TUM School of Computation, Information and Technology
Technische Universität München

The Hitchhiker's Guide to Machine Learning for Biomedical Image Analysis

Florian Kofler

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:
Prof. Dr. Julia Schnabel

Prüfer*innen der **Dissertation:**
1. Prof. Dr. Björn Menze
2. Prof. Dr. Roland Wiest

Die Dissertation wurde am 27.06.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 14.02.2023 angenommen.

ABSTRACT

The Hitchiker's Guide to Machine Learning for Biomedical Image Analysis is a dissertation for tackling biomedical image analysis problems with machine learning. It covers the *whole iterative workflow* from *dataset curation*, *network training*, *evaluation*, and *interpretation* of results to refining the model in three parts.

The first part covers typical challenges in hyperparameter tuning. It features recommendations for tuning important parameters, such as *cropping*, *patch shape*, *patch size*, *normalization*, and *data preprocessing* with a special focus on Computed Tomography (CT) and Magnetic Resonance (MR) imaging data.

The second part illustrates methods to *operationalize model performance*. It covers an overview of *established metrics* and remarks regarding their interpretation. Therefore, the concept of *peak ground truth* is introduced. This is the point at which the model starts to mimic the errors in the (human) annotation, and an increase in similarity metrics results in lower model performance. To provide an alternative to *established metrics*, software architecture for conducting *Pyschophysics* experiments with clinicians is introduced.

The third part comprises model performance optimization strategies. Besides *ensembling*, it lays out concepts for *dataset curation* with an exemplary brain tumor database. Further, guidelines for making informed decisions regarding *network architecture* and *loss function* are provided.

Peer-reviewed academic contributions and an outlook complete the dissertation.

ZUSAMMENFASSUNG

The Hitchiker's Guide to Machine Learning for Biomedical Image Analysis ist eine Dissertation zur Bearbeitung biomedizinischer Bildanalyseprobleme mittels maschinellem Lernen. In drei Teilen wird der gesamte *iterative Arbeitsablauf* vom *Kuratieren von Datensätzen*, *Netzwerktraining*, *Evaluation und Interpretation von Ergebnissen*, bis hin zum *Verfeinern von Modellen* behandelt.

Der erste Teil beschreibt typische Herausforderungen beim *Feinjustieren von Hyperparametern*. Er beinhaltet Empfehlungen zum Finden geeigneter Einstellungen für wichtige Parameter wie *Cropping*, *Patch Shape*, *Patch Size* und *Normalisierung*, sowie *Datenpräprozessierung* mit einem speziellen Focus auf CT und MR Bilddaten.

Der zweite zweite Teil beschreibt Methoden um die Leistungsfähigkeit von Modellen zu operationalisieren. Er inkludiert eine Übersicht über etablierte Metriken und Hinweise zu deren Interpretation. Hierzu wird das Konzept von *peak ground truth* eingeführt. Dies ist der Punkt an dem das Modell beginnt die Fehler in der Annotierung zu reproduzieren und somit die Steigerung einer Ähnlichkeitsmetrik zu einer Verschlechterung der Leistungsfähigkeit des Modells führt. Als Alternative zu *etablierten Metriken* wird eine Softwarearchitektur zur Durchführung von *Psychophysics*-Experimenten mit Ärzten vorgestellt.

Im Dritten Teil finden sich Strategien zur Optmierung der Leistungsfähigkeit von modellen. Neben *Ensembling* wird ein Konzept zur *Kuratierung von Datensätzen*, am Beispiel einer Gehirntumordatenbank erläutert. Des Weiteren beinhaltet er Richtlinien um informierte Entscheidungen bezüglich *Netzwerkarchitektur* und *Lossfunktion* zu treffen.

Akademische Veröffentlichen, sowie ein Ausblick komplettieren die Dissertation.

ACKNOWLEDGMENTS

I am deeply thankful to everyone who contributed to the success of this thesis and supported me over the last few years. This includes my advisors, colleagues, academic- and industry collaborators, examiners, working-, bachelor- and master's students, friends, medical - and administrative staff who keep the world turning.

Looking at the publications' long author lists and acknowledgments, it becomes clear to me, that this thesis is not the work of a sole individual, but the result of a great team effort. Without the support of these largely overlapping groups, it would not have been possible to achieve my doctorate.

I especially want to express my sincere gratitude to Prof. Dr. Bjoern Menze for providing guidance and inspiration while providing me the freedom to explore various endeavors.

Prof. Dr. Claus Zimmer and Dr. Benedikt Wiestler for answering all my medical questions and providing me with general support whenever needed.

Prof. Dr. Jan Kirschke in his role as project leader of the *RadioTumor* project, especially for promoting our *Pyschophysics* experiments.

A huge thank you goes to all my (post-) doctoral fellows for all the discussions and experiences we could share together.

Further, I acknowledge all my funding sources, especially, the SFB 824, subproject B12 and Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81.

Finally, I am also thankful for everyone who was patient with me in inspiring and sometimes demanding discussions. This includes especially my family and friends who I want to thank for their unwavering support.

PUBLICATIONS

This cumulative dissertation is based on the following peer-reviewed publications:

[1] Florian Kofler, Christoph Berger, Diana Waldmannstetter, Jana Lipkova, Ivan Ezhov, Giles Tetteh, Jan Kirschke, Claus Zimmer, Benedikt Wiestler, and Bjoern H Menze. "BraTS Toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice." In: *Frontiers in neuroscience* (2020), p. 125.

[2] Florian Kofler, Ivan Ezhov, Lucas Fidon, Carolin M Pirkl, Johannes C Paetzold, Egon Burian, Sarthak Pati, Malek El Husseini, Fernando Navarro, Suprosanna Shit, et al. "Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles." In: *Frontiers in Neuroscience* 15 (2021).

[3] Florian Kofler, Johannes C Paetzold, Ivan Ezhov, Suprosanna Shit, Daniel Krahulec, Jan S Kirschke, Claus Zimmer, Benedikt Wiestler, and Bjoern H Menze. "A baseline for predicting glioblastoma patient survival time with classical statistical models and primitive features ignoring image information." In: *International MICCAI Brainlesion Workshop*. Springer. 2019, pp. 254–261.

The following manuscripts are not subject to evaluation but complement the content of the above:

[1] Florian Kofler, Ivan Ezhov, Fabian Isensee, Fabian Balsiger, Christoph Berger, Maximilian Koerner, Johannes Paetzold, Hongwei Li, Suprosanna Shit, Richard McKinley, et al. "Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient." In: *arXiv preprint arXiv:2103.06205* (2021).

[2] Florian Kofler, Suprosanna Shit, Ivan Ezhov, Lucas Fidon, Izabela Horvath, Rami Al-Maskari, Hongwei Bran Li, Harsharan Bhatia, Timo Loehr, Marie Piraud, et al. "blob loss: instance imbalance aware loss functions for semantic segmentation." In: *International Conference on Information Processing in Medical Imaging*. Springer Nature Switzerland Cham. 2023, pp. 755–767.

[3] Florian Kofler et al. *Deep Quality Estimation: Creating Surrogate Models for Human Quality Ratings*. 2022. DOI: 10.48550/ARXIV.2205.10355. URL: https://arxiv.org/abs/2205.10355.

[4] Florian Kofler et al. *Approaching Peak Ground Truth*. 2023. arXiv: 2301.00243 [cs.LG].

FOREWORD

The Hitchhiker's Guide to Machine Learning for Biomedical Image Analysis? While there are already hitchhiker's guides to broader topics such as the *galaxy*, this hitchhiker's guide focuses on a slightly narrower topic, namely machine learning for biomedical image analysis.

It is meant to be a compact and accessible guide for tackling biomedical image analysis problems with a focus on, but not limited to, segmentation projects. I summarize some of the learnings, pitfalls, and mitigation strategies for model training, evaluation, and interpretation of results that I discovered during my doctorate at the Technical University of Munich (TUM). The guide is written with hitchhikers in mind, meaning individuals coming from outside bio-informatics entering the field (like myself at the start of my doctorate). However, I hope that it will be appealing to wider audiences, be it hitchhikers or not, including you.

ACRONYMS

BTK	BraTS Toolkit
CPU	Central Processing Unit
CT	Computed Tomography
CNN	Convolutional Neural Network
DSC	Sørensen–Dice coefficient
GUI	Graphical User Interface
GPU	Graphical Processing Unit
HU	Hounsfield Units
ITI	Intra-Trial Tnterval
MAE	Mean Absolute Error
MSE	Mean Square Error
ML	Machine Learning
MR	Magnetic Resonance
MS	Multiple sclerosis
RF	Radio Frequency
TUM	Technical University of Munich

CONTENTS

Part I

CHALLENGES IN HYPERPARAMETER TUNING

When it comes to hyperparameter tuning each imaging modality features its own challenges in biomedical image analysis. Besides modality-specific aspects, this chapter features some general remarks.

1 GENERAL CHALLENGES AND MITIGATION STRATEGIES

Especially for beginners, the process of identifying *good* hyperparameters for model training that provide the best, potentially random, fit on the validation - and test sets often feels completely random. However, there are a few heuristics that often produce satisfying results. Before diving into image-modality-specific facets of hyperparameter tuning, some general strategies for selecting important hyperparameters should be discussed.

Cropping: Modern Graphical Processing Units (GPUs) feature big amounts of memory, enabling huge crop sizes. Sometimes it is possible to skip cropping entirely and train on whole image volumes. Even when cropping is not necessary to fit data into memory, it should be considered a tuneable hyperparameter, and it might make sense to experiment with smaller crop sizes. Also, multiple crops from the same volume should be considered. Especially when dealing with a multi-fragmented segmentation problem with a high variance within the connected components, such as segmentation of Multiple sclerosis (MS) lesions, cropping becomes a powerful augmentation strategy. When computing the loss on the whole image volume, larger-sized instances dominate the loss calculation. Cropping will lead to patches with only small lesions, forcing the Convolutional Neural Network (CNN) to pick up these small lesions to achieve a good loss, see Chapter 8 and Figure 1.1.

