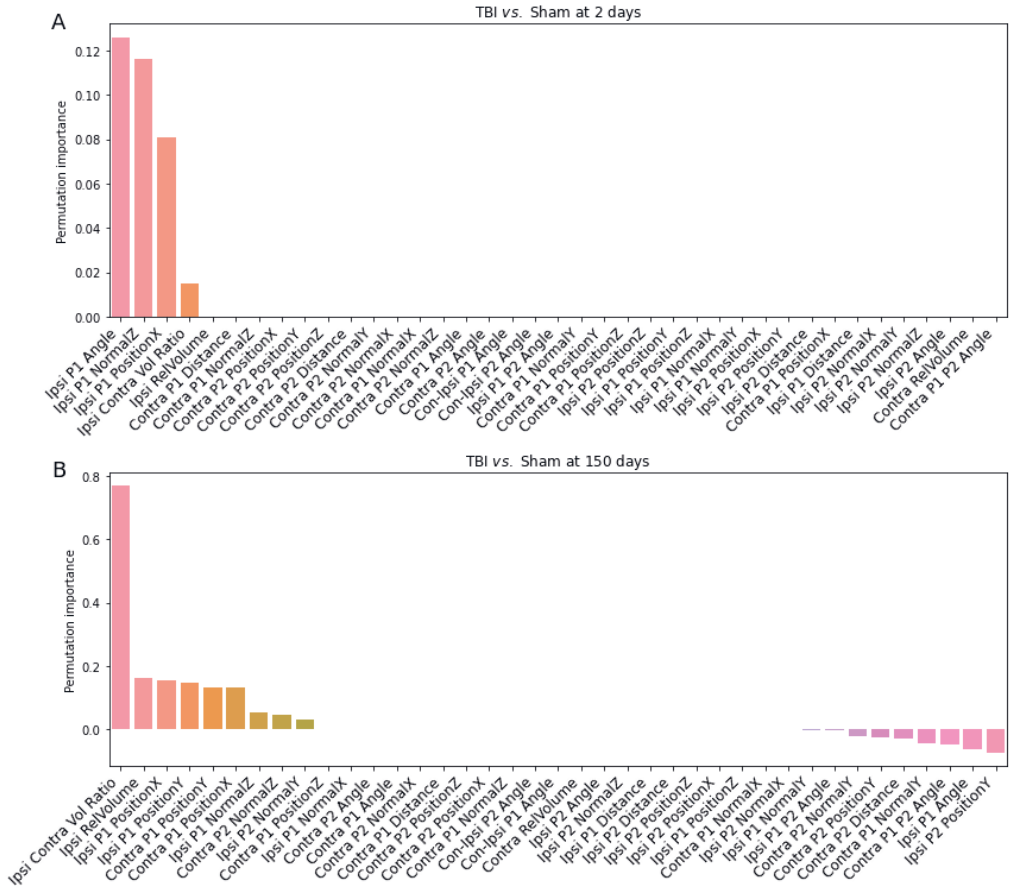

Figure 35. Parameter importance for the **(A)** 2 d and **(B)** 150 d time point classification of Sham vs. TBI animals in the EpiBioS4Rx dataset. While for the 2 d time point the most important parameters describe the orientation and positioning of the hippocampus through the plane, for later time points the hippocampal volume was the most important factor. Image reproduced under CC license (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

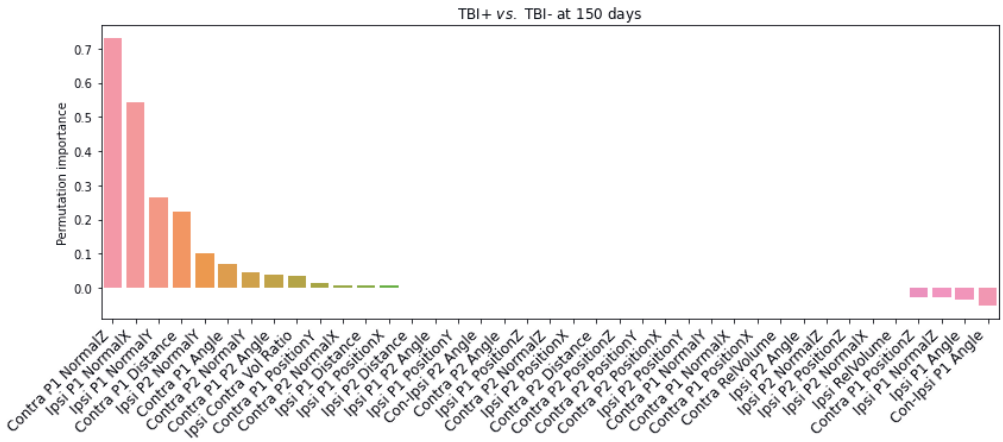

Figure 36. The 150 d time point classification of the epilepsy (TBI+) vs. no epilepsy (TBI-) animals. The most important parameters discriminating the two groups were the p1 parameters of the ipsilateral and contralateral hippocampi. Image reproduced under CC license (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)
.

10.4 TBI+ *VS.* TBI- CLASSIFICATION

Mass univariate analysis As described in the TBI vs. Sham analysis, many of the parameters were dependent on the timepoint variable, and unsurprisingly this was verified again in the TBI+ *vs.* TBI- analysis. However, no parameter displayed a significant dependence on TBI+ in the multivariate ANOVA analysis, neither when considered individually nor in combination with the timepoint, across both datasets.

Classification Next, we trained a random forest to discriminate between TBI+ and TBI- rats. The 150 d timepoint in the EpiBioS4Rx dataset revealed a high balanced accuracy of 0.7998 ($p = 0.0054$) and an F1-score of 0.6944 ($p = 0.00779$), as reported in more detail in Table 11. For all other timepoints for both datasets, we did not obtain classifiers that would have been significantly better than the chance level (50 % balanced accuracy).

The two most important parameters according to permutation importance were the contralateral p^1_{nz} and the ipsilateral p^1_{nx}. These are followed by the ipsilateral p^1_{ny}, the contralateral $|p^1_n|$, and parameters of lower importance as displayed in Figure 36. Parameter importances for all other timepoints are reported in the appendix.

11 DISCUSSION

We have presented a multi-task deep neural network, MU-Net, for the simultaneous skull-stripping and segmentation of mouse brain MRI. While Roy, Knutsen, Korotcov, Bosomtwi, Dardzinski, Butman & Pham (2018) and Hsu *et al.* (2020) proposed Convolutional Neural Networks (CNNs) for skull-stripping mouse MRI, to our knowledge this work represents the first CNN performing both region segmentation and skull-stripping. To define the structure of MU-Net we selected the best performing network among a number of architectures. This network was found to achieve better segmentation accuracy on the validation set compared to state-of-the-art multi-atlas segmentation procedures, with a markedly lower segmentation time ($0.35\,\mathrm{s}$ vs $117\,\mathrm{minutes}$). After validating this approach on a large test dataset including both Wild Type (WT) as well as Huntington (HT) mice, we adapted this approach to the task of hippocampal segmentation in rat brain MRI after TBI. We quantitatively evaluated the segmentation performance on this data and compared it with registration-based methods through a variety of metrics (Dice score, volume similarity, Hausdorff distance, compactness score, precision, and recall). Finally, we utilized these segmentation masks to build a classifier to identify rats developing epilepsy 5 months post-Traumatic Brain Injury (TBI), with a balanced accuracy of about 80%.

11.1 SEGMENTATION PERFORMANCE

11.1.1 Convolutional neural networks

We evaluated the performance of MU-Net on a large and heterogeneous test set of 1,782 mice from 10 different studies of Huntington disease, with varying ages and genetic backgrounds (WT as well as HT Q175 and R6/2 variants). In this test set, we measured average Dice scores of 0.978, 0.906 and 0.937 for the brain mask, striati and cortex, rivaling human-level performance. The performance of MU-Net-R on the hippocampi (Dice scores of 0.921 and 0.928) remained in line with the validation score of MU-Net (0.925) in the same Region of Interest (ROI), when evaluated on the EpiBioS4Rx data. We registered a lowe performance on the EPITARGET data (0.836 and 0.838), likely as a consequence of the smaller training set. Similarly, we trained MU-Net for the segmentation of high resolution mouse MRIs of the MRM Neat atlas into 37 ROIs measuring an average Dice score of 0.820. The lower Dice overlap can be understood as a consequence of a smaller training set, and a larger number of ROIs, increasing the difficilty of the task. Furthermore, as mentioned in

chapter 2, the comparison of Dice scores across different ROIs is of questionable significance.

To put the Dice scores we have reported in context, Dice scores between two human experts have ranged from 0.80 to 0.90, depending on ROI, for mouse brain MRI segmentation (Ali *et al.* 2005). For different segmentation tasks in brain MRI in general, including human data, inter- and intra-rater Dice score have ranged between 0.75 to 0.96 (Ali *et al.* 2005, Yushkevich *et al.* 2006, Entis *et al.* 2012). The Dice scores of MU-Net exceeded the above mentioned scores between two human experts, suggesting human-level segmentation performance. In addition, the Dice score of MU-Net for skull-stripping was higher than Dice score from the skull-stripping CNN implemented by Roy, Knutsen, Korotcov, Bosomtwi, Dardzinski, Butman & Pham (2018) (0.949). Obviously, comparing previously reported Dice scores to our segmentation accuracy measures must be done with care as these vary across different studies, segmentation tasks, and datasets, and the confounding factors include image resolution, presence of artifacts and noise, rater expertise, and the choice of ROIs.

The bimodal nature of the segmentation was not explained by the distinction between wild type and Huntington animals, sex, or the specific animal model. The distinction by each of the studies that contributed to this dataset also failed to explain this observation (Fig. 41). It's possible that the observed bimodality could be owed to the presence of different human annotators generating the ground truth data.

Overall, it appears that the employment of deep neural networks for the segmentation of animal MRI is a promising strategy for the reduction of both rater bias and segmentation time.

11.1.2 Registration-based segmentation

We observed that the segmentation accuracy of atlas-based methods can vary markedly, based on the specific use case depending on the number of manually drawn ROIs, voxel-size, and image quality. The best performance was achieved using advanced registration-based methods (Ma *et al.* 2014) on the high resolution data (Ma *et al.* 2008) with a densely labeled atlas of 37 ROIs, and a lower performance using a majority voting rule on a sparsely outlined atlas with a low resolution along the fronto-caudal direction. The least satisfactory results were displayed by single-atlas segmentation, with a marked decrease in segmentation quality in terms of precision, Dice score, and HD95 distance compared to all other methods. Conversely, STEPS and majority voting both displayed similarly excellent performance on the EpiBioS4Rx and EPITARGET

datasets, with Dice coefficients compatible with human inter-rater agreement (Ali *et al.* 2005, Yushkevich *et al.* 2006, Entis *et al.* 2012).

It would appear then that while STEPS and majority voting can achieve comparable performances, STEPS would be more robust, while the performance of majority voting may be affected by multiple factors. Two of these may be the different MRI setup, featuring a different coil and generating visibly different images, or the usage of different tools to perform diffeomorphic registration (FSL FNIRT (Andersson *et al.* 2007) for the CR data, and ANTs (Avants, Tustison, Song, Cook, Klein & Gee 2011) on the UEF data).

The robustness of CNNs is further evidenced by the lack of any meaningful difference when operating on volume that have or have not undergone bias correction. As the bias field simply results in a smooth gradient applied over the entire volume, it's effect on the performance of CNNs was negligible, whereas it's generally required in registration-based segmentation.

11.1.3 Comparing registration and CNNs

The advantages of CNNs with respect to atlas-based region segmentation (De Feo & Giove 2019, Bai *et al.* 2012, Ma *et al.* 2014) are clear. First, compared to atlas-based segmentation MU-Net is much faster and produces accurate results without pre-processing. While the inference time for our CNNs remains lower than one second per scan, registration required 40 minutes for each volume pair in EpiBioS4Rx . The long registration times are primarily owed to diffeomorphic registration, whereas the preliminary similarity registration and the following label fusion steps are generally faster. Eliminating the diffeomorphic registration step however would result in a marked decrease in segmentation quality. CNNs allow us to combine speed and accuracy, at the cost of a lengthy training procedure. By using early stopping in MU-Net-R the training time was brought down to be comparable to the total time required for registration-based segmentation, while generating models that can later be employed to quickly label a much larger volume of data.

Second, we found MU-Net and MU-Net-R to be equally or significantly more accurate than the state-of-the-art STEPS multi-atlas segmentation (Ma *et al.* 2014) and majority voting segmentation in all experiments. In particular, MU-Net was significantly more accurate on anisotropic, relatively quick to acquire MR images favored in pre-clinical drug and biomarker discovery applications. Furthermore, MU-Net and MU-Net-R performed better than or equally well compared to STEPS on isotropic, high-resolution MR images with relatively long acquisition times, favored in basic research.

Finally, while the overall performance of STEPS, majority voting, and MU-Net-R was of similar quality on the EpiBioS4Rx and EPITARGET datasets,

MU-Net-R was the only method to perform equally well on the hippocampus ipsilateral and contralateral to the lesion in TBI rats. MU-Net-R segmentation resulted in a marked reduction in the differences between the two hemispheres in each metric quantifying the agreement between the segmentation mask and the respective ground truth. Offering a lower performance on the hippocampus ipsilateral to the lesion, registration-based segmentation appears to be more likely to introduce random or systematic errors in the shape and positioning of the hippocampus, supporting the choice of using MU-Net-R for the segmentation of the hippocampus when looking for geometric biomarkers of epilepsy.

A small bias was still measured, both quantitatively and qualitatively. In our qualitative evaluation of MU-Net-R segmentations, we detected a larger number of slices evaluated with scores of 3 (Moderate edits required, 20–50% of the area would need to be changed) or higher for the ipsilateral hippocampus. This difference was likely owed to the small size of the training set; the presence of lesions manifesting in different shapes and sizes implies a larger degree of variability, requiring more training data to capture this variability. Taking this small difference into account, the vast majority of slices required no correction regardless of ROI, with only 2% of slices requiring moderate or larger edits.

11.2 TRAINING AND ARCHITECTURE

In literature both 3D and 2D implementations of CNNs are available for different segmentation tasks (Roy, Conjeti, Navab & Wachinger 2018, Milletari *et al.* 2016, Çiçek *et al.* 2016), and other architectural variants have been proposed: Roy, Conjeti, Navab & Wachinger (2018) added dense connections (Huang *et al.* 2017) in the convolution blocks of U-Net while keeping the number of output channels constant; Han & Ye (2018) proposed two variants based on signal processing arguments for the reduction of artifacts in a sparse image reconstruction task. We, however, found that a more complex model did not improve and in fact lowered the accuracy of our results, perhaps given the simplicity of the task. Thus, in agreement with Isensee *et al.* (2018), we found that a 2D approach was preferable to 3D approach in the presence of anisotropic voxels. We also found the Dice loss to be sufficient to effectively train our model without the addition of a cross-entropy loss. However, as we did not perform any fine tuning of hyperparameters for any of our models, it is possible that after sufficient fine tuning the performance of one of these alternative approaches might be improved.

To ensure the network generalizes to a wide age range, our results indicate that the distinctive features present before adulthood need to be adequately represented in the training data. This is evidenced by the degraded performance

observed when testing networks trained on 5-week old mice on the volumes acquired from older ones, and vice-versa. As mice are typically weaned at 3-4 weeks and attain sexual maturity at 8-12 weeks (Dutta & Sengupta 2016), 5-week old mice are not adults. In contrast, training solely on male mice did not significantly influence MU-Net performance on female animals. We studied why the Dice coefficient distributions were bi-modal with the large test set (see figure 24). The bi-modal nature of the distributions appears not to be explained by differences between different studies, genders, or genotypes (see Figures 39 and 41 in the appendix). We cannot offer a definitive explanation for the cause of these bi-modal distributions, however, we speculate that it is a sum of several factors, including intra-rater segmentation variability.

Interestingly, MU-Nets trained on automatic STEPS multi-atlas segmentation maps achieved higher Dice score with the ground truth than STEPS, highlighting the generalization ability of MU-Net. This supports the use of atlas based segmentation methods to augment MRI segmentation datasets suggested in Roy, Conjeti, Navab & Wachinger (2018), leveraging unlabeled data. The results obtained by training on STEPS segmentation maps alone remain, however, of insufficient quality to eliminate the need for manual annotations in the training data, as the CNN attempts to replicate any form of systematic error present in the atlas-based labeling procedure.

MU-Net-R was adjusted to reduce the number of parameters to better manage with the small training sets. The size of all kernels was reduced from the 5×5 to 3×3, and the overall number of convolution operations has been decreased in the first two blocks of the neural network, whereas MU-Net used 64 kernels for each convolution. We further replaced the Dice loss of MU-Net with the generalized Dice loss (see (Sudre *et al.* 2017)). Interestingly, while CNNs on average performed equally well or better than STEPS, they did so with a larger average precision but a lower average recall. This would indicate that STEPS prediction contained more false positives, labeling background voxels as belonging to ROIs, and conversely MU-Net's prediction favored false negatives and were biased towards the background. It follows that while designed to balance the different classes based on their size, the generalized Dice loss would still fail to entirely balance for the large background class. Future developments might benefit from implementing other loss function developed for highly unbalanced classes such as the Tversky loss (Abraham & Khan 2019) or the boundary loss (Kervadec, Bouchtiba, Desrosiers, Granger, Dolz & Ayed 2019).

11.3 BIOMARKERS

Over the course of weeks to months, TBI induces substantial atrophy to the brain, accompanied by a ventricle enlargement and changes in the position and orientation of periventricular structures such as the hippocampus. Here, our objective was to identify prognostic biomarkers for PTE using parameters derived from the repositioning of the hippocampus during the post-TBI aftermath. We extracted a set of parameters describing the position and orientation of the ipsilateral and contralateral hippocampus over the course of 5 post-TBI months, which overlaps with the evolution of PTE. Our data show that these parameters can discriminate between the sham-operated and TBI animals with a high sensitivity and specificity. At the more advanced stages of post-TBI structural alterations (i.e., 5 months post-TBI), the changed position and orientation of the hippocampus also differentiate the epileptic from non-epileptic animals.

11.3.1 TBI *vs.* sham classification

Previous histologic studies have shown progressive pathology in the principal cells and interneurons of the dentate gyrus and hippocampus proper after lateral Fluid Percussion Injury (FPI) -induced TBI (Pitkänen & Mcintosh 2006), and revealed progressive diffusion changes in the ipsilateral hippocampus for up to 3 months post-injury, which were substantially milder contralaterally (Immonen *et al.* 2009). As expected the hippocampal parameters were highly effective in differentiating TBI rats from the sham-operated animals at all time points. Interestingly the parameters contributing to the classification at different imaging time points varied. At early 2 d post-TBI time point, the most important parameters described the orientation of $\boldsymbol{P^1}$, that is, the plane cutting the ipsilateral hippocampus into two halves along its septo-temporal axis, indicating rotation. At 7 d, 21 d, 30 d and 150 d the reduction in ipsilateral/contralateral volume ratio (v_r) showed the best predictive value. The low-performance of the classifier at 9 d timepoint was likely due to overfitting. This was evident as the performance markedly increased when we limited the classification to the top 10 parameters by importance as calculated on the 7 d timepoint in the EPITARGET dataset, which was the closest timepoint to 9 d imaging in the EpiBioS4Rx cohort. Another factor contributing to increasing the difficulty at the 9-d timepoint would be the type of anesthesia: it appears that the use of a penthobarbital-based anesthesia cocktail for EPITARGET could have reduced the acute post-impact swelling compared to EpiBioS4Rx (4% isoflurane), resulting in a different time course, and acting as a confounder in the EpiBioS4Rx data.

11.3.2 Epileptic *vs.* non-epileptogenic classification

Next we assessed whether the parameters characterizing MRIs captured at an early post-injury time point could differentiate between the animals which were going to develop epilepsy from those who will not. No single variable was significantly effective in discriminating between TBI+ and TBI- animals after Bonferroni's correction, suggesting that a linear univariate model is not sufficiently complex to discriminate between the two categories. However, random forests effectively discriminated between the epileptogenic from the non-epileptogenic animals at the 150 d timepoint, based primarily on the geometric parameters describing both $\boldsymbol{P^1}$ planes, indicating the rotation of both hippocampi along their longitudinal septo-temporal axis. Unlike in sham *vs* TBI classification, volume reduction only played a minor role.The parameter importance evaluation seems to indicate that the ipsilateral hippocampus would be rotating outward, in the lateral direction. The most important feature however was found in the contralateral hippocampus, being deformed to bring its $\boldsymbol{P^1}$ away from the vertical direction. This change should not be understood as a rigid transantion, but as one of the overall results of a more complex deformation.

Our present findings are in agreement with previous studies, showing that hippocampal changes assessed at 6 to 9 months after TBI differentiate between the epileptogenic and non-epileptogenic animals, even though no statistics for differentiation accuracy was provided by earlier reports (Kharatishvili *et al.* 2007, Hayward *et al.* 2010, Shultz *et al.* 2013). Interestingly, in the same EPITARGET cohort analyzed here, thalamic diffusion changes already differentiated epileptogenic from non-epileptogenic animals during the first postinjury weeks (Manninen *et al.* 2021). Instead, the severity of cortical damage or its progression was without prognostic value (Manninen *et al.* 2020). These data suggest that whole-brain multimodal MRI analysis could prove highly valuable in the study of epileptogenesis.

11.4 LIMITATIONS

An obvious limitation of our approach is its specialization for the specific MRI contrast the algorithm is trained on. Making MU-Net to be more robust to marked changes in the image acquisition could be achieved by expanding the training data to be more variable or/and utilizing techniques such as domain adaptation, transfer learning (Valverde *et al.* 2021*a*) or image translation to minimize the amount of new training data for the model to generalize to new type of MRI acquisition (Armanious *et al.* 2020, Zhuang *et al.* 2020). This research line is one of the most important areas for future research in MRI

segmentation with deep learning. However, MU-Net successfully generalized to a variety of transgenic mice in an age range wider than that of the training set.

Another limitation of this study is the number of ROIs as mouse brain atlases with extremely detailed segmentation featuring over 700 ROIs currently exist (Nie *et al.* 2019). However, atlases such as (Nie *et al.* 2019) are constructed by specialized procedures and do not contain manual labels for all images used in the atlas construction. Therefore, these atlases are not directly applicable for training segmentation neural networks.

Even though the size of the entire EpiBioS4Rx and EPITARGET were larger than in most previous preclinical studies of TBI, the training datasets were small due to the cost of manual segmentation of MRIs. Their limited size as well as that of the NeAt dataset are another limitation of our work. It is reasonable to believe that larger training datasets would further enhance the performance of our neural network. This also provides an explanation for the higher standard deviation for HD95 distances for MU-Net compared to STEPS on NeAt data.

The size and features of our datasets also affected the search for biomarkers. While the full EPITARGET dataset was large, including 170 animals imaged at 3 timepoints, MRI was performed with a low resolution and at no timepoint later than 21 d. These shortcomings likely explain the lower number of parameters detected as significantly dependent on the time point in the repeated measures ANOVA, and prevented us from replicating our findings at 150 d in this cohort. Conversely, EpiBioS4Rx suffers from the smaller dataset size, including only 43 animals, which is likely one of the causes of the overfitting problems. It is also possible that for a larger, high-resolution dataset a biomarker for PTE might have been detected at earlier time points.