Patch shape: Broader contextual information, such as anatomy, only becomes visible in larger-sized patches. For instance, zoomed-in tissue from the left and right lung is indistinguishable. Encoding such contextual information can aid in improving performance. Cubes are probably the most intuitive way to cut images into patches; however, more contextual information can be captured by employing flatter cuboid shapes. If the image resolution differs between planes, it makes sense to align the long side of the cuboid with the high-resolution plane. The bodies of vertebrates, such as humans, feature several symmetries. Human reading of imaging data often involves comparing sides, e.g., for detecting MS lesions intensities are compared between the left and right brain hemispheres. Therefore, images often feature higher resolutions in the axial plane. Consequently, to enable the CNN to pick up this information, the axial plane should resemble the long side of the patch shape for cropping, as illustrated in Figure 1.1 for the example of in MS lesion segmentation.

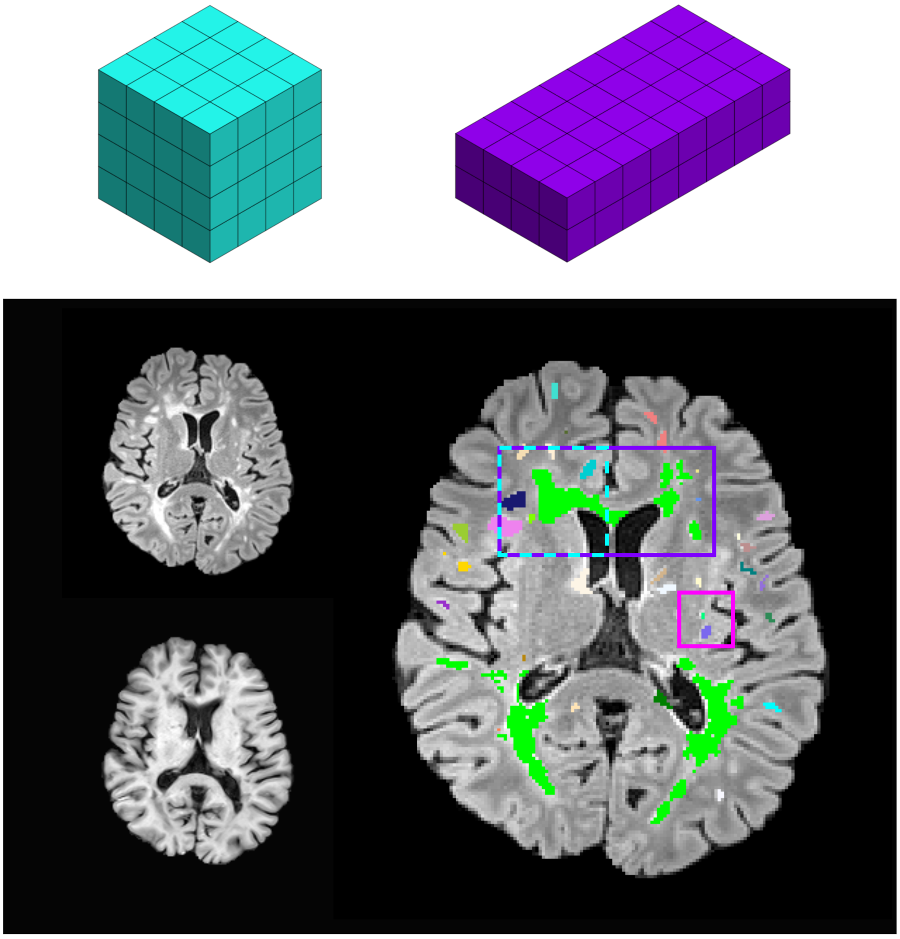


Figure 1.1: Patch shape and size. *Top:* The teal 4*x*4*x*4 cube and purple 8*x*4*x*2 cuboid, pictured on the top, result in the same volume. *Bottom Left:* Raw FLAIR (top) T1 (bottom) image. *Bottom Right:* Magnified FLAIR image with human annotation of MS lesions overlayed on the FLAIR image. Corresponding crops to the teal cube and purple cuboid patch are illustrated in the respective colors. The purple cuboid allows encoding more contextual information in the *axial* plane while maintaining the same memory footprint. Using smaller crops, such as the pink one, incentivizes the network to also pickup smaller lesions. There is a target conflict in encoding context and learning about such micro-structures. Employing *blob loss*, see Chapter 8, allows maintaining both targets.

Balancing patch vs. batch size: GPU memory is a precious commodity. Hence, *patch* and *batch* sizes need to be balanced carefully. Training with smaller-sized crops of the whole image enables increasing *batch size*, thus capturing more variance in the data. The greater the variance between exams, the more beneficial effects can be expected from increasing *batch* size. Using a *patch size* where humans can solve the image analysis problem at hand is often a good starting point for experimentation.

Normalization / thresholding: Normalization and windowing can greatly affect model performance. Therefore, it makes sense to conduct dedicated experiments to explore the optimal method for the problem at hand. In some rare cases applying normalization strategies might turn out counterproductive; thus, training a network using raw input values is also worth a shot. [1]

Scarce training data: Traditional Machine Learning (ML) splits data into training, validation, and a test set. For instance, researchers often use 70 percent of the data for training and distribute the remaining 30 percent equally across validation and test set. During training, the model is continuously evaluated on the validation set. Finally, the best-performing model on the validation set is evaluated on the test set. In biomedical imaging, one is frequently confronted with the scarce availability of training data due to the high data acquisition cost. Additionally, annotation often requires, hard to come by, domain experts, and to make it worse, images feature high variance. The combination of these effects regularly renders the formation of representative subsets impossible. Consequently, it makes sense to experiment with diverging from the traditional ML paradigm sketched above. In practice, skipping model selection on the validation set favoring training with additional data often outperforms conventional approaches. Another field-proven mitigation strategy is ensembling. There are many ways to implement ensembling. For instance, Isensee et al. [16] suggest splitting the training data into five folds and building an ensemble by training on the resulting subsets, see Chapter 8.

Multimodal imaging data: Many biomedical imaging problems, such as glioma segmentation, require multimodal imaging to capture multifaceted aspects of the problem at hand. Multimodal imaging comprises multiple challenges. As the imaging modalities are usually recorded sequentially, they reflect changes within the subject, such as movements or decay of contrast agent. Furthermore, the imaging resolution might vary between modalities. It is advisable to co-register the images and harmonize the resolution to tackle these challenges. In this process, it is often advisable to resample to an isotropic resolution, enabling more straightforward volume calculations. Further, the use of an atlas should be evaluated if available.

[1] For instance, skipping normalization led to a double-digit improvement in Sørensen–Dice coefficient (DSC) for *c-Fos* data [18].

Registration tools: Over the years, multiple tools were established for registration. NiftyReg *NiftyReg* provides a quick and easy to use baseline [36]. While CNN-based approaches, such as as *VoxelMorph* [6], promise very quick registrations in a *"quasi-ballistic"* manner, Advanced Normalization Tools (ANTs) [1] and greedy [46] are worth considering when more fine-grained control is required.

2 MAGNETIC RESONANCE IMAGING TECHNICALITIES

Basic Functionality: Magnetic Resonance (MR) imaging represents a non-invasive method enabling sophisticated insights into biological organisms. Therefore, MR scanners generate a strong and uniform magnetic field. Protons then align with this so-called *Bo* field and spin with its frequency along their own axis. Next, a Radio Frequency (RF) pulse orthogonal to the *Bo* field is generated. Its frequency matches the frequency of the spinning protons to flip them out of orientation with the base magnetic field. This enables the generation of an image by recording the time the protons need for realignment with the *Bo* magnetic field with sophisticated coils. Broadhouse [9] provide a gentle introduction into the complex topic of MR physics.

Normalization / thresholding: In conventional MR imaging, the absolute intensities of voxels are not interpretable. Consequently, voxel intensities can only be interpreted in relation to each other. As the same subject in a different scanner, even of the same kind, might generate entirely different intensity values applying normalization is advisable. Only experimentation can reveal whether conventional *minimum / maximum normalization* or the more outlier-resistant *percentile-based normalization* as used by Isensee et al. [16] is more suited to the problem at hand.

Patch and batch size: *Patch* and *batch* size need to be carefully balanced. MR images often feature higher axial resolution, hence training with cuboid patches to capture anatomic contextual information is advisable.

Image acquisition artifacts: MR images are prone to image acquisition artifacts such as spikes, ghosting, and magnetic field inhomogeneities. While N4 bias field correction methods can address field inhomogeneities [45], libraries such as *TorchIo* can simulate image acquisition artifacts to achieve greater robustness [33].

Multimodal data: MR imaging problems often combine multiple imaging protocols to aggregate information, while each imaging modality is a potential source of error. The details of image protocols tend to vary between scanners, making it difficult to curate datasets from multiple scanners. As an MR exam usually requires several minutes and sometimes exceeds one hour, state changes within the subject and movement need to be corrected, also see Chapter 1. Figure 2.1 illustrates a full pipeline to tackle these problems for a multimodal brain tumor imaging protocol.