11.5 APPLICATIONS

Given the limitations discussed above there are two immediate use cases for segmentation with CNNs. The first is when dealing with large volumes of data acquired with the same pipeline. A laboratory generating hundreds of scans using the same sequences and hardware may benefit from either applying these methods "as is" or as the starting point for manual segmentation in a semi-automated approach. The latter is also helpful in dealing with the problem of data drift, that is the gradual shift of the domain of the data over time. In MRI this can be a results of changes in protocols, equipment, operators, or experimental subjects. Even in the absence of a more sophisticated method this can be amended by periodic retraining of the network utilizing the most recent manually-reviewed data.

The second use case is to replace registration-based methods where these would be hampered by the presence of significant anatomical alterations, or for the segmentation of object classes presenting an highly unstable appearance, such as the lesions themselves. In these cases registration-based methods may simply be entirely ineffective, and CNNs provide a valuable alternative.

11.6 CONCLUSION

The employment of CNNs for the segmentation of mouse brain MRI provides a number of benefits for preclinical researchers. Beyond allowing for the employment of large datasets in a time-efficient manner, the ability to generalize and abstract from the training data results in segmentation that are more robust to anatomical alterations, although robustness to changes in imaging modalities remains an open problem. We have demonstrated this robustness both in the case of Huntington and epilepsy animal models, efficiently producing human-level segmentation maps in both healthy and non-healthy animals.

In the future, we can expect these methods to reduce the confounding effect of intra- and inter-rater variability inherent in manual segmentation procedures while streamlining animal MRI experimental pipelines, without relying on the accurate alignment of a set of atlases with the target image.

Although we have limited our analysis to a specific network architecture, we do not assume this robustness to anatomical alternations to be a unique feature of our neural network, or even limited to the U-Net-like architectures. Because CNNs eliminate reliance on registration and replace it with encoding knowledge in the parameters of the network itself, and given their intrinsic properties of spatial invariance, we expect the general category of segmentation CNNs to be more robust to anatomical change than registration-based methods.

The method presented here for the extraction of geometric parameters characterizing the hippocampal anatomy is clinically translatable, and identified subjects at high-risk of post-traumatic epilepsy after experimental TBI. As the orientation of hippocampus in the temporal lobe and its exposure to TBI-related mechanical forces in humans differs from that in rat lateral FPI model, the parameters derived from the animal study are likely not directly applicable to human brain. However, our study provides a testable hypothesis that parameters reporting on the position and orientation of hippocampus bilaterally signal on the severity of brain pathology, and provide prognostic biomarkers for post-traumatic epileptogenesis beyond experimental models, paving the way towards subject stratification for antiepileptogenesis studies.

11 APPENDIX

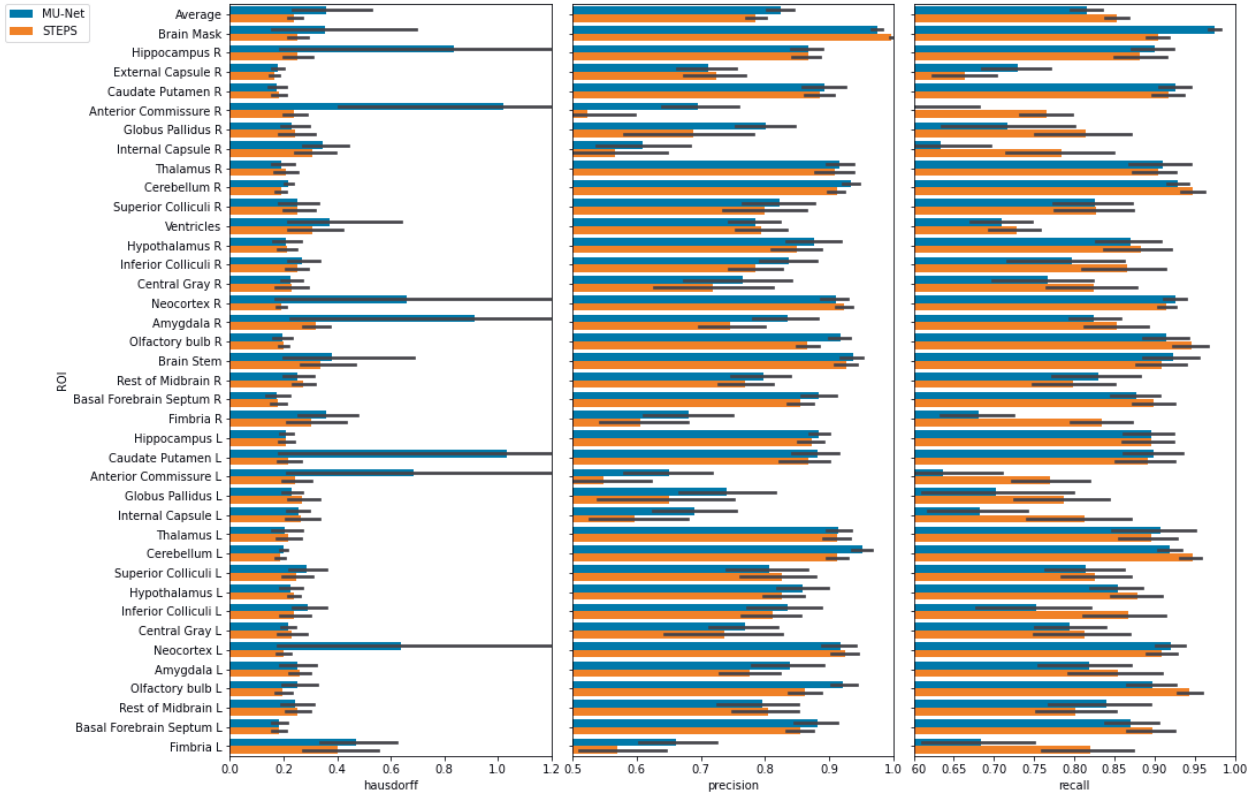

Figure 37. Hausdorff, precision and recall metrics for MU-Net and STEPS segmentation on the NeAt dataset. Image reproduced under CC license (De Feo *et al.* 2021).

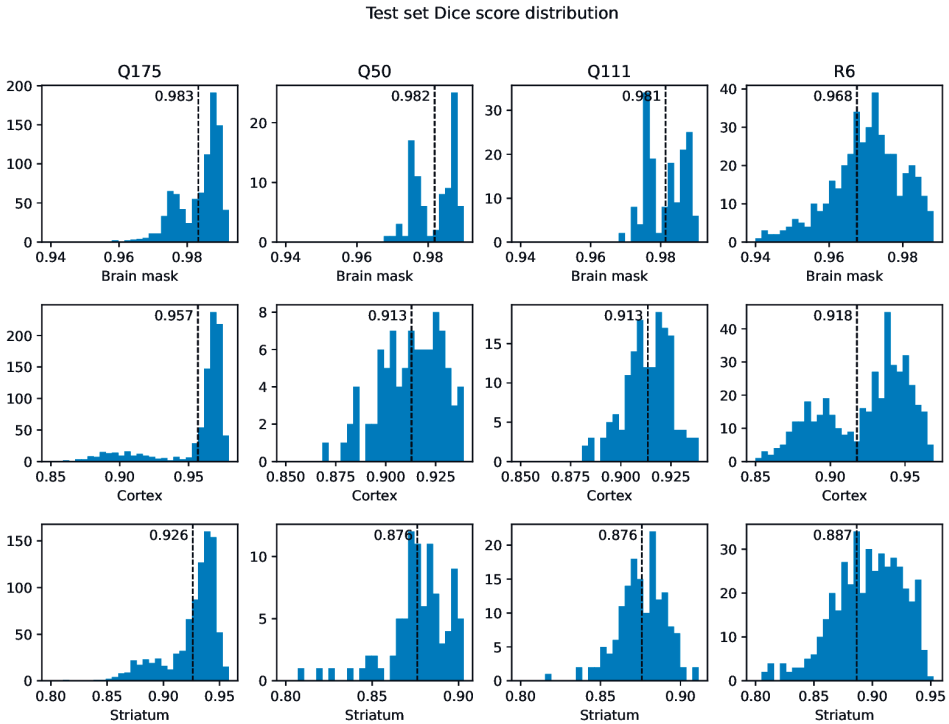

Figure 38. Breakdown of test score dice score distribution, by genotype, part 1. Image reproduced under CC license (De Feo *et al.* 2021).

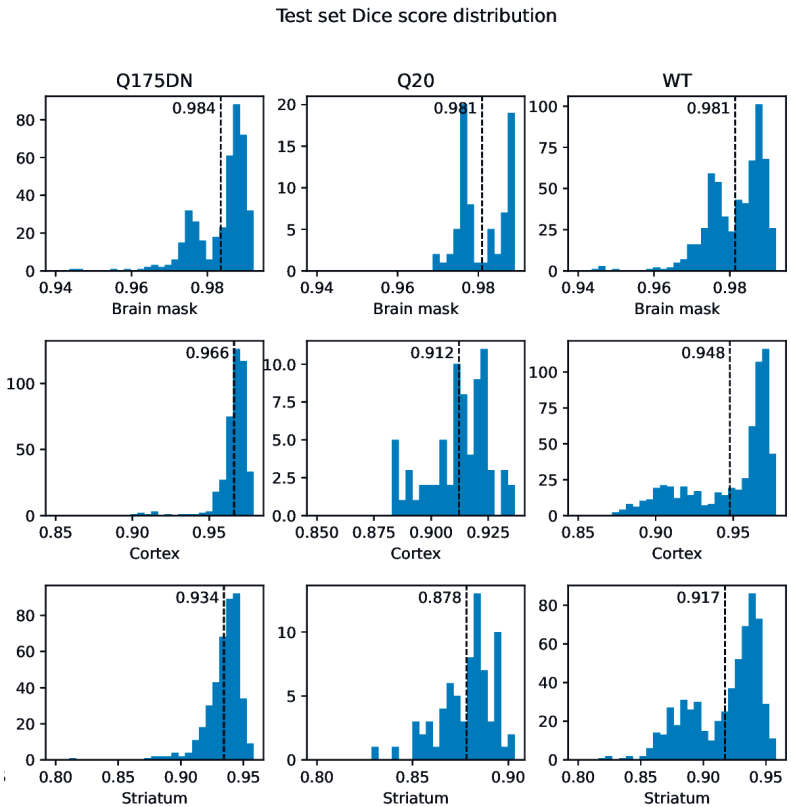

Figure 39. Breakdown of test score dice score distribution, by genotype, part 2. Image reproduced under CC license (De Feo *et al.* 2021).

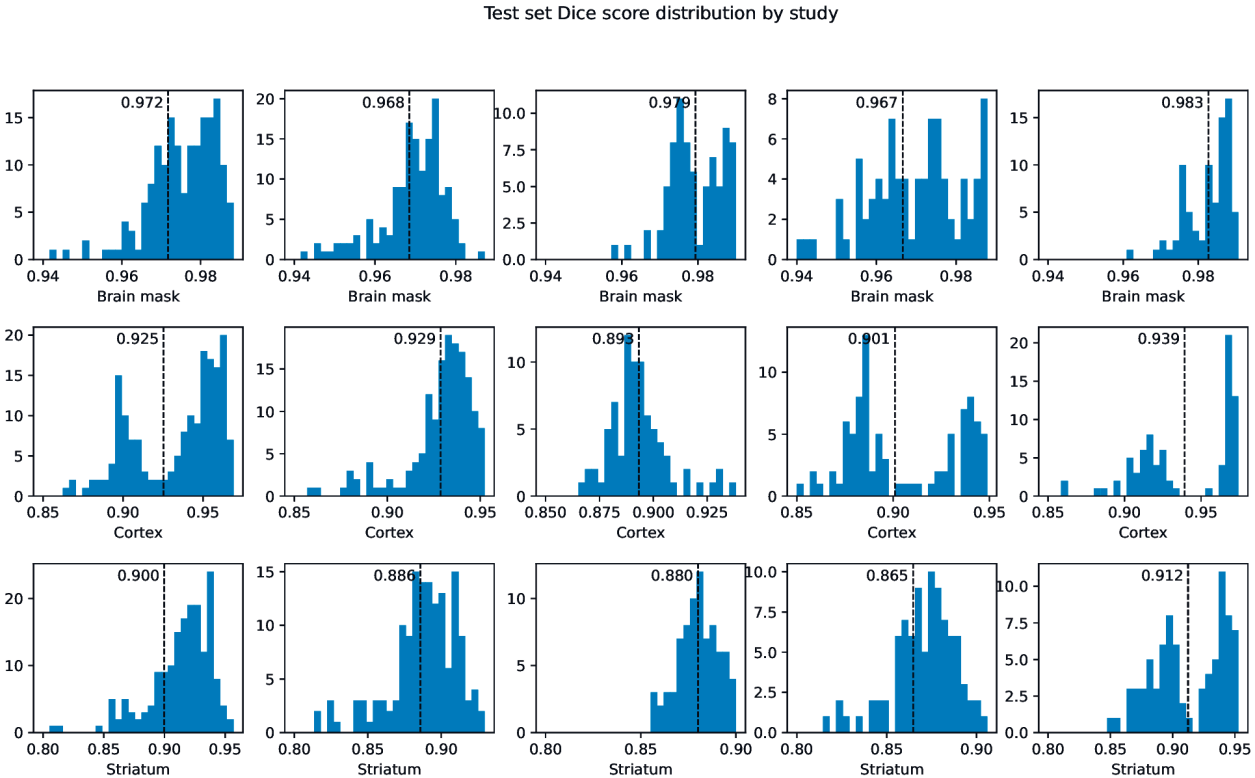

Figure 40. Breakdown of test dice score distribution, divided by each study included in the testing dataset, part 1. Image reproduced under CC license (De Feo *et al.* 2021).