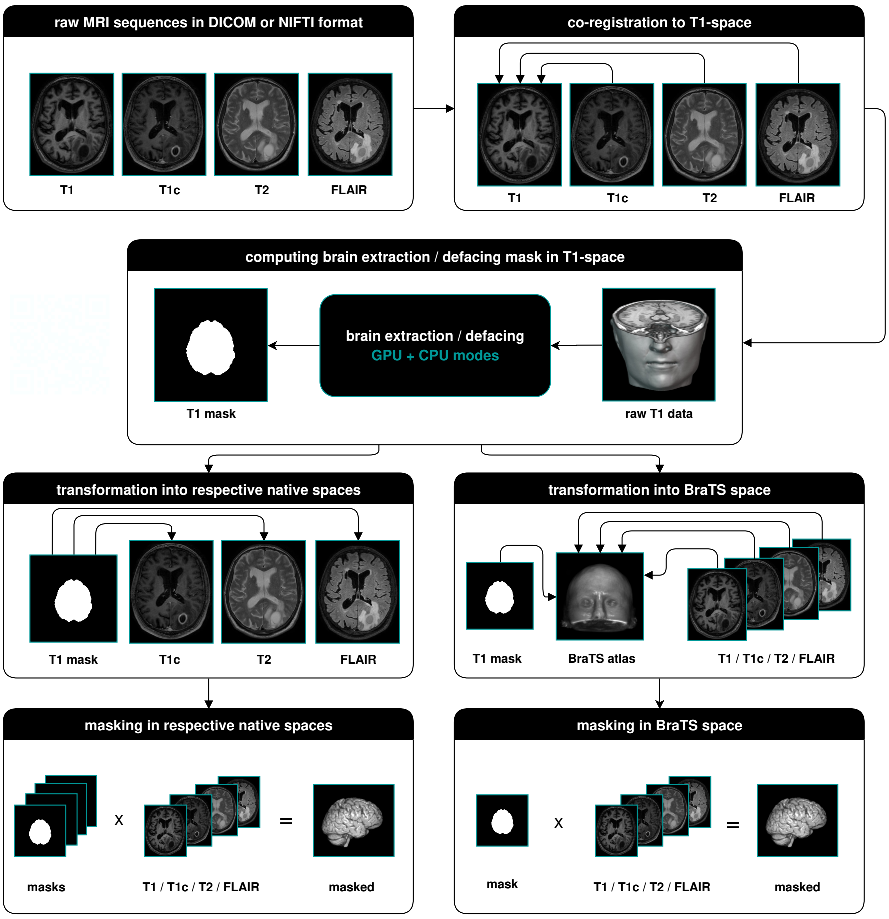

raw MRI sequences in DICOM or NIFTI format
T1
T1c
T2
FLAIR
co-registration to T1-space
T1
T1c
T2
FLAIR
computing brain extraction / defacing mask in T1-space
brain extraction / defacing
GPU + CPU modes
T1 mask
raw T1 data
transformation into respective native spaces
T1 mask
T1c
T2
FLAIR
transformation into BraTS space
T1 mask
BraTS atlas
T1 / T1c / T2 / FLAIR
masking in respective native spaces
masks
x
T1 / T1c / T2 / FLAIR
=
masked
masking in BraTS space
mask
x
T1 / T1c / T2 / FLAIR
=
masked

Figure 2.1: Preprocessing pipeline for multimodal brain tumor segmentation based on T1, T1c, T2, and FLAIR images. The images are first co-registered to the T1 image, which provides the best anatomic information. Then they are skullstripped in native T1 space and registered to the BraTS atlas. The skullstripping masks are then morphed into the BraTS and native T1 space and multiplied with the images. This way, the images are available in native T1 space and BraTS space for further downstream processing, such as CNN training. The figure appears in [19].

3 COMPUTED TOMOGRAPHY TECHNICALITIES

Basic Functionality: Wilhelm Röntgen discovered the X-ray spectrum of electromagnetic waves. In 1895 he developed X-ray imaging and recorded the famous image of his wife's hand by recording the radiation passing through her body with a photographic plate. One year later, X-ray found its' way into the medical domain with the first surgery performed in England by John Hall-Edwards. Around the same time, the collective consciousness of the hazards of radiation associated with X-rays began to develop. Several decades later, in the 1970s, the first commercial CT machines became available. These machines rotated the radiation source and receiver around the subject. Then image reconstruction algorithms create three-dimensional representations from the acquired data [38]. Even though modern designs are more sophisticated, the core idea remains the same.

Comparison to MR: Both MR - and CT imaging provide non-invasive means to derive insights regarding the internals of subjects. Both techniques can be combined with contrast agents to obtain further knowledge. Compared to MR images, modern CT images feature higher temporal and spatial resolution. However, CT imaging is considered harmful due to the hazards associated with radiation. Therefore, its usage should be carefully evaluated. Nevertheless, it can be applied when MR use is contraindicated due to ferromagnetic material within the subject of interest.

Normalization / thresholding: CT images are quantitative, meaning the values of the observed image intensities are measured on the Hounsfield scale, and certain Hounsfield Units (HU) can be directly attributed to specific substances [39]. Therefore, conventional normalization is often counterproductive as it removes this vital information. Instead, it is advisable to experiment with windowing. Windows specific to the tissue of interest are good starting points for experimentation. For example, for segmentation of lung lesions, a lung window within a certain range of HU makes sense [13], compare fig. 3.1.

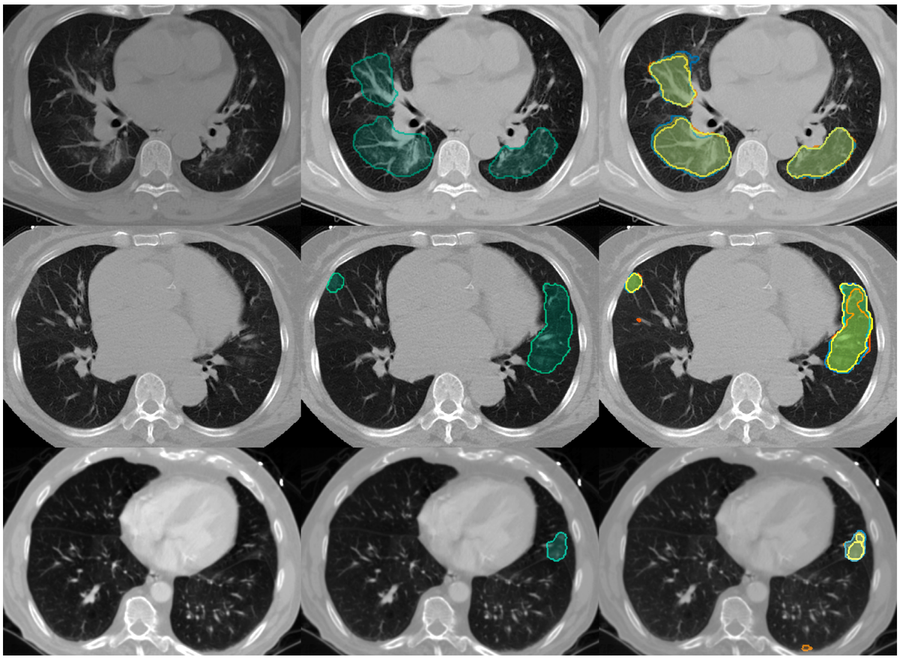


Figure 3.1: Covid-19 lung lesions in CT imaging. The images is thresholded with a lung window from *-1000* to *500* HUs for better visibility of lung lesions. In the right column, multiple segmentation candidates are represented in different colors. The middle column illustrates a fusion of these. that successfully removed the false positive lesions. The figure appears in [20].

Patch and batch size: One of the biggest challenges in CT segmentation are the great regions of interest combined with the high spatial resolution. With single exams regularly featuring hundreds of megabytes, this combination results in high data loads, which need to be fed to the neural networks. For instance, spine and multi-organ segmentation pipelines both require analysis of the whole thorax. Furthermore, (larger) anatomical context can be relevant to distinguish between vertebrae and organs, respectively. Due to the above-mentioned quantitative imaging intensities, differences between images are usually low compared to other image modalities. Thus, bigger patch sizes often help to improve performance and should be prioritized over batch size. GPUs with high memory capacity are advantageous for these processing requirements. Post-processing pipelines incorporating top-down (anatomical) knowledge can mitigate this to a certain extent.

Part II

OPERATIONALIZING MODEL PERFORMANCE

This chapter deals with methods to operationalize model performance. Even though the chapter focuses on segmentation tasks, many of the following concepts easily translate to other machine learning problems.

4 ESTABLISHED SIMILARITY METRICS IN ML

Computing similarity metrics between model output and ground truth annotations is the standard way to operationalize model performance. There are numerous established similarity metrics for measuring segmentation quality. Most metrics are based on volumetry, like the prominent DSC. The DSC computes the relation of correctly and incorrectly classified voxels and is defined in eq. (4.1).

$$DSC = \frac{2TP}{2TP + FP + FN}. \tag{4.1}$$

Where *TP* represents true positive voxels, *FP* false positive voxels, and *FN* false negative voxels.

While most volumetry-based metrics, such as *Jaccard coefficient* can be derived from this, distance-based metrics, such as Hausdorff distance and surface Dice similarity coefficient, [32] mark some of the few exceptions.

Taha and Hanbury [44], as well as Reinke et al. [34], review current metrics and nicely illustrate the particularities that come with their usage. For training a CNN, a differentiable loss function is required. However, for some metrics, differentiable implementations do not exist. Ma et al. [28] provide an overview of available loss implementations for segmentation tasks.

5 INTERPRETATION OF SIMILARITY METRICS

In the (biomedical) ML community, it is common practice to market a technical innovation by claiming an improvement in segmentation quality by demonstrating a small improvement in similarity metrics such as DSC. Many segmentation challenges, such as BraTS, LiTS, KiTS, etc., emerged that decorate winners based on such tiny improvements. Newer research casts doubt on this procedure [20, 29, 34]. However, it remains an open research question whether these improvements translate to real-world benefits for the application of ML in clinical workflows. Over the years challenge organizers developed consciousness of this problem and started combining multiple metrics to get a broader understanding of the models' performance [29].

Nevertheless, for the interpretation of similarity metrics, it is necessary to reflect upon the nature of the employed annotations. For biomedical problems, often human annotations provide the *gold standard* to measure model performance. It is important to note the limitations of human annotators. Humans and computers both learn systematic errors. However, unlike humans, computers do not suffer from sleepiness, laziness, and other distractions. Thus, computers are not prone to random errors. This, combined with a computer's unlimited diligence, can lead to computers outperforming humans, as reported in [22]. Therefore, human annotations should be interpreted as such and cannot (and should not) be considered as *ground truth* in the sense of classical ML. This has wide-reaching implications for CNN training, as illustrated in Figure 5.1. [1].

[1] Meanwhile, the *Peak Ground Truth* concept introduced here has been developed further and is published in the proceedings of the *International Symposium on Biomedical Imaging (ISBI)* [25]

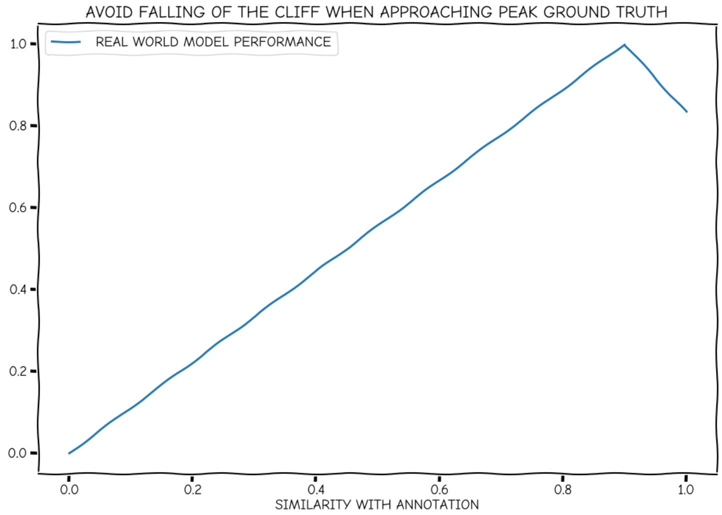


Figure 5.1: Implications for CNN training. When employing a similarity metric in the loss function, the network is incentivized to maximize similarity with potentially erroneous human annotations. Consequently, increasing similarity with such annotations only corresponds to increased *real world model performance* until a certain point, nicknamed *Peak Ground Truth*. Increasing similarity beyond *Peak Ground Truth* leads to *falling off the cliff* and translates to worse *real world model performance*. In this exemplary sketch *Peak Ground Truth* is located around .9 on the x-axis. Its real location and shape depend on the underlying annotation quality. Researchers are well-advised to familiarize themselves with the label quality to get a feeling, until which point it makes sense to hunt for improvements in similarity metrics.

6 SEGMENTATION QUALITY ACCORDING TO CLINICIANS

As explained in Chapter 4, improving upon similarity metrics does not necessarily translate to *real world performance improvements*. So how can *real world performance improvements* be achieved? A key step towards performance optimization is setting up performance monitoring. Doctors are ultimately responsible and liable for treatment decisions in current clinical practice. Hence, the clinicians' quality estimate should be the gold standard to measure, and any ML solution to a medical problem must gain the clinicians' trust.

6.1 PSYCHOPHYSICS APPROACHES

To get an understanding of clinicians' response to model outputs, it makes sense to dive into *Psychophysics*. *Psychophysics* systematically and quantitatively analyzes human behavior in interaction with stimulus material. Typically psychophysical experiments take place in dark and sound-proof experimental cabins shielding the participants from disturbing external stimuli. The stimuli are presented with the help of Psychtoolbox Psychtoolbox [8] or similar libraries in a Matlab environment.

In contrast, collecting the data in the clinicians' natural habitat makes sense to maximize external validity. Here, this means collecting the data directly at the diagnostic workstations. As installing additional software on these machines embedded in the clinical network is problematic, a web browser-based solution is preferable. In recent years multiple frameworks established themselves to display stimulus material in a web browser, such as PsyToolkit [42], jsPsych [10], lab.js and OpenSesame [30]. These can be combined with various backends like JATOS [26] or commercial solutions such as GORILLA to record the data and organize the experiments.

6.2 PLATFORM FOR PSYCHOPHYSICS EXPERIMENTS

To comply with the project-specific requirements, such as data privacy regulation regarding clinician and patient data, a platform for running *Psychophysics* experiments is developed. Figure 6.1 visualizes its software architecture.

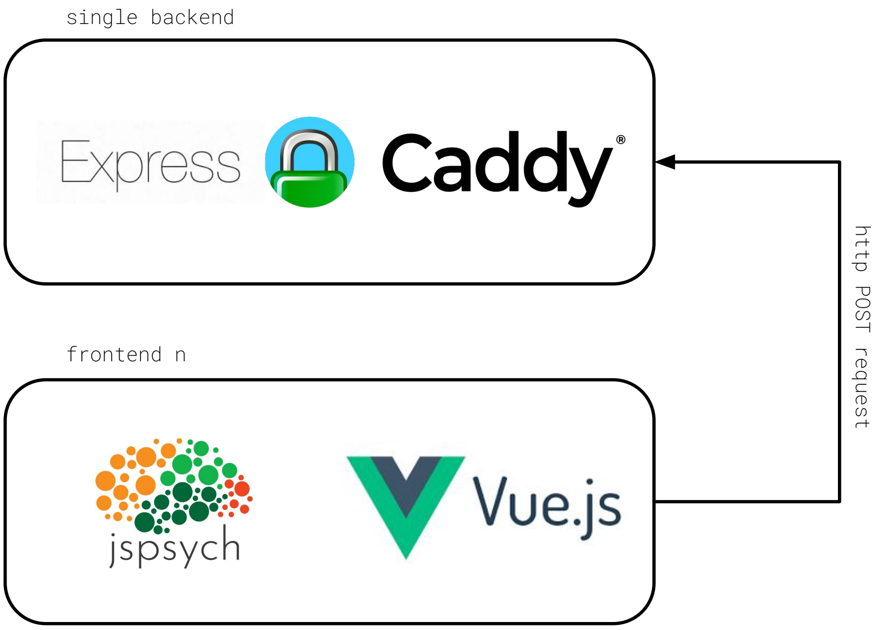


Figure 6.1: Software architecture for the psychophysics experiments. The stimulus material is presented on the frontend via jsPsych [10] embedded into a Vue.js web app. Vue.js is selected over other frontend frameworks as it features a low memory footprint and is heavily optimized for speed. This promises to record reaction times most accurately. Incorporating jsPsych [10] allows to easily follow best practices in experimental Psychology. For instance, trial sequence, trial - and Intra-Trial Tnterval (ITI) duration can be randomized using built-in functionality. The clinicians' responses are sent via *http* to a Express.js backend, that is protected by Caddy acting as a reverse proxy. This way, a single backend server is able to serve multiple studies running at the same time. For flexibility, the deployment happens in containerized fashion via *docker compose*.

Using the aforementioned software architecture comprises several advantages. First, it allows hosting on-premise to comply with data protection regulations. Second, the clinicians can participate in the experiments directly from their acquainted diagnostic workstations and do not have to adjust to a new environment. Third, it enables detailed reaction time analysis, as it is typically done in *Psychophysics* experiments, see an example illustrated in Figure 6.2.

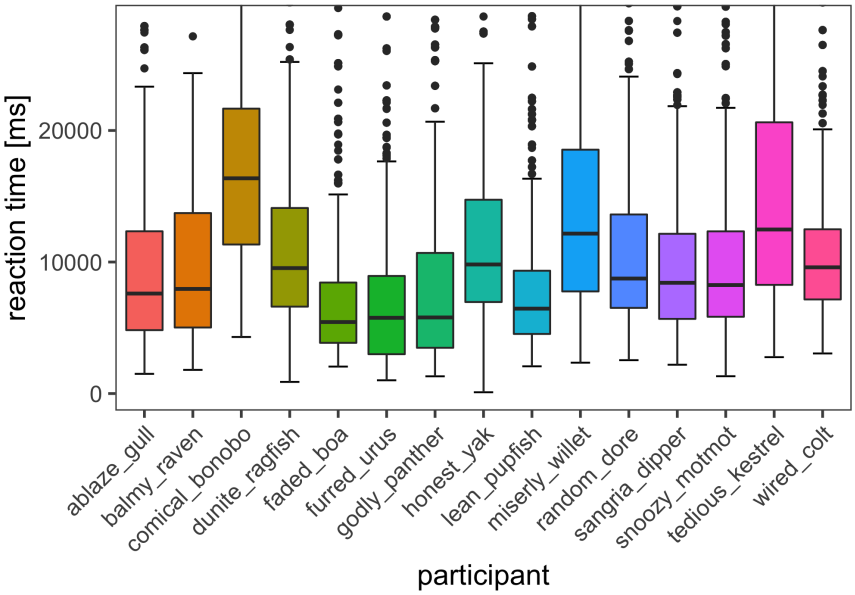


Figure 6.2: Exemplary comparison of reaction times between participants of a glioma segmentation rating quality experiment from [22]. Participants' reaction times vary heavily, with some participants taking more than double the time of other participants to respond.

6.3 ACADEMIC IMPACT

The platform allowed the collection of human expert ratings, complimenting established metrics, for several research projects. All experiments followed the same scheme, illustrated in Figure 6.3.

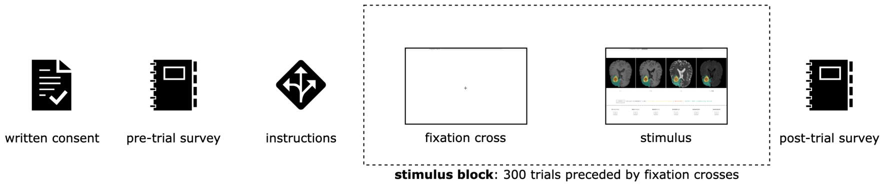


Figure 6.3: Chronological sequence of the experiments from left to right. Initially, the participants are instructed to conduct the evaluations in a suitable environment for reading medical exams. To begin the experiments start with declaring consent. After this, participants answer a survey with questions regarding their age, gender, and various items to measure their expertise. Subsequently, the stimulus trials are presented in random sequence to account for order effects. Following the assignment of a rating, the experiment automatically progresses to the presentation of the next trial. At the end of the experiment, participants have the opportunity to provide feedback during a post-trial survey. This figure also appeared in [22].