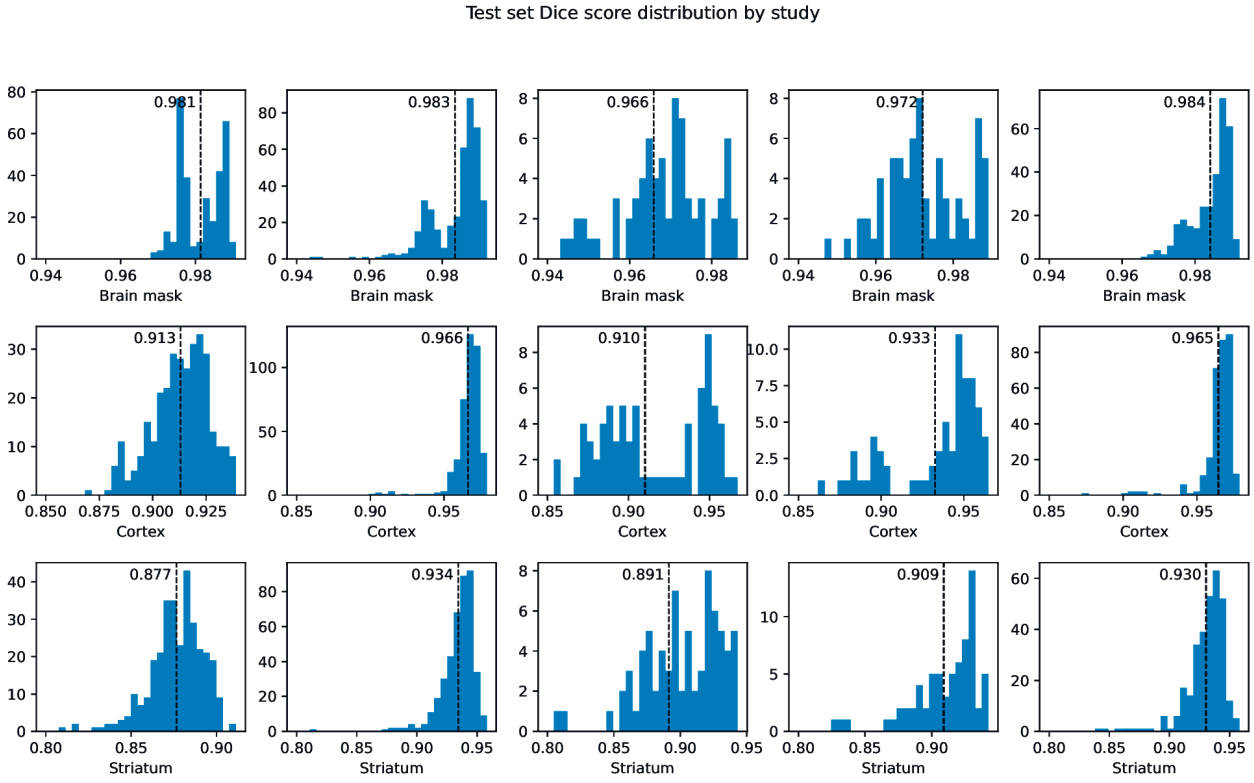

Figure 41. Breakdown of test dice score distribution, divided by each study included in the testing dataset, part 2. Image reproduced under CC license (De Feo *et al.* 2021).

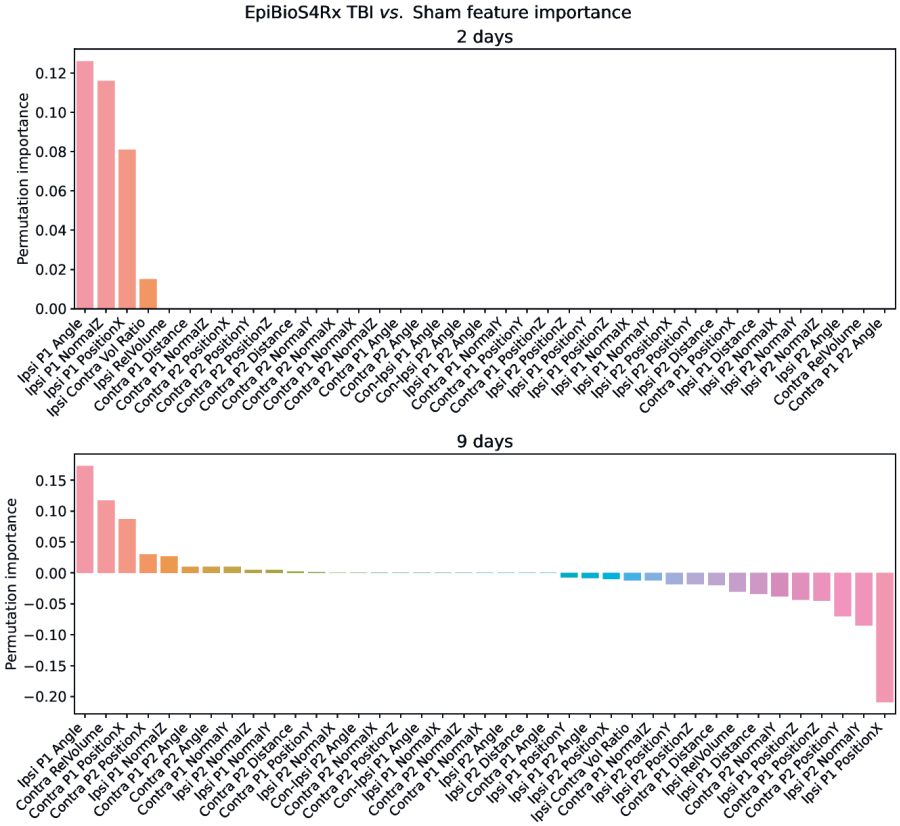

Figure 42. Parameter importance for the 2 and 9 d time points classification of Sham vs. TBI animals in the EpiBios4Rx dataset. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

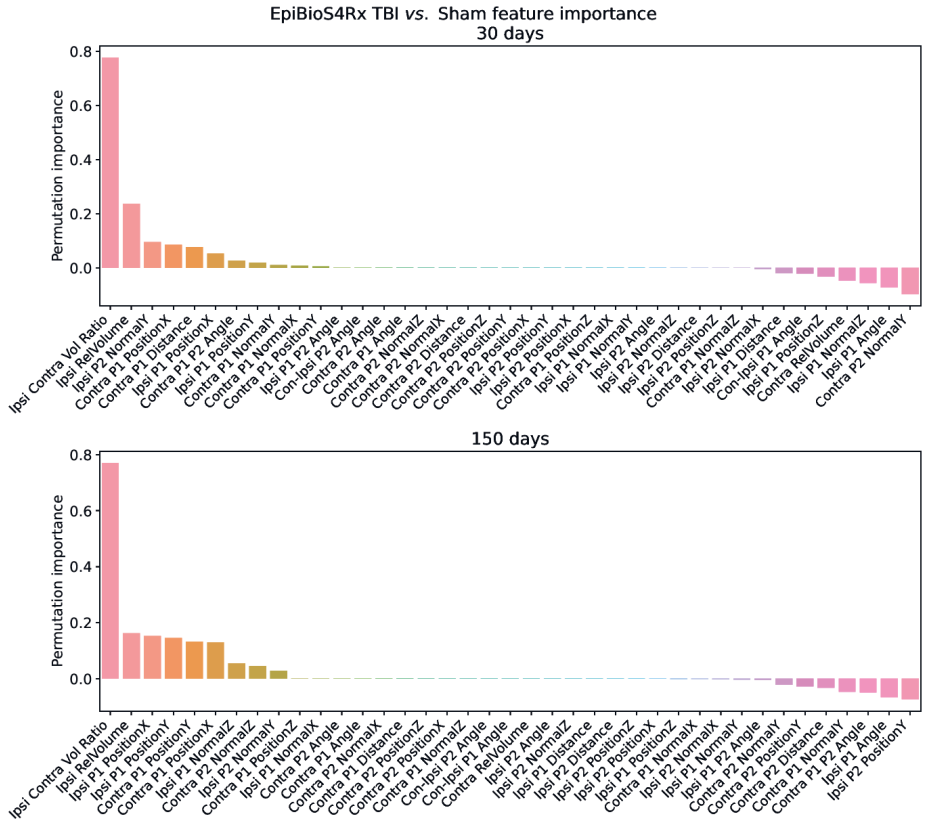

Figure 43. Parameter importance for the 30 and 150 d time points classification of Sham vs. TBI animals in the EpiBios4Rx dataset. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