A selection of these studies is already published:

[1] Florian Kofler, Ivan Ezhov, Fabian Isensee, Fabian Balsiger, Christoph Berger, Maximilian Koerner, Johannes Paetzold, Hongwei Li, Suprosanna Shit, Richard McKinley, et al. "Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient." In: *arXiv preprint arXiv:2103.06205* (2021).

[2] Florian Kofler et al. *Deep Quality Estimation: Creating Surrogate Models for Human Quality Ratings*. 2022. DOI: 10.48550/ARXIV.2205.10355. URL: https://arxiv.org/abs/2205.10355.

[3] Hongwei Li, Johannes C Paetzold, Anjany Sekuboyina, Florian Kofler, Jianguo Zhang, Jan S Kirschke, Benedikt Wiestler, and Bjoern Menze. "DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 795–803.

[4] Maximilian Möller, Matthias Kohl, Stefan Braunewell, Florian Kofler, Benedikt Wiestler, Jan S Kirschke, Björn H Menze, and Marie Piraud. "Reliable Saliency Maps for Weakly-Supervised Localization of Disease Patterns." In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 2020, pp. 63–72.

Part III

OPTIMIZATION STRATEGIES

Regardless of the machine learning problem at hand, a few optimization strategies have proven successful over the test of time. In this chapter, we take a closer look at three established strategies and how they can be implemented in specific examples.

7 DATASET CURATION

Dataset curation is an essential strategy to optimize model performance. While even seemingly fundamental changes to cornerstones of CNN training such as replacing network architecture or loss function often have negligible effect improving the dataset has an almost 100 percent success guarantee. The first step towards improvement is to start measuring (accurately). Thus obtaining a decent test set should be the number one priority; only then will one be able to evaluate further measures correctly. In the following pages, we want to explore a practical example of how a dataset curation strategy could be implemented considering a complex set of requirements.

7.1 PRACTICAL EXAMPLE: BRAIN TUMOR DATABASE

We set out to explore how an international and multi-institutional brain tumor database could be formed. The dataset is supposed to unite glioma data in MR imaging with demographics and clinical metadata such as IDH methylation status [48].

7.1.1 *Specific requirements*

For many institutions, patient data falls under strict data protection regulations. This leads to several downstream implications. First of all, many institutions require patient data to be anonymized or pseudonymized on-premise. As the local workforce often lacks programming knowledge, a Graphical User Interface (GUI) is desirable. Hospital machines often have no internet access, feature a variety of outdated operating systems, and lack GPU computing capabilities. Consequently, a cross-platform solution with an elegant update solution and an option for Central Processing Unit (CPU) based processing is mandatory.

7.1.2 *Proposed Solution*

We conceptualize a brain tumor database called *BraTum DB*. Figure 7.1 illustrates the branding of the database.

Figure 7.1: Branding of the brain tumor database. The name combines the acronym *TUM* for the *Technical University of Munich* with the first syllables of the words *brain tumor*. Additionally, the two letters *DB* are short for *database*. For the logo, we choose an elephant as elephants are generally attributed as peaceful animals with good long-term memory.

To harmonize the data for import into the database, we design a preprocessing pipeline codenamed *Project Elephant*. The pipeline incorporates a front- and backend. Physicians can trigger the preprocessing on the backend using a GUI that also serves for defining pseudonyms and entering metadata.

The preprocessing pipeline outputs skull-stripped as well as defaced images. These are registered to *BraTS* and *T1 native-space*. Further, a JSON file is created that contains the metadata and overview images in *png* format so the data can be quickly inspected without dedicated software. The co-registrations are implemented via ANTs [1]. Users can conduct brain extraction with HD-BET [17] on GPU or CPU. Moreover, Robex [15] is available as a fallback for a smaller computational footprint.

The exams are assigned a pseudonym on-site using a lookup table to render reconstruction of the patient data impossible. For harmonization purposes, each exam is then assigned a second easy-to-remember pseudonym consisting of adjectives and nouns when ingested into the database. [1]

Figure 7.2 visualizes the software architecture of the preprocessor.

[1] The same naming scheme is utilized in Figure 6.2.

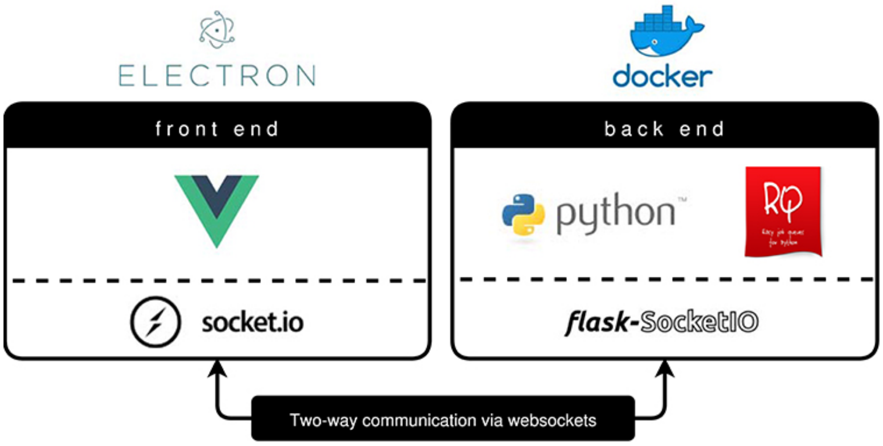


Figure 7.2: Software architecture for *Project Elephant*. To ensure cross-platform capabilities the frontend is written in Electron JS, while the Python backend runs inside a docker. Because of the long computation times, two-way communication between front- and backend is implemented via WebSocket(s). Moreover, computation jobs are queued using Python RQ, and workers are automatically spawned according to the host system's capabilities. The figure is taken from the BraTS Toolkit manuscript [19].

7.1.3 *Automated medical reporting as incentive*

Additionally, we design a tool codenamed *Kraken* to provide physicians with an incentive to contribute data to the brain tumor database. Therefore, we create medical reports in a fully-automatic fashion. The reports comprise an automatic tumor segmentation obtained from BraTS Toolkit Kofler et al. [19] in NIfTI format, as well as a *PDF* file including visualization of the tumor segmentation and volumetry calculations. Once the report is generated, it is shipped to the user via an email address provided during the upload. Figure 7.3 illustrates branding and distributed software architecture of the *Kraken*.

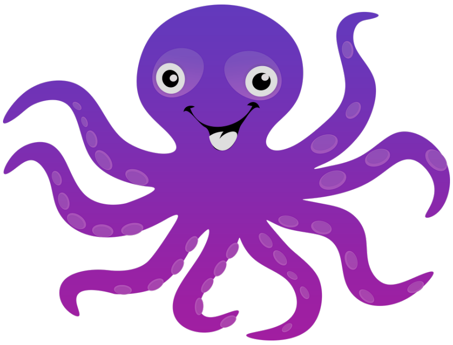


Figure 7.3: Software architecture and branding. The *Kraken* can be deployed using containerization software such as *docker compose* or *Kubernetes*. The backend service listens to *http* requests via aiohttp deployed behind Caddy acting as a reverse proxy for security. Resumable uploads are integrated via Uppy. Similar to *Project Elephant*, the computing jobs are put on a Python Redis Queue, however, this time to enable distribution of the jobs to multiple machines. The reports themselves can be created using web tech and are then rendered to *PDF* format using a headless *Chromium* browser running inside a container. The software is written agnostic to the glioma use case, so it can be easily adapted to other use cases, e.g., during the Covid-19 pandemic, it was used to create lung lesion reports based on CT images. We chose the codename *Kraken* for this project as octopi are regarded as highly intelligent creatures. Further, their tentacles can work independently, representing distributed computing capabilities. In our case, the *Kraken's* tentacles reach out to other institutions to collect data and generate reports in a parallel fashion. As krakens are often perceived as threatening, we try to mitigate this with a friendly appearance.

7.2 ACADEMIC IMPACT

As *BraTum DB* could not be realized due to organizational issues, *Project Elephant*, meaning the preprocessing pipeline was published as a preprocessing module for BraTS Toolkit (BTK) [19]. BTK is a holistic tool for brain tumor segmentation consisting of three modules. Besides the aforementioned preprocessing tool, it comprises a segmentation module to generate glioma segmentations with algorithms collected within the scope of the BraTS challenge [2–5, 31]. Furthermore, the *fusionator* module is included to combine segmentations via majority voting or *SIMPLE* fusion Langerak et al. [27].

To this date BTK generated thousands of brain tumor segmentations for several research projects. For Technical University of Munich (TUM) internal projects, the computations could be distributed across multiple machines using the above-mentioned *Kraken* platform. A few of these research projects already made it to publication:

[1] Florian Kofler, Ivan Ezhov, Lucas Fidon, Carolin M Pirkl, Johannes C Paetzold, Egon Burian, Sarthak Pati, Malek El Husseini, Fernando Navarro, Suprosanna Shit, et al. "Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles." In: *Frontiers in Neuroscience* 15 (2021).

[2] KJ Paprottka, S Kleiner, C Preibisch, F Kofler, F Schmidt-Graf, C Delbridge, D Bernhardt, SE Combs, J Gempt, B Meyer, et al. "Fully automated analysis combining [18F]-FET-PET and multiparametric MRI including DSC perfusion and APTw imaging: a promising tool for objective evaluation of glioma progression." In: *European journal of nuclear medicine and molecular imaging* 48.13 (2021), pp. 4445–4455.

[3] Carolin M Pirkl, Laura Nunez-Gonzalez, Florian Kofler, Sebastian Endt, Lioba Grundl, Mohammad Golbabaee, Pedro A Gómez, Matteo Cencini, Guido Buonincontri, Rolf F Schulte, et al. "Accelerated 3D whole-brain T1, T2, and proton density mapping: feasibility for clinical glioma MR imaging." In: *Neuroradiology* 63.11 (2021), pp. 1831–1851.