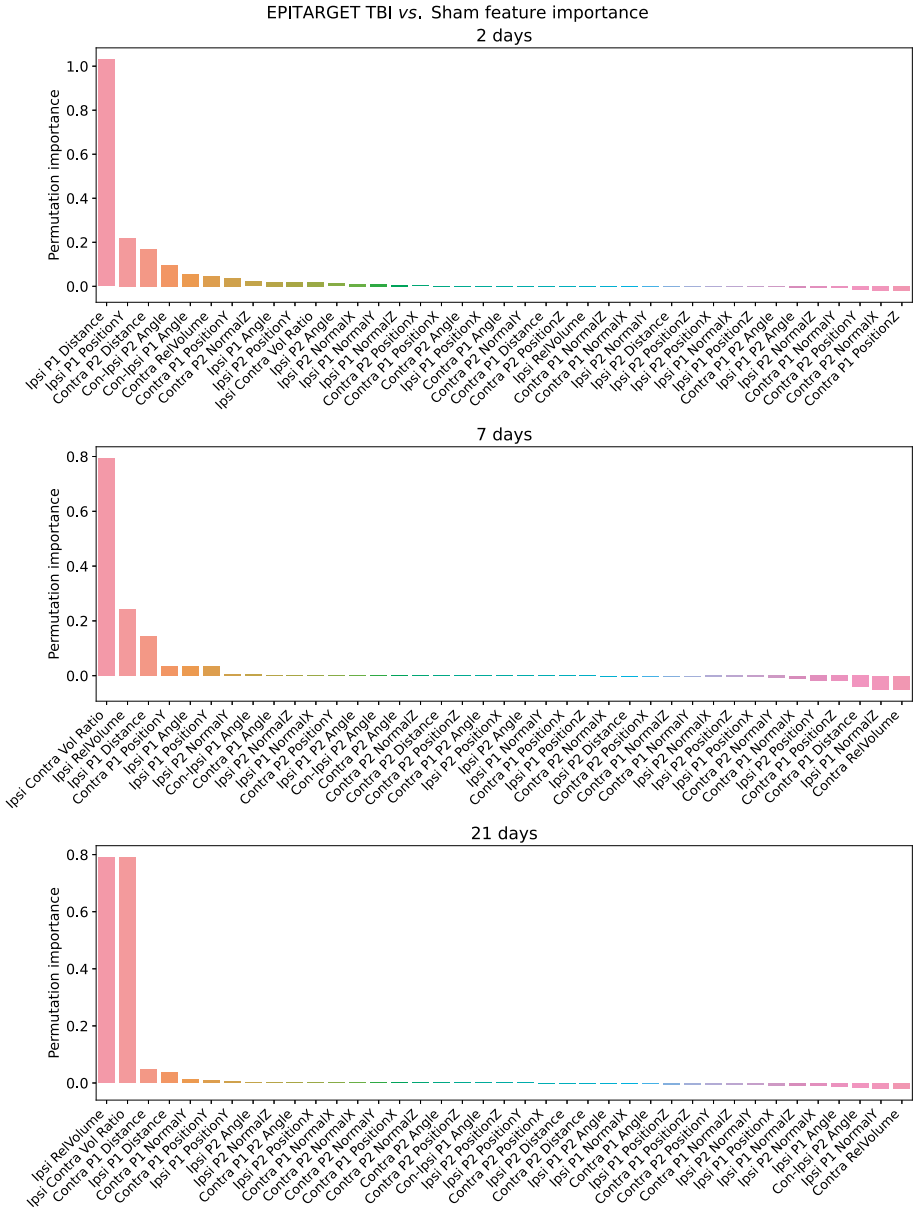

Figure 44. Parameter importance for the 2, 7 and 21 d time points classification of Sham vs. TBI animals in the EPITARGET dataset. Image reproduced under CC license (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022).

12 List of publications

- Riccardo De Feo and Federico Giove: **Towards an efficient segmentation of small rodents brain: a short critical review**. In: *Journal of neuroscience methods* 323 (2019), pp. 82–89 (De Feo & Giove 2019)

- Riccardo De Feo, Artem Shatillo, Alejandra Sierra, Juan Miguel Valverde, Olli Gröhn, Federico Giove, and Jussi Tohka: **Automated joint skull-stripping and segmentation with Multi-Task U-Net in large mouse brain MRI databases.**. In: *NeuroImage*, 2021, p. 117734. (De Feo *et al.* 2021)

- Riccardo De Feo, Elina Hämäläinen, Eppu Manninen, Riikka Immonen, Juan Miguel Valverde, Xavier Ekolle Ndode-Ekane, Olli Gröhn, Asla Pitkanen, and Jussi Tohka: **Convolutional neural networks enable robust automatic segmentation of the rat hippocampus in MRI after traumatic brain injury**. Accepted for publication in: *Frontiers in Neurology* (De Feo, Hämäläinen, Manninen, Immonen, Valverde, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

- Riccardo De Feo, Eppu Manninen, Karthik Chary, Elina Hämäläinen, Riikka Immonen, Pedro Andrade, Xavier Ekolle Ndode-Ekane, Olli Gröhn Asla Pitkänen, Jussi Tohka: **Hippocampal position and orientation as prognostic biomarkers for posttraumatic epileptogenesis: An experimental study in a rat lateral fluid percussion model**. In: *Epilepsia*, 2022 Apr 22. (De Feo, Manninen, Chary, Hämäläinen, Immonen, Andrade, Ndode-Ekane, Gröhn, Pitkänen & Tohka 2022)

- Valverde, Juan Miguel, Artem Shatillo, Riccardo De Feo, and Jussi Tohka: **Automatic Cerebral Hemisphere Segmentation in Rat MRI with Ischemic Lesions via Attention-based Convolutional Neural Networks**. In: *Neuroinformatics*, 2022: 1-14. (Valverde *et al.* 2022)

- Manninen, Eppu, Karthik Chary, Riccardo De Feo, Elina Hämäläinen, Pedro Andrade, Tomi Paananen, Alejandra Sierra, Jussi Tohka, Olli Gröhn, and Asla Pitkänen. **Acute Hippocampal Damage as a Prognostic Biomarker for Cognitive Decline but Not for Epileptogenesis after Experimental Traumatic Brain Injury**. *Biomedicines*, 10, no. 11 (2022): 2721. (Manninen *et al.* 2022)

- Valverde, J. M., Shatillo, A., De Feo, R., & Tohka, J.: **Automatic Cerebral Hemisphere Segmentation in Rat MRI with Ischemic**

Lesions via Attention-based Convolutional Neural Networks. In: *Neuroinformatics*, 2022 Sep 30:1-4. (Valverde *et al.* 2022)

- Timmins, K. M., van der Schaaf, I. C., Bennink, E., Ruigrok, Y. M., An, X., Baumgartner, M., ... & Kuijf, H. J.: **Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM Challenge**. In: *NeuroImage*, 2021, p. 118216. (Timmins *et al.* 2021)

- Valverde, J. M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R., & Tohka, J.: **Transfer Learning in Magnetic Resonance Brain Imaging: A Systematic Review**. In: *Journal of Imaging*, 2021, 7(4), 66. (Valverde *et al.* 2021*b*)

- Valverde, J. M., Shatillo, A., De Feo, R., Gröhn, O., Sierra, A., & Tohka, J.: **RatLesNetv2: A Fully Convolutional Network for Rodent Brain Lesion Segmentation**. In: *Frontiers in neuroscience*, 2020. (Valverde *et al.* 2020)

- Valverde, J. M., Shatillo, A., De Feo, R., Gröhn, O., Sierra, A., & Tohka, J.: **Automatic Rodent Brain MRI Lesion Segmentation with Fully Convolutional Networks**. In: *International Workshop on Machine Learning in Medical Imaging*, 2019, Springer, Cham (pp. 195-202). (Valverde *et al.* 2019)

- Di Trani, M. G., Nezzo, M., Caporale, A. S., De Feo, R., Miano, R., Mauriello, A., ... & Capuani, S: **Performance of Diffusion Kurtosis Imaging Versus Diffusion Tensor Imaging in Discriminating Between Benign Tissue, Low and High Gleason Grade Prostate Cancer**. In: *Academic radiology*, 2018. (Di Trani, Nezzo, Caporale, De Feo, Miano, Mauriello, Bove, Manenti & Capuani 2019)

- Di Trani, M. G., Manganaro, L., Antonelli, A., Guerreri, M., De Feo, R., Catalano, C., & Capuani, S.: **Apparent diffusion coefficient assessment of brain development in normal fetuses and ventriculomegaly**. In: *Frontiers of Physics*, 2019, n.7: 160. doi: 10.3389/fphy. (Di Trani, Manganaro, Antonelli, Guerreri, De Feo, Catalano & Capuani 2019)

13 Declarations

CODE AVAILABILITY STATEMENT

MU-Net code and trained models are freely available at `https://github.com/Hierakonpolis/MU-Net`. A tutorial of usage of MU-Net is available at `https://github.com/Hierakonpolis/NN4Kubiac`. The code for MU-NEt-R used in our work is freely available at `https://github.com/Hierakonpolis/MU-Net-R`, and the code utilized for biormarker research at `https://github.com/Hierakonpolis/RatHippocampusGeometry`. All code is released under the MIT license.

DATA AVAILABILITY STATEMENT

The Charles River (CR) dataset is property of Charles River Discovery Services, and the test dataset is property of CHDI 'Cure Huntington's Disease Initiative' foundation, which kindly decided to make this dataset available for these experiments. The MRM NeAt dataset is freely available at `https://github.com/dancebean/mouse-brain-atlas`. All the Dice scores between MU-Net and manual segmentations are available as supplementary files to the published paper (De Feo *et al.* 2021). The MRI data for the UEF datasets is stored on UEF servers and will be made available upon request.

ETHICS STATEMENT

Animal experiments in the CR datasets were carried out according to the United States National Institute of Health (NIH) guidelines for the care and use of laboratory animals, and approved by the National Animal Experiment Board. For University of Eastern Finland (UEF) datasets, all experiments were approved by the Animal Ethics Committee of the Provincial Government of Southern Finland and were performed in accordance with the guidelines of the European Community Directives 2010/63/EU.

FUNDING

My work has received funding from the European Union's Horizon 2020 Framework Programme under the Marie Skłodowska Curie grant agreement No #691110 (MICROBRADAM), the European Social Fund (grant S21770), and from the North Savo Regional Fund (grant 65211916).

COMPETING INTERESTS STATEMENT

The author declares no competing interests.

LICENSE

REFERENCES

Abbasi, B. & Goldenholz, D. M. (2019), 'Machine learning applications in epilepsy', *Epilepsia* **6**0(10), 2037–2047.

Abraham, N. & Khan, N. M. (2019), A novel focal tversky loss function with improved attention u-net for lesion segmentation, *in* '2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)', IEEE, pp. 683–687.

Aggarwal, M., Zhang, J., Miller, M. I., Sidman, R. L. & Mori, S. (2009), 'Magnetic resonance imaging and micro-computed tomography combined atlas of developing and adult mouse brains for stereotaxic surgery', *Neuroscience* **1**62(4), 1339–1350.

Ali, A. A., Dale, A. M., Badea, A. & Johnson, G. A. (2005), 'Automated segmentation of neuroanatomical structures in multispectral mr microscopy of the mouse brain', *Neuroimage* **2**7(2), 425–435.

Andersson, J. L., Jenkinson, M., Smith, S. *et al.* (2007), 'Non-linear registration aka spatial normalisation fmrib technial report tr07ja2', *FMRIB Analysis Group of the University of Oxford* .

Annegers, J. F., Hauser, W. A., Coan, S. P. & Rocca, W. A. (1998), 'A population-based study of seizures after traumatic brain injuries', *New England Journal of Medicine* **3**38(1), 20–24.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S. & Yang, B. (2020), 'Medgan: Medical image translation using gans', *Computerized Medical Imaging and Graphics* p. 101684.

Ashburner, J. (2007), 'A fast diffeomorphic image registration algorithm', *Neuroimage* **3**8(1), 95–113.

Ashburner, J. & Friston, K. J. (2005), 'Unified segmentation', *Neuroimage* **2**6(3), 839–851.

Ashburner, J. & Friston, K. J. (2011), 'Diffeomorphic registration using geodesic shooting and gauss–newton optimisation', *NeuroImage* **5**5(3), 954–967.

Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. (2008), 'Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain', *Medical image analysis* **1**2(1), 26–41.

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* **5**4(3), 2033–2044.

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A. & Gee, J. C. (2011), 'An open source multivariate framework for n-tissue segmentation with evaluation

on public data', *Neuroinformatics* **9**(4), 381–400.
Avants, B. B., Tustison, N. & Song, G. (2009), 'Advanced normalization tools (ants)', *Insight j* **2**, 1–35.
Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J. & Gee, J. C. (2010), 'The optimal template effect in hippocampus studies of diseased populations', *Neuroimage* **4**9(3), 2457–2466.
Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017), 'Segnet: A deep convolutional encoder-decoder architecture for image segmentation', *IEEE transactions on pattern analysis and machine intelligence* **3**9(12), 2481–2495.
Bae, M. H., Pan, R., Wu, T. & Badea, A. (2009), 'Automated segmentation of mouse brain images using extended mrf', *Neuroimage* **4**6(3), 717–725.
Bai, J., Trinh, T. L. H., Chuang, K.-H. & Qiu, A. (2012), 'Atlas-based automatic mouse brain image segmentation revisited: model complexity vs. image registration', *Magnetic resonance imaging* **3**0(6), 789–798.
Beg, M. F., Miller, M. I., Trouvé, A. & Younes, L. (2005), 'Computing large deformation metric mappings via geodesic flows of diffeomorphisms', *International journal of computer vision* **6**1(2), 139–157.
Billot, B., Greve, D., Van Leemput, K., Fischl, B., Iglesias, J. E. & Dalca, A. V. (2020), 'A learning strategy for contrast-agnostic mri segmentation', *arXiv preprint arXiv:2003.01995* .
Bragin, A., Li, L., Almajano, J., Alvarado-Rojas, C., Reid, A. Y., Staba, R. J. & Engel Jr, J. (2016), 'Pathologic electrographic changes after experimental traumatic brain injury', *Epilepsia* **5**7(5), 735–745.
Breiman, L. (2001), 'Random forests', *Machine learning* **4**5(1), 5–32.
Bribiesca, E. (2008), 'An easy measure of compactness for 2d and 3d shapes', *Pattern Recognition* **4**1(2), 543–554.
Campbell, J. N., Gandhi, A., Singh, B. & Churn, S. B. (2014), 'Traumatic brain injury causes a tacrolimus-sensitive increase in non-convulsive seizures in a rat model of post-traumatic epilepsy', *International journal of neurology & brain disorders* **1**(1), 1.
Cardoso, M. J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N. C., Ourselin, S., Initiative, A. D. N. *et al.* (2013), 'Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation', *Medical image analysis* **1**7(6), 671–684.
Cardoso, M. J., Modat, M., Ourselin, S., Keihaninejad, S. & Cash, D. (2012), Steps: multi-label similarity and truth estimation for propagated segmentations, *in* 'Mathematical Methods in Biomedical Image Analysis (MMBIA), 2012 IEEE Workshop on', IEEE, pp. 153–158.
Chou, N., Wu, J., Bingren, J. B., Qiu, A. & Chuang, K.-H. (2011), 'Robust automatic rodent brain extraction using 3-d pulse-coupled neural networks

(pcnn)', *IEEE Transactions on Image Processing* **2**0(9), 2554–2564.
Christensen, J. (2012), 'Traumatic brain injury: risks of epilepsy and implications for medicolegal assessment', *Epilepsia* **5**3, 43–47.
Chuang, N., Mori, S., Yamamoto, A., Jiang, H., Ye, X., Xu, X., Richards, L. J., Nathans, J., Miller, M. I., Toga, A. W. *et al.* (2011), 'An mri-based atlas and database of the developing mouse brain', *Neuroimage* **5**4(1), 80–89.
Chung, E., Romano, J. P. *et al.* (2013), 'Exact and asymptotically robust permutation tests', *Annals of Statistics* **4**1(2), 484–507.
Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016), 3d u-net: learning dense volumetric segmentation from sparse annotation, *in* 'International conference on medical image computing and computer-assisted intervention', Springer, pp. 424–432.
Collins, D., Neelin, P., Peters, T. & Evans, A. (1994), 'Automatic 3d intersubject registration of mr volumetric data in standardized talairach space j comput assist tomogr 1994; 18 (2): 192-205'.
Conn, A. R., Gould, N. I. & Toint, P. L. (2000), *Trust region methods*, SIAM.
Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **2**0(3), 273–297.
Dalca, A. V., Balakrishnan, G., Guttag, J. & Sabuncu, M. R. (2018), 'Unsupervised learning for fast probabilistic diffeomorphic registration', *arXiv preprint arXiv:1805.04605* .
De Feo, R. & Giove, F. (2019), 'Towards an efficient segmentation of small rodents brain: a short critical review', *Journal of neuroscience methods* **3**23, 82–89.
De Feo, R., Hämäläinen, E., Manninen, E., Immonen, R., Valverde, J. M., Ndode-Ekane, X. E., Gröhn, O., Pitkänen, A. & Tohka, J. (2022), 'Convolutional neural networks enable robust automatic segmentation of the rat hippocampus in mri after traumatic brain injury', *Frontiers in Neurology* **1**3.
URL: *https://www.frontiersin.org/article/10.3389/fneur.2022.820267*
De Feo, R., Manninen, E., Chary, K., Hämäläinen, E., Immonen, R., Andrade, P., Ndode-Ekane, X. E., Gröhn, O., Pitkänen, A. & Tohka, J. (2022), 'Hippocampal position and orientation as prognostic biomarkers for posttraumatic epileptogenesis: An experimental study in a rat lateral fluid percussion model', *Epilepsia* .
De Feo, R., Shatillo, A., Sierra, A., Valverde, J. M., Gröhn, O., Giove, F. & Tohka, J. (2021), 'Automated joint skull-stripping and segmentation with multi-task u-net in large mouse brain mri databases', *NeuroImage* p. 117734.
Delora, A., Gonzales, A., Medina, C. S., Mitchell, A., Mohed, A. F., Jacobs, R. E. & Bearer, E. L. (2016), 'A simple rapid process for semi-automated brain extraction from magnetic resonance images of the whole mouse head',

Journal of neuroscience methods **2**57, 185–193.
Di Trani, M. G., Manganaro, L., Antonelli, A., Guerreri, M., De Feo, R., Catalano, C. & Capuani, S. (2019), 'Apparent diffusion coefficient assessment of brain development in normal fetuses and ventriculomegaly', *Frontiers in Physics* **7**, 160.
Di Trani, M. G., Nezzo, M., Caporale, A. S., De Feo, R., Miano, R., Mauriello, A., Bove, P., Manenti, G. & Capuani, S. (2019), 'Performance of diffusion kurtosis imaging versus diffusion tensor imaging in discriminating between benign tissue, low and high gleason grade prostate cancer', *Academic Radiology* **2**6(10), 1328–1337.
Dice, L. R. (1945), 'Measures of the amount of ecologic association between species', *Ecology* **2**6(3), 297–302.
Dorr, A., Lerch, J. P., Spring, S., Kabani, N. & Henkelman, R. M. (2008), 'High resolution three-dimensional brain atlas using an average magnetic resonance image of 40 adult c57bl/6j mice', *Neuroimage* **4**2(1), 60–69.
Douglass, M. J. (2020), 'Book review: Hands-on machine learning with scikit-learn, keras, and tensorflow, by aurélien géron'.
Dulla, C. G. & Pitkänen, A. (2021), 'Novel approaches to prevent epileptogenesis after traumatic brain injury', *Neurotherapeutics* pp. 1–20.
Dutta, S. & Sengupta, P. (2016), 'Men and mice: relating their ages', *Life sciences* **1**52, 244–248.
Engel Jr, J., Pitkänen, A., Loeb, J. A., Edward Dudek, F., Bertram III, E. H., Cole, A. J., Moshé, S. L., Wiebe, S., Jensen, F. E., Mody, I. *et al.* (2013), 'Epilepsy biomarkers', *Epilepsia* **5**4, 61–69.
Entis, J. J., Doerga, P., Barrett, L. F. & Dickerson, B. C. (2012), 'A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution mri', *Neuroimage* **6**0(2), 1226–1235.
Fu, Z., Lin, L. & Jin, C. (2016), Symmetric image normalization for mouse brain magnetic resonance microscopy, *in* 'International Conference on Advances in Mechanical Engineering and Industrial Informatics. Zheng Zhou, China'.
Fu, Z., Lin, L., Tian, M., Wang, J., Zhang, B., Chu, P., Li, S., Pathan, M. M., Deng, Y. & Wu, S. (2017), 'Evaluation of five diffeomorphic image registration algorithms for mouse brain magnetic resonance microscopy', *Journal of microscopy* **2**68(2), 141–154.
Greenham, S., Dean, J., Fu, C. K. K., Goman, J., Mulligan, J., Tune, D., Sampson, D., Westhuyzen, J. & McKay, M. (2014), 'Evaluation of atlas-based auto-segmentation software in prostate cancer patients', *Journal of medical radiation sciences* **6**1(3), 151–158.
Greenhouse, S. W. & Geisser, S. (1959), 'On methods in the analysis of profile data', *Psychometrika* **2**4(2), 95–112.
Gupta, P. K., Sayed, N., Ding, K., Agostini, M. A., Van Ness, P. C., Yablon,

S., Madden, C., Mickey, B., D'Ambrosio, R. & Diaz-Arrastia, R. (2014), 'Subtypes of post-traumatic epilepsy: clinical, electrophysiological, and imaging features', *Journal of neurotrauma* **3**1(16), 1439–1443.

Gutierrez-Becker, B., Mateus, D., Peter, L. & Navab, N. (2017), 'Guiding multimodal registration with learned optimization updates', *Medical image analysis* **4**1, 2–17.

Haltiner, A. M., Temkin, N. R. & Dikmen, S. S. (1997), 'Risk of seizure recurrence after the first late posttraumatic seizure', *Archives of physical medicine and rehabilitation* **7**8(8), 835–840.

Han, Y. & Ye, J. C. (2018), 'Framing u-net via deep convolutional framelets: Application to sparse-view ct', *IEEE transactions on medical imaging* **3**7(6), 1418–1429.

Hawrylycz, M., Baldock, R. A., Burger, A., Hashikawa, T., Johnson, G. A., Martone, M., Ng, L., Lau, C., Larsen, S. D., Nissanov, J. *et al.* (2011), 'Digital atlasing and standardization in the mouse brain', *PLoS computational biology* **7**(2), e1001065.

Hayward, N. M., Immonen, R., Tuunanen, P. I., Ndode-Ekane, X. E., Gröhn, O. & Pitkänen, A. (2010), 'Association of chronic vascular changes with functional outcome after traumatic brain injury in rats', *Journal of neurotrauma* **2**7(12), 2203–2219.

Hjornevik, T., Leergaard, T. B., Darine, D., Moldestad, O., Dale, A. M., Willoch, F. & Bjaalie, J. G. (2007), 'Three-dimensional atlas system for mouse and rat brain imaging data', *Frontiers in Neuroinformatics* **1**, 4.

Hoffmann, M., Billot, B., Greve, D. N., Iglesias, J. E., Fischl, B. & Dalca, A. V. (2021), 'Synthmorph: learning contrast-invariant registration without acquired images', *IEEE Transactions on Medical Imaging* .

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. *et al.* (2020), 'Automatic skull stripping of rat and mouse brain mri data using u-net', *Frontiers in neuroscience* **1**4.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017), Densely connected convolutional networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 4700–4708.

Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. (1993), 'Comparing images using the hausdorff distance', *IEEE Transactions on pattern analysis and machine intelligence* **1**5(9), 850–863.

Huttunen, J. K., Airaksinen, A. M., Barba, C., Colicchio, G., Niskanen, J.-P., Shatillo, A., Sierra Lopez, A., Ndode-Ekane, X. E., Pitkänen, A. & Gröhn, O. H. (2018), 'Detection of hyperexcitability by functional magnetic resonance imaging after experimental traumatic brain injury', *Journal of neurotrauma* **3**5(22), 2708–2717.

Iglesias, J. E. & Sabuncu, M. R. (2015), 'Multi-atlas segmentation of biomedical images: a survey', *Medical image analysis* **2**4(1), 205–219.

Immonen, R. J., Kharatishvili, I., Niskanen, J.-P., Gröhn, H., Pitkänen, A. & Gröhn, O. H. (2009), 'Distinct mri pattern in lesional and perilesional area after traumatic brain injury in rat—11 months follow-up', *Experimental neurology* **2**15(1), 29–40.

Immonen, R., Smith, G., Brady, R. D., Wright, D., Johnston, L., Harris, N. G., Manninen, E., Salo, R., Branch, C., Duncan, D. *et al.* (2019), 'Harmonization of pipeline for preclinical multicenter mri biomarker discovery in a rat model of post-traumatic epileptogenesis', *Epilepsy research* **1**50, 46–57.

Ioffe, S. & Szegedy, C. (2015), 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', *arXiv preprint arXiv:1502.03167* .

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S. *et al.* (2018), 'nnu-net: Self-adapting framework for u-net-based medical image segmentation', *arXiv preprint arXiv:1809.10486* .

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017), Image-to-image translation with conditional adversarial networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1125–1134.

Jaccard, P. (1912), 'The distribution of the flora in the alpine zone. 1', *New phytologist* **1**1(2), 37–50.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. (2012), 'Fsl', *Neuroimage* **6**2(2), 782–790.