[4] Marie Franziska Thomas, Florian Kofler, Lioba Grundl, Tom Finck, Hongwei Li, Claus Zimmer, Björn Menze, and Benedikt Wiestler. "Improving Automated Glioma Segmentation in Routine Clinical Use Through Artificial Intelligence-Based Replacement of Missing Sequences With Synthetic Magnetic Resonance Imaging Scans." In: *Investigative Radiology* 57.3 (2022), pp. 187–193.

[5] Andrey Zhylka, Nico Sollmann, Florian Kofler, Ahmed Radwan, Alberto De Luca, Jens Gempt, Benedikt Wiestler, Bjoern Menze, Sandro M Krieg, Claus Zimmer, et al. "Tracking the Corticospinal Tract in Patients With High-Grade Glioma: Clinical Evaluation of Multi-Level Fiber Tracking and Comparison to Conventional Deterministic Approaches." In: *Frontiers in oncology* 11 (2021), pp. 761169–761169.

8 TECHNICAL INNOVATION

Besides curating a better training set, researchers can advance model performance due to technical innovation. Besides ensembling, fine-tuning the network architecture and loss function to fit the problem at hand are among the most promising strategies.

8.1 ENSEMBLING

Ensembling and test-time augmentations usually lead to more robust model performance. Normally, *the bigger, the better* is a good rule of thumb for building ensembles. For instance, in BraTS glioma segmentation, the fusion of multiple segmentation algorithms turns out beneficial, even though the algorithms represent ensembles themselves. Here *SIMPLE* [27], a multi-iterative approach analyzing similarity between individual algorithms, outperforms a basic majority voting [19]. It is important to note that ensembling does not require developing different network architectures. Instead, Isensee et al. [16] propose training an ensemble of identical architectures on different subsets of the training data. Fort, Hu, and Lakshminarayanan [12] take this one step further and demonstrate that starting with only different random initialization can be sufficient.

8.2 NETWORK ARCHITECTURE

One of the *state-of-the-art* network architectures for segmentation tasks represents the U-Net [35], as illustrated in Figure 8.1. Beyond segmentation tasks, adding *skip connections*, as featured in the *U-Net*, should be considered when fine details from the input image matter.

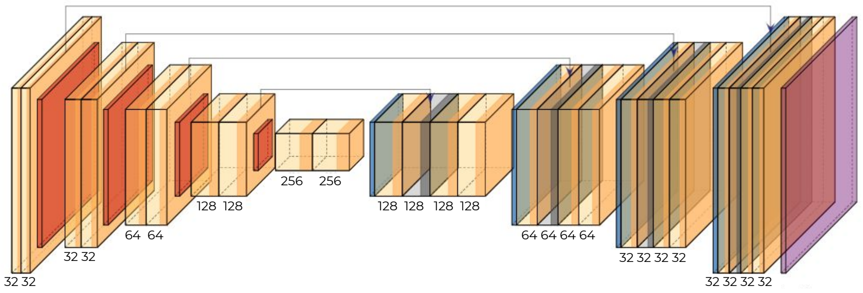


Figure 8.1: U-net network architecture. Just like in a conventional *Auto Encoder*, first, the *encoder* part of the CNN creates increasingly abstract representations of the input image (left part of the network). The decoder then reconstructs the images from the *latent space*. The innovation of the *U-net* [35] lies in adding *skip connections*. These directly connect layers from the *encoder* to the *decoder* and, thus, preserve fine details from the input image. The figure appears in the supplementary materials of [24].

Apart from this, a suitable network architecture has to be identified through experimentation. Bengio [7] provide practical recommendations for setting hyperparameter in CNNs training. Besides *standard* parameters, such as *learning rate*, *activation functions, number of parameters, network depth, upsampling methods* etc., should all be considered as tuneable hyperparameters of the network.

Apart from grid searching for a good set of hyperparameters, researchers develop new approaches to find good-performing sets of hyperparameters. A good starting point for diving into the field of *neural architecture search* is provided by Elsken, Metzen, and Hutter [11].

8.3 LOSS FUNCTION

Theoretical considerations might help when choosing a loss function. For instance, for training regression networks, Mean Square Error (MSE) loss might be preferred to Mean Absolute Error (MAE) loss when avoiding outliers is a priority, as it penalizes heavier for higher deviations from the ground truth. For segmentation networks, most researchers try to optimize DSC, and therefore, default to *(soft) Dice* loss. Isensee et al. [16] demonstrated that an equally weighted *Dice and Binary Cross Entropy* compound loss [16] often outperforms such a plain *(soft) Dice* loss.

In reinforcement learning, one of the main challenges in network training is operationalizing the problem so that the agent is correctly rewarded or penalized for learning meaningful behavior. In biomedical segmentation, we are often confronted with a similar problem, as many segmentation problems are poorly represented by optimizing DSC. MS lesion segmentation represents a particularly prominent example of this, as illustrated in Figure 8.2.

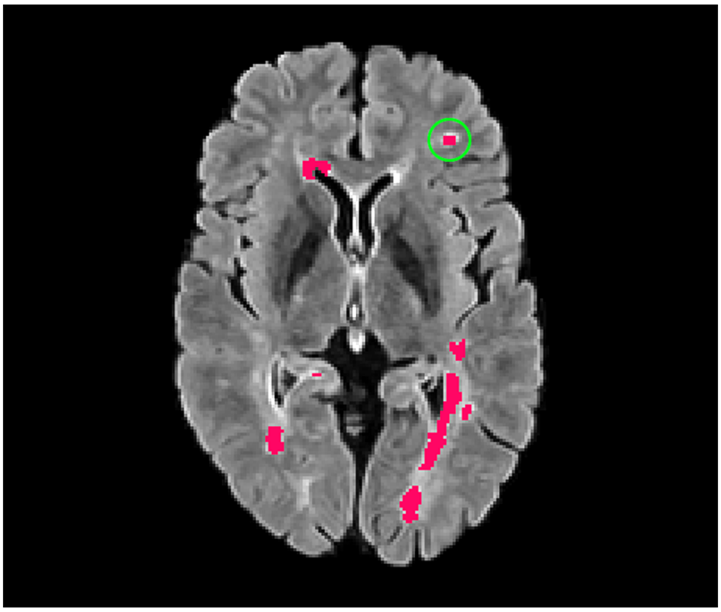


Figure 8.2: Multiple sclerosis (MS) lesions vs. Sørensen–Dice coefficient (DSC). The DSC between the segmentation with and without the MS lesion encircled in green is *0.9806*. Consequently, when computing the soft Dice loss, the network is hardly incentivized to pick up the highlighted lesion. Even though the lesion might be relevant from a clinical perspective and affect treatment decisions. The figure appears in [24].

MS lesions are largely inhomogeneous regarding the volume, texture, and other features. In ML terminology, the *instances* (lesions) are highly *imbalanced*. Optimizing DSC does not account for this and prioritizes big lesions. Researchers try to solve such problems by hand-crafting problem-specific losses.

For MS lesion segmentation, *Tversky loss* [37] promise to allow tuning the network's precision and sensitivity to individual preferences. Building upon this, Hashemi et al. [14] propose *assymetric loss functions*. Zhang et al. [47] introduce an auxiliary task with fixed-size spheres for each lesion. In contrast, Shirokikh et al. [40] account for lesion volume by using a weight map inversely to lesion size.

As can be learned from MS lesions segmentation, it is highly advisable to develop an understanding of the underlying biomedical problem to avoid framing an ill-posed ML problem. Usually, achieving a *voxel-perfect* segmentation has not the highest priority in clinical practice. [1]

Another good example of this is the segmentation of tubular structures such as blood vessels or street maps. Here the accurate repre-

[1] Especially when accounting for uncertainties within the annotations, see also chapter 5.

sentation of the network structure is most important. Therefore, Shit et al. [41] propose *centerlineDice* to preserve these details better. They achieve this by computing the *skeleton* of the structures of interest and adding a specific term for it in the loss function.

As always, before beginning to custom-tailor a loss function it makes sense to dive into the literature. Therefore, Ma et al. [28] provide a broad, but not exhaustive, overview of loss functions available for semantic segmentation tasks.

8.4 ACADEMIC IMPACT

A segmentation quality estimation method for segmentation ensembles is published [20]:

[1] Florian Kofler, Ivan Ezhov, Lucas Fidon, Carolin M Pirkl, Johannes C Paetzold, Egon Burian, Sarthak Pati, Malek El Husseini, Fernando Navarro, Suprosanna Shit, et al. "Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles." In: *Frontiers in Neuroscience* 15 (2021).

Further, *blob loss* tackles the problem of *instance imbalance* by providing existing loss functions with *instance imbalance awareness*. The *blob loss* manuscript is published in the proceedings of the *international conference on Information Processing in Medical Imaging (IPMI 2023)* and further available on arXiv:

[1] Florian Kofler, Suprosanna Shit, Ivan Ezhov, Lucas Fidon, Izabela Horvath, Rami Al-Maskari, Hongwei Bran Li, Harsharan Bhatia, Timo Loehr, Marie Piraud, et al. "blob loss: instance imbalance aware loss functions for semantic segmentation." In: *International Conference on Information Processing in Medical Imaging*. Springer Nature Switzerland Cham. 2023, pp. 755–767.

Applying *blob loss* helped to win the first place in the *segmentation of lacunes* subtask at the Where is VALDO - Vascular Lesions Detection Challenge 2021. Further, a third place was secured in the *segmentation of cerebral microbleeds* subtask [43].