Jenkinson, M. & Smith, S. (2001), 'A global optimisation method for robust affine registration of brain images', *Medical image analysis* **5**(2), 143–156.

Johnson, G. A., Badea, A., Brandenburg, J., Cofer, G., Fubara, B., Liu, S. & Nissanov, J. (2010), 'Waxholm space: an image-based reference for coordinating mouse brain research', *Neuroimage* **5**3(2), 365–372.

Johnson, G. A., Calabrese, E., Badea, A., Paxinos, G. & Watson, C. (2012), 'A multidimensional magnetic resonance histology atlas of the wistar rat brain', *Neuroimage* **6**2(3), 1848–1856.

Jonckers, E., Van Audekerke, J., De Visscher, G., Van der Linden, A. & Verhoye, M. (2011), 'Functional connectivity fmri of the rodent brain: comparison of functional connectivity networks in rat and mouse', *PloS one* **6**(4), e18876.

Karimi, D. & Salcudean, S. E. (2019), 'Reducing the hausdorff distance in medical image segmentation with convolutional neural networks', *arXiv preprint arXiv:1904.10030* .

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J. & Ayed, I. B. (2019), Boundary loss for highly unbalanced segmentation, *in* 'International conference on medical imaging with deep learning', PMLR, pp. 285–296.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ayed, I. B. (2019),

'Constrained-cnn losses for weakly supervised segmentation', *Medical image analysis* **5**4, 88–99.

Kharatishvili, I., Immonen, R., Gröhn, O. & Pitkänen, A. (2007), 'Quantitative diffusion mri of hippocampus as a surrogate marker for post-traumatic epileptogenesis', *Brain* **1**30(12), 3155–3168.

Kharatishvili, I., Nissinen, J., McIntosh, T. & Pitkänen, A. (2006), 'A model of posttraumatic epilepsy induced by lateral fluid-percussion brain injury in rats', *Neuroscience* **1**40(2), 685–697.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kjonigsen, L. J., Lillehaug, S., Bjaalie, J. G., Witter, M. P. & Leergaard, T. B. (2015), 'Waxholm space atlas of the rat brain hippocampal region: three-dimensional delineations based on magnetic resonance and diffusion tensor imaging', *Neuroimage* **1**08, 441–449.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P. *et al.* (2009), 'Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration', *Neuroimage* **4**6(3), 786–802.

Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. (2010), 'Elastix: a toolbox for intensity-based medical image registration', *IEEE transactions on medical imaging* **2**9(1), 196–205.

Kochunov, P., Lancaster, J. L., Thompson, P., Woods, R., Mazziotta, J., Hardies, J. & Fox, P. (2001), 'Regional spatial normalization: toward an optimal target', *Journal of computer assisted tomography* **2**5(5), 805–816.

Kovačević, N., Henderson, J., Chan, E., Lifshitz, N., Bishop, J., Evans, A., Henkelman, R. & Chen, X. (2004), 'A three-dimensional mri atlas of the mouse brain with estimates of the average and variability', *Cerebral cortex* **1**5(5), 639–645.

Lancelot, S., Roche, R., Slimen, A., Bouillot, C., Levigoureux, E., Langlois, J.-B., Zimmer, L. & Costes, N. (2014), 'A multi-atlas based method for automated anatomical rat brain mri segmentation and extraction of pet activity', *PloS one* **9**(10), e109113.

Lapinlampi, N., Andrade, P., Paananen, T., Hämäläinen, E., Ekolle Ndode-Ekane, X., Puhakka, N. & Pitkänen, A. (2020), 'Postinjury weight rather than cognitive or behavioral impairment predicts development of posttraumatic epilepsy after lateral fluid-percussion injury in rats', *Epilepsia* **6**1(9), 2035–2052.

Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C. A., Pappata, S., Molko, N. & Chabriat, H. (2001), 'Diffusion tensor imaging: concepts and applications', *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **1**3(4), 534–546.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *nature* **5**21(7553), 436.

Lee, J., Lyu, I. & Styner, M. (2014), Multi-atlas segmentation with particle-based group-wise image registration, *in* 'Medical Imaging 2014: Image Processing', Vol. 9034, International Society for Optics and Photonics, p. 903447.

Leung, K. K., Barnes, J., Modat, M., Ridgway, G. R., Bartlett, J. W., Fox, N. C., Ourselin, S., Initiative, A. D. N. *et al.* (2011), 'Brain maps: an automated, accurate and robust brain extraction technique using a template library', *Neuroimage* **5**5(3), 1091–1108.

Li, J., Liu, X., Zhuo, J., Gullapalli, R. P. & Zara, J. M. (2013), 'An automatic rat brain extraction method based on a deformable surface model', *Journal of neuroscience methods* **2**18(1), 72–82.

Li, Q., Cheung, C., Wei, R., Hui, E. S., Feldon, J., Meyer, U., Chung, S., Chua, S. E., Sham, P. C., Wu, E. X. *et al.* (2009), 'Prenatal immune challenge is an environmental risk factor for brain and behavior change relevant to schizophrenia: evidence from mri in a mouse model', *PloS one* **4**(7), e6354.

Li, S. Z. (1994), Markov random field models in computer vision, *in* 'European conference on computer vision', Springer, pp. 361–370.

Liang, S., Wu, S., Huang, Q., Duan, S., Liu, H., Li, Y., Zhao, S., Nie, B. & Shan, B. (2017), 'Rat brain digital stereotaxic white matter atlas with fine tract delineation in paxinos space and its automated applications in dti data analysis', *Magnetic resonance imaging* **4**3, 122–128.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. & Han, J. (2019), 'On the variance of the adaptive learning rate and beyond', *arXiv preprint arXiv:1908.03265* .

Livne, M., Rieger, J., Aydin, O. U., Taha, A. A., Akay, E. M., Kossen, T., Sobesky, J., Kelleher, J. D., Hildebrand, K., Frey, D. *et al.* (2019), 'A u-net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease', *Frontiers in neuroscience* **1**3, 97.

Lundberg, S. M. & Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Proceedings of the 31st international conference on neural information processing systems', pp. 4768–4777.

Ma, D., Cardoso, M. J., Modat, M., Powell, N., Wells, J., Holmes, H., Wiseman, F., Tybulewicz, V., Fisher, E., Lythgoe, M. F. *et al.* (2014), 'Automatic structural parcellation of mouse brain mri using multi-atlas label fusion', *PloS one* **9**(1), e86576.

Ma, D., Cardoso, M., Modat, M., Powell, N., Holmes, H., Lythgoe, M. & Ourselin, S. (2012), Multi atlas segmentation applied to in vivo mouse brain mri, MICCAI 2012 Workshop on Multi-Atlas Labeling.

Ma, Y., Hof, P., Grant, S., Blackband, S., Bennett, R., Slatest, L., McGuigan,

M. & Benveniste, H. (2005), 'A three-dimensional digital atlas database of the adult c57bl/6j mouse brain by magnetic resonance microscopy', *Neuroscience* **1**35(4), 1203–1215.

Ma, Y., Smith, D., Hof, P. R., Foerster, B., Hamilton, S., Blackband, S. J., Yu, M. & Benveniste, H. (2008), 'In vivo 3d digital atlas database of the adult c57bl/6j mouse brain by magnetic resonance microscopy', *Frontiers in neuroanatomy* **2**, 1.

Maas, A. L., Hannun, A. Y. & Ng, A. Y. (2013), Rectifier nonlinearities improve neural network acoustic models, *in* 'Proc. icml', Vol. 30, p. 3.

Manninen, E., Chary, K., De Feo, R., Hämäläinen, E., Andrade, P., Paananen, T., Sierra, A., Tohka, J., Gröhn, O. & Pitkänen, A. (2022), 'Acute hippocampal damage as a prognostic biomarker for cognitive decline but not for epileptogenesis after experimental traumatic brain injury', *Biomedicines* **1**0(11), 2721.

Manninen, E., Chary, K., Lapinlampi, N., Andrade, P., Paananen, T., Sierra, A., Tohka, J., Gröhn, O. & Pitkänen, A. (2020), 'Early increase in cortical t2 relaxation is a prognostic biomarker for the evolution of severe cortical damage, but not for epileptogenesis, after experimental traumatic brain injury', *Journal of neurotrauma* **3**7(23), 2580–2594.

Manninen, E., Chary, K., Lapinlampi, N., Andrade, P., Paananen, T., Sierra, A., Tohka, J., Gröhn, O. & Pitkänen, A. (2021), 'Acute thalamic damage as a prognostic biomarker for post-traumatic epileptogenesis', *Epilepsia* **6**2(8), 1852–1864.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016), V-net: Fully convolutional neural networks for volumetric medical image segmentation, *in* '2016 Fourth International Conference on 3D Vision (3DV)', IEEE, pp. 565–571.

Murugavel, M. & Sullivan Jr, J. M. (2009), 'Automatic cropping of mri rat brain volumes using pulse coupled neural networks', *Neuroimage* **4**5(3), 845–854.

Ndode-Ekane, X. E., Santana-Gomez, C., Casillas-Espinosa, P. M., Ali, I., Brady, R. D., Smith, G., Andrade, P., Immonen, R., Puhakka, N., Hudson, M. R. *et al.* (2019), 'Harmonization of lateral fluid-percussion injury model production and post-injury monitoring in a preclinical multicenter biomarker discovery study on post-traumatic epileptogenesis', *Epilepsy research* **1**51, 7–16.

Nie, B., Liu, H., Chen, K., Jiang, X. & Shan, B. (2014), 'A statistical parametric mapping toolbox used for voxel-wise analysis of fdg-pet images of rat brain', *PloS one* **9**(9), e108295.

Nie, B., Wu, D., Liang, S., Liu, H., Sun, X., Li, P., Huang, Q., Zhang, T., Feng, T., Ye, S. *et al.* (2019), 'A stereotaxic mri template set of mouse brain with fine sub-anatomical delineations: Application to memri studies of 5xfad mice', *Magnetic Resonance Imaging* **5**7, 83–94.

Nie, J. & Shen, D. (2013), 'Automated segmentation of mouse brain images using multi-atlas multi-roi deformation and label fusion', *Neuroinformatics* **1**1(1), 35–45.

Nissinen, J., Andrade, P., Natunen, T., Hiltunen, M., Malm, T., Kanninen, K., Soares, J. I., Shatillo, O., Sallinen, J., Ndode-Ekane, X. E. *et al.* (2017), 'Disease-modifying effect of atipamezole in a model of post-traumatic epilepsy', *Epilepsy research* **1**36, 18–34.

Noh, H., Hong, S. & Han, B. (2015), Learning deconvolution network for semantic segmentation, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 1520–1528.

Oguz, I., Zhang, H., Rumple, A. & Sonka, M. (2014), 'Rats: rapid automatic tissue segmentation in rodent brain mri', *Journal of neuroscience methods* **2**21, 175–182.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B. *et al.* (2018), 'Attention u-net: Learning where to look for the pancreas', *arXiv preprint arXiv:1804.03999* .

Pagani, M., Damiano, M., Galbusera, A., Tsaftaris, S. A. & Gozzi, A. (2016), 'Semi-automated registration-based anatomical labelling, voxel based morphometry and cortical thickness mapping of the mouse brain', *Journal of neuroscience methods* **2**67, 62–73.

Papandreou, G., Chen, L.-C., Murphy, K. P. & Yuille, A. L. (2015), Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 1742–1750.

Papp, E. A., Leergaard, T. B., Calabrese, E., Johnson, G. A. & Bjaalie, J. G. (2014), 'Waxholm space atlas of the sprague dawley rat brain', *Neuroimage* **9**7, 374–386.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017), 'Automatic differentiation in pytorch'.

Paxinos, G. & Franklin, K. B. (2004), *The mouse brain in stereotaxic coordinates*, Gulf professional publishing.