Part IV

ACADEMIC CONTRIBUTIONS

9

BRATS TOOLKIT

BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice

Authors: *Florian Kofler*, Christoph Berger, Diana Waldmannstetter, Jana Lipkova, Ivan Ezhov, Giles Tetteh, Jan Kirschke, Claus Zimmer, Benedikt Wiestler, Bjoern H. Menze

Venue: *Frontiers in neuroscience 2020: 125* [19]

Abstract: Despite great advances in brain tumor segmentation and clear clinical need, translation of state-of-the-art computational methods into clinical routine and scientific practice remains a major challenge. Several factors impede successful implementations, including data standardization and preprocessing. However, these steps are pivotal for the deployment of state-of-the-art image segmentation algorithms. To overcome these issues, we present BraTS Toolkit. BraTS Toolkit is a holistic approach to brain tumor segmentation and consists of three components: First, the BraTS Preprocessor facilitates data standardization and preprocessing for researchers and clinicians alike. It covers the entire image analysis workflow prior to tumor segmentation, from image conversion and registration to brain extraction. Second, BraTS Segmentor enables orchestration of BraTS brain tumor segmentation algorithms for generation of fully-automated segmentations. Finally, Brats Fusionator can combine the resulting candidate segmentations into consensus segmentations using fusion methods such as majority voting and iterative SIMPLE fusion. The capabilities of our tools are illustrated with a practical example to enable easy translation to clinical and scientific practice.

Contribution: Project conception and coordination, implementation, data analysis, manuscript preparation

frontiers
in Neuroscience

TECHNOLOGY AND CODE
published: 29 April 2020
doi: 10.3389/fnins.2020.00125

BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice

Florian Kofler[1,2]*, *Christoph Berger*[1], *Diana Waldmannstetter*[1], *Jana Lipkova*[1], *Ivan Ezhov*[1], *Giles Tetteh*[1], *Jan Kirschke*[2], *Claus Zimmer*[2], *Benedikt Wiestler*[2†] and *Bjoern H. Menze*[1†]

[1] *Image-Based Biomedical Modeling, Department of Informatics, Technical University of Munich, Munich, Germany,*
[2] *Department of Neuroradiology, Klinikum rechts der Isar, Munich, Germany*

OPEN ACCESS

Edited by:
Kamran Avanaki,
Wayne State University, United States

Reviewed by:
Suyash P. Awate,
Indian Institute of Technology Bombay, India
Guido Gerig,
NYU Tandon School of Engineering, United States

Correspondence:
Florian Kofler
florian.kofler@tum.de

†These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *30 September 2019*
Accepted: *31 January 2020*
Published: *29 April 2020*

Citation:
Kofler F, Berger C, Waldmannstetter D, Lipkova J, Ezhov I, Tetteh G, Kirschke J, Zimmer C, Wiestler B and Menze BH (2020) BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. Front. Neurosci. 14:125. doi: 10.3389/fnins.2020.00125

Despite great advances in brain tumor segmentation and clear clinical need, translation of state-of-the-art computational methods into clinical routine and scientific practice remains a major challenge. Several factors impede successful implementations, including data standardization and preprocessing. However, these steps are pivotal for the deployment of state-of-the-art image segmentation algorithms. To overcome these issues, we present BraTS Toolkit. BraTS Toolkit is a holistic approach to brain tumor segmentation and consists of three components: First, the BraTS Preprocessor facilitates data standardization and preprocessing for researchers and clinicians alike. It covers the entire image analysis workflow prior to tumor segmentation, from image conversion and registration to brain extraction. Second, BraTS Segmentor enables orchestration of BraTS brain tumor segmentation algorithms for generation of fully-automated segmentations. Finally, Brats Fusionator can combine the resulting candidate segmentations into consensus segmentations using fusion methods such as majority voting and iterative SIMPLE fusion. The capabilities of our tools are illustrated with a practical example to enable easy translation to clinical and scientific practice.

Keywords: brain tumor segmentation, anonymization, MRI data preprocessing, medical imaging, brain extraction, BraTS, glioma

1. INTRODUCTION

Advances in deep learning have led to unprecedented opportunities for computer-aided image analysis. In image segmentation, the introduction of the U-Net architecture (Ronneberger et al., 2015) and subsequently developed variations like the V-Net (Milletari et al., 2016) or the 3D U-Net (Çiçek et al., 2016) have yielded algorithms for brain tumor segmentation that achieve a performance comparable to experienced human raters (Dvorak and Menze, 2015; Menze et al., 2015a; Bakas et al., 2018). A recent retrospective analysis of a large, multi-center cohort of glioblastoma patients convincingly demonstrated that objective assessment of tumor response via U-Net-based segmentation outperforms the assessment by human readers in terms of predicting patient survival (Kickingereder et al., 2019; Kofler et al., 2019), suggesting a potential benefit of implementing these algorithms into clinical routine.

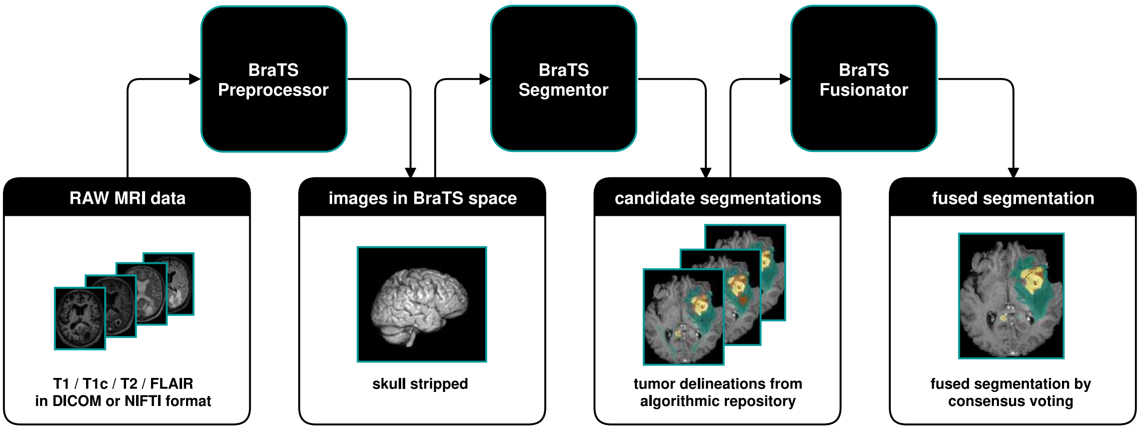


FIGURE 1 | Illustration of a typical dataflow to get from raw MRI scans to segmented brain tumors by combining the three components of the BraTS Toolkit. After preprocessing the raw MRI scans using the BraTS Preprocessor, the data is passed to the BraTS Segmentor, where arbitrary state-of-the-art models from the BraTS algorithmic repository can be used for segmentation. With BraTS Fusionator, multiple candidate segmentations may then be fused to obtain a consensus segmentation. As the Toolkit is designed to be completely modular and with clearly defined interfaces, each component can be replaced with custom solutions if required.

Recent works present diverse approaches toward brain tumor segmentation and analysis. Jena and Awate (2019) introduced a Deep-Neural-Network for image segmentation with uncertainty estimates based on Bayesian decision theory. Shboul et al. (2019) deployed feature-guided radiomics for glioblastoma segmentation and survival prediction. Jungo et al. (2018) analyzed the impact of inter-rater variability and fusion techniques for ground truth generation on uncertainty estimation. Shah et al. (2018) combined strong and weak supervision in training of their segmentation network to reduce overall supervision cost. Cheplygina et al. (2019) created an overview of Machine Learning methods in medical image analysis employing less or unconventional kinds of supervision.

In earlier years researchers experimented with a variety of approaches to tackle brain tumor segmentation (Prastawa et al., 2003; Menze et al., 2010, 2015b; Geremia et al., 2012), however in recent years the field is increasingly dominated by convolutional neural networks (CNN). This is also reflected in the contributions to the Multimodel Brain Tumor Segmentation Benchmark (BraTS) challenge (Bakas et al., 2018). The BraTS challenge (Menze et al., 2015a; Bakas et al., 2017) was introduced in 2012 at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), evaluating different algorithms for automated brain tumor segmentation. Therefore, every year the BraTS organizers provide a set of MRI scans, consisting of T1, T1c, T2, and FLAIR images from low- and high-grade glioma patients, coming with the corresponding ground truth segmentations.

Nonetheless, the computational methods presented in the BraTS challenge have not found their way into clinical and scientific practice. While the individual reasons vary, there are some key obstacles that impede the successful implementation of these algorithms. First of all, the availability of data for training, especially of high-quality, well-annotated data, is limited. Additionally, data protection as well as ethical barriers, complicate the development of centralized solutions, making local solutions strongly preferable. Furthermore, there are knowledge and skill barriers, when it comes to the conduction of setting up necessary preprocessing of data, while time and resources are limited.

While individual solutions for several of these problems exist, such as containerization for simplified distribution of code or public datasets, these are oftentimes fragmented and hence difficult to combine. Centralizing these efforts holds promise for making advances in image analysis easily available for broad implementation. Here we introduce three components to tackle these problems. First *BraTS preprocessor* facilitates data standardization and preprocessing for researchers and clinicians alike. Building upon that, varying tumor segmentations can be obtained from multiple algorithms with *BraTS Segmentor*. Finally, *BraTS Fusionator* can fuse these candidate segmentations into consensus segmentations by majority voting and iterative SIMPLE (Langerak et al., 2010) fusion. Together our tools represent *BraTS Toolkit* and enable a holistic approach integrating all the steps necessary for brain tumor image analysis.

2. METHODS

We developed BraTS Toolkit to get from raw DICOM data to fully automatically generate tumor segmentations in NIFTI format. The toolkit consists of three modular components. **Figure 1** visualizes how a typical brain tumor segmentation pipeline can be realized using the toolkit. The data is first preprocessed using the BraTS Preprocessor, then candidate segmentations are obtained from the BraTS Segmentor and finally fused via the BraTS Fusionator. Each component can be replaced with custom solutions to account for local requirements[1]. A key design principle of the software is that all data processing happens locally to comply with data privacy and protection regulations.

BraTS Toolkit comes as a python package and can be deployed either via Python or by using the integrated command line interface (CLI). As the software is subject to ongoing development and improvement this work focuses on more abstract descriptions of the software's fundamental design principles. To ease deployment in scientific and clinical practice an up-to-date user guide with installation and usage instructions can be found here: https://neuronflow.github.io/BraTS-Toolkit/.