Paxinos, G. & Watson, C. (1986), *The rat brain in stereotaxic coordinates / George Paxinos, Charles Watson*, 2nd ed. edn, Academic Press Sydney.

Paxinos, G. & Watson, C. (2006), *The rat brain in stereotaxic coordinates: hard cover edition*, Elsevier.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*

12, 2825–2830.
Pelt, D. M. & Sethian, J. A. (2018), 'A mixed-scale dense convolutional neural network for image analysis', *Proceedings of the National Academy of Sciences* **1**15(2), 254–259.
Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J. & Nichols, T. E. (2011), *Statistical parametric mapping: the analysis of functional brain images*, Elsevier.
Pitkänen, A. & Immonen, R. (2014), 'Epilepsy related to traumatic brain injury', *Neurotherapeutics* **1**1(2), 286–296.
Pitkänen, A. & Mcintosh, T. K. (2006), 'Animal models of post-traumatic epilepsy', *Journal of neurotrauma* **2**3(2), 241–261.
Power, A., Burda, Y., Edwards, H., Babuschkin, I. & Misra, V. (2022), 'Grokking: Generalization beyond overfitting on small algorithmic datasets', *arXiv preprint arXiv:2201.02177* .
Reid, A. Y., Bragin, A., Giza, C. C., Staba, R. J. & Engel Jr, J. (2016), 'The progression of electrophysiologic abnormalities during epileptogenesis after experimental traumatic brain injury', *Epilepsia* **5**7(10), 1558–1567.
Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 234–241.
Rosenblatt, F. (1958), 'The perceptron: a probabilistic model for information storage and organization in the brain.', *Psychological review* **6**5(6), 386.
Roy, A. G., Conjeti, S., Navab, N. & Wachinger, C. (2018), 'Quicknat: Segmenting mri neuroanatomy in 20 seconds', *arXiv preprint arXiv:1801.04161* .
Roy, S., Knutsen, A., Korotcov, A., Bosomtwi, A., Dardzinski, B., Butman, J. A. & Pham, D. L. (2018), A deep learning framework for brain extraction in humans and animals with traumatic brain injury, *in* 'Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on', IEEE, pp. 687–691.
Rumple, A., McMurray, M., Johns, J., Lauder, J., Makam, P., Radcliffe, M. & Oguz, I. (2013), '3-dimensional diffusion tensor imaging (dti) atlas of the rat brain', *PLoS One* **8**(7), e67334.
Rundo, L., Han, C., Nagano, Y., Zhang, J., Hataya, R., Militello, C., Tangherloni, A., Nobile, M. S., Ferretti, C., Besozzi, D. *et al.* (2019), 'Use-net: Incorporating squeeze-and-excitation blocks into u-net for prostate zonal segmentation of multi-institutional mri datasets', *Neurocomputing* **3**65, 31–43.
Saenz-Aguirre, A., Zulueta, E., Fernandez-Gamiz, U., Lozano, J. & Lopez-Guede, J. M. (2019), 'Artificial neural network based reinforcement learning for wind turbine yaw control', *Energies* **1**2(3), 436.
Santhakumar, V., Bender, R., Frotscher, M., Ross, S. T., Hollrigel, G. S., Toth, Z. & Soltesz, I. (2000), 'Granule cell hyperexcitability in the early post-traumatic rat dentate gyrus: the 'irritable mossy cell'hypothesis', *The Journal*

of physiology **5**24(1), 117–134.
Sawiak, S., Wood, N., Williams, G., Morton, A. & Carpenter, T. (2009), Spmmouse: A new toolbox for spm in the animal brain, *in* 'ISMRM 17th Scientific Meeting & Exhibition, April', pp. 18–24.
Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., Hirsch, E., Jain, S., Mathern, G. W., Moshé, S. L. *et al.* (2017), 'Ilae classification of the epilepsies: position paper of the ilae commission for classification and terminology', *Epilepsia* **5**8(4), 512–521.
Schwarz, A. J., Danckaert, A., Reese, T., Gozzi, A., Paxinos, G., Watson, C., Merlo-Pich, E. V. & Bifone, A. (2006), 'A stereotaxic mri template set for the rat brain with tissue class distribution maps and co-registered anatomical atlas: application to pharmacological mri', *Neuroimage* **3**2(2), 538–550.
Schweinhardt, P., Fransson, P., Olson, L., Spenger, C. & Andersson, J. L. (2003), 'A template for spatial normalisation of mr images of the rat brain', *Journal of neuroscience methods* **1**29(2), 105–113.
Shultz, S. R., Cardamone, L., Liu, Y. R., Hogan, R. E., Maccotta, L., Wright, D. K., Zheng, P., Koe, A., Gregoire, M.-C., Williams, J. P. *et al.* (2013), 'Can structural or functional changes following traumatic brain injury in the rat predict epileptic outcome?', *Epilepsia* **5**4(7), 1240–1250.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. *et al.* (2017), 'Mastering the game of go without human knowledge', *nature* **5**50(7676), 354–359.
Sled, J. G., Zijdenbos, A. P. & Evans, A. C. (1998), 'A nonparametric method for automatic correction of intensity nonuniformity in mri data', *IEEE transactions on medical imaging* **1**7(1), 87–97.
Smith, S. M. (2002*a*), 'Fast robust automated brain extraction', *Human brain mapping* **1**7(3), 143–155.
Smith, S. M. (2002*b*), 'Fast robust automated brain extraction', *Human brain mapping* **1**7(3), 143–155.
Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. (2017), Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, *in* 'Deep learning in medical image analysis and multimodal learning for clinical decision support', Springer, pp. 240–248.
Swartz, B. E., Houser, C. R., Tomiyasu, U., Walsh, G. O., DeSalles, A., Rich, J. R. & Delgado-Escueta, A. (2006), 'Hippocampal cell loss in posttraumatic human epilepsy', *Epilepsia* **4**7(8), 1373–1382.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015), Going deeper with convolutions, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1–9.
Taha, A. A. & Hanbury, A. (2015), 'Metrics for evaluating 3d medical image

segmentation: analysis, selection, and tool', *BMC medical imaging* **1**5(1), 1–28.

Thirion, J.-P. (1998), 'Image matching as a diffusion process: an analogy with maxwell's demons', *Medical image analysis* **2**(3), 243–260.

Timmins, K. M., van der Schaaf, I. C., Bennink, E., Ruigrok, Y. M., An, X., Baumgartner, M., Bourdon, P., De Feo, R., Di Noto, T., Dubost, F. *et al.* (2021), 'Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge', *Neuroimage* **2**38, 118216.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A. & Gee, J. C. (2010), 'N4itk: improved n3 bias correction', *IEEE transactions on medical imaging* **2**9(6), 1310–1320.

Uberti, M. G., Boska, M. D. & Liu, Y. (2009), 'A semi-automatic image segmentation method for extraction of brain volume from in vivo mouse head magnetic resonance imaging using constraint level sets', *Journal of neuroscience methods* **1**79(2), 338–344.

Vallat, R. (2018), 'Pingouin: statistics in python', *The Journal of Open Source Software* **3**(31), 1026.

Valverde, J. M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R. & Tohka, J. (2021*a*), 'Transfer learning in magnetic resonance brain imaging: A systematic review', *Journal of Imaging* **7**(4), 66.

Valverde, J. M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R. & Tohka, J. (2021*b*), 'Transfer learning in magnetic resonance brain imaging: a systematic review', *Journal of imaging* **7**(4), 66.

Valverde, J. M., Shatillo, A., De Feo, R., Gröhn, O., Sierra, A. & Tohka, J. (2019), Automatic rodent brain mri lesion segmentation with fully convolutional networks, *in* 'International Workshop on Machine Learning in Medical Imaging', Springer, pp. 195–202.

Valverde, J. M., Shatillo, A., De Feo, R., Gröhn, O., Sierra, A. & Tohka, J. (2020), 'Ratlesnetv2: A fully convolutional network for rodent brain lesion segmentation', *Frontiers in neuroscience* **1**4, 1333.

Valverde, J. M., Shatillo, A., De Feo, R. & Tohka, J. (2022), 'Automatic cerebral hemisphere segmentation in rat mri with ischemic lesions via attention-based convolutional neural networks', *Neuroinformatics* pp. 1–14.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E. & Yu, T. (2014), 'scikit-image: image processing in python', *PeerJ* **2**, e453.

Veraart, J., Leergaard, T. B., Antonsen, B. T., Van Hecke, W., Blockx, I., Jeurissen, B., Jiang, Y., Van der Linden, A., Johnson, G. A., Verhoye, M. *et al.* (2011), 'Population-averaged diffusion tensor imaging atlas of the sprague dawley rat brain', *Neuroimage* **5**8(4), 975–983.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. *et al.* (2020), 'Scipy 1.0: fundamental algorithms for scientific computing in python', *Nature methods* **1**7(3), 261–272.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors (2020), 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', *Nature Methods* **1**7, 261–272.

Wachinger, C., Reuter, M. & Klein, T. (2018), 'Deepnat: Deep convolutional neural network for segmenting neuroanatomy', *NeuroImage* **1**70, 434–445.

Wang, L., Zhang, D., Guo, J. & Han, Y. (2020), 'Image anomaly detection using normal data only by latent space resampling', *Applied Sciences* **1**0(23), 8660.

Wang, X., Wang, Y., Zhang, C., Liu, C., Yang, H.-F., Hu, W.-H., Zhang, J.-G. & Zhang, K. (2016), 'Endogenous cannabinoid system alterations and their role in epileptogenesis after brain injury in rat', *Epilepsy research* **1**28, 35–42.

Warfield, S. K., Zou, K. H. & Wells, W. M. (2004), 'Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation', *IEEE transactions on medical imaging* **2**3(7), 903–921.

World Health Organization (2019), *Epilepsy: a public health imperative*, World Health Organization, Geneva.

Xie, L., Qi, Y., Subashi, E., Liao, G., Miller-DeGraff, L., Jetten, A. M. & Johnson, G. A. (2015), '4d mri of polycystic kidneys from rapamycin-treated glis3-deficient mice', *NMR in Biomedicine* **2**8(5), 546–554.

Yang, X., Kwitt, R., Styner, M. & Niethammer, M. (2017), 'Quicksilver: Fast predictive image registration–a deep learning approach', *NeuroImage* **1**58, 378–396.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C. & Gerig, G. (2006), 'User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability', *Neuroimage* **3**1(3), 1116–1128.

Zeiler, M. D. & Fergus, R. (2014), Visualizing and understanding convolutional networks, *in* 'European conference on computer vision', Springer, pp. 818–833.

Zeng, X., Huang, R., Zhong, Y., Sun, D., Han, C., Lin, D., Ni, D. & Wang, Y. (2021), Reciprocal learning for semi-supervised segmentation, *in* 'International Conference on Medical Image Computing and Computer-

Assisted Intervention', Springer, pp. 352–361.
Zhan, K., Shi, J., Wang, H., Xie, Y. & Li, Q. (2017), 'Computational mechanisms of pulse-coupled neural networks: a comprehensive review', *Archives of Computational Methods in Engineering* **2**4(3), 573–588.
Zhang, T. & Suen, C. Y. (1984), 'A fast parallel algorithm for thinning digital patterns', *Communications of the ACM* **2**7(3), 236–239.
Zhang, Y., Brady, M. & Smith, S. (2001), 'Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm', *IEEE transactions on medical imaging* **2**0(1), 45–57.
Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. (2020), 'A comprehensive survey on transfer learning', *Proceedings of the IEEE* p. in press.

RICCARDO DE FEO

Convolutional neural networks are proving to be an essential tool in image analysis. In this work a novel methodological approach is explored for the simultaneous segmentation and skull stripping of rodent brain MRI, and applied in a wide range of datasets of different sizes, in healthy and pathological brains, and finally applied to the task of discriminating between epileptic and non-epileptic rats as a consequence of traumatic brain injury.

UNIVERSITY OF
EASTERN FINLAND

uef.fi

**PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND**
Dissertations in Health Sciences

ISBN 978-952-61-4785-7
ISSN 1798-5706