Users that prefer an easier approach can alternatively use the BraTS Preprocessor's graphical user interface (GUI) to take care of the data preprocessing[2]. The GUI is constantly improved in a close feedback loop with radiologists from the department of Neuroradiology at Klinikum Rechts der Isar (Technical University of Munich) to address the needs of clinical practitioners. Depending on the community's feedback, we plan to additionally provide graphical user interfaces for BraTS Segmentor and BraTS Fusionator in the future. Therefore, BraTS Toolkit features update mechanisms to ensure that users have access to the latest features.

2.1. Component One: BraTS Preprocessor

BraTS Preprocessor provides image conversion, registration, and anonymization functionality. The starting point to use BraTS Preprocessor is to have T1, T1c, T2, and FLAIR imaging data in NIFTI format. DICOM files can be converted to NIFTI format using the embedded dcm2niix conversion software (Li et al., 2016).

The main output of BraTS Preprocessor consists of the anonymized image data of all four modalities in BraTS space. Moreover, it generates the original input images converted to BraTS space, anonymized data in native space, defacing/skullstripping masks for anonymization, registration matrices to convert between BraTS and native space and overview images of the volumes' slices in png format. **Figure 2** depicts the data-processing in detail.

BraTS Preprocessor handles standardization and preprocessing of brain MRI data using a classical front- and back end software architecture. **Figure 3** illustrates the GUI variant's software architecture, which enables users without programming knowledge to handle MRI data pre-processing steps.

Advanced Normalization Tools (ANTs) (Avants et al., 2011) are deployed for linear registration and transformation of images into BraTS space, independent of the selected mode. In order to achieve proper anonymization of the image data there are four different processing modes to account for different local requirements and hardware configurations:

1. GPU brain-extraction mode
2. CPU brain-extraction mode
3. GPU defacing mode (under development)
4. CPU defacing mode

Brain extraction is implemented by means of HD-BET (Isensee et al., 2019) using GPU or CPU, respectively. HD-BET is a deep learning based brain extraction method, which is trained on glioma patients and therefore particularly well-suited for our task. In case the available RAM is not sufficient the CPU mode automatically falls back to ROBEX (Iglesias et al., 2011). ROBEX is another robust, but slightly less accurate, skull-stripping method that requires less RAM than HD-BET, when running on CPU.

Alternatively, the BraTS Preprocessor features GPU and CPU defacing modes for users who find brain-extraction too destructive. Defacing on the CPU is implemented via Freesurfer's mri-deface (Fischl, 2012), while deep-learning based defacing on the GPU is currently under development.

2.2. Component Two: BraTS Segmentor

The Segmentor module provides a standardized control interface for the BraTS algorithmic repository[3] (Bakas et al., 2018). This repository is a collection of Docker images, each containing a Deep Learning model and accompanying code designed for the BraTS challenge. Each model has a rigidly defined interface to hand data to the model and retrieve segmentation results from the model. This enables the application of state-of-the-art models for brain tumor segmentation on new data without the need to install additional software or to train a model from scratch. However, even though the algorithmic repository provides unified models, it is still up to the interested user to download and run each Docker image individually as well as manage the input and output. This final gap in the pipeline is closed by the Segmentor, which enables less experienced users to download, run and evaluate any model in the BraTS algorithmic repository. It provides a front end to manage all available containers and run them on arbitrary data, as long as the data conforms to the BraTS format. To this end, the Segmentor provides a command line interface to process data with any or all of the available Docker images in the repository while ensuring proper handling of the files. Its modular structure also allows anyone to extend the code, include other Docker containers or include it as a Python package.

2.3. Component Three: BraTS Fusionator

The Segmentor module can generate multiple segmentations for a given set of images which usually vary in accuracy and without prior knowledge, a user might be unsure which segmentation is the most accurate. The Fusionator module provides two methods to combine this arbitrary number of segmentation candidates into one final fusion which represents the consensus of all available segmentations. There are two main methods offered: Majority voting and the selective and iterative method for performance level estimation (SIMPLE) proposed by

[1] As an example users who do not want to generate tumor segmentations on their own hardware using the BraTS Segmentor, can alternatively try our experimental web technology based solution nicknamed the Kraken: https://neuronflow.github.io/kraken/.

[2] For an up-to-date installation and user guide please refer to: https://neuronflow.github.io/BraTS-Preprocessor/.

[3] https://github.com/BraTS/Instructions/blob/master/Repository_Links.md#brats-algorithmic-repository

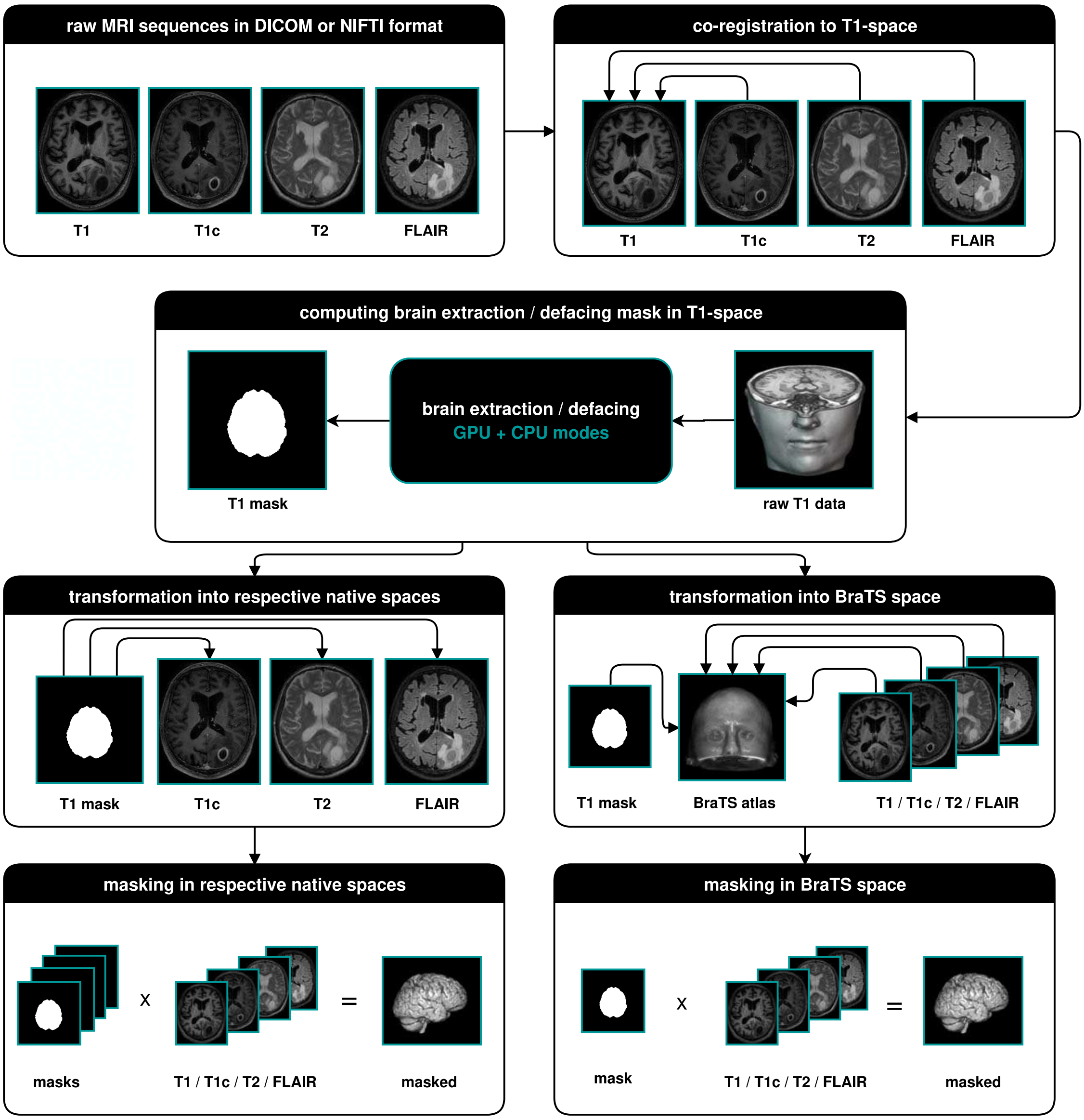

raw MRI sequences in DICOM or NIFTI format
T1
T1c
T2
FLAIR
co-registration to T1-space
T1
T1c
T2
FLAIR
computing brain extraction / defacing mask in T1-space
T1 mask
brain extraction / defacing
GPU + CPU modes
raw T1 data
transformation into respective native spaces
T1 mask
T1c
T2
FLAIR
transformation into BraTS space
T1 mask
BraTS atlas
T1 / T1c / T2 / FLAIR
masking in respective native spaces
masks
x
T1 / T1c / T2 / FLAIR
=
masked
masking in BraTS space
mask
x
T1 / T1c / T2 / FLAIR
=
masked

FIGURE 2 | Illustration of the data-processing. We start with a T1, T1c, T2, and FLAIR volume. In a first step we co-register all modalities to the T1 image. Depending on the chosen mode, we then compute the brain segmentation or defacing mask in T1-space. To morph the segmented images in native space, we transform the mask to the respective native spaces and multiply it with the volumes. For obtaining the segmented images in BraTS space, we transform the masks and volumes to the BraTS space using a brain atlas. We then apply the masks to the volumes.

Langerak et al. (2010). Both methods take all available candidate segmentations produced by the algorithms of the repository and combine each label to generate a final fusion. In majority voting, a class is assigned to a given voxel if at least half of the candidate segmentations agree that this voxel is of a certain class. This is repeated for each class to generate the complete segmentation. The SIMPLE fusion works as follows: First, a majority vote fusion with all candidate segmentations is performed. Secondly, each candidate segmentation is compared to the current consensus fusion and the resulting overlap score (a standard DICE measure in the proposed method) is used as a weight for the majority voting. This causes the candidate segmentations with higher estimated accuracy to have a higher influence on the final result. Lastly, each candidate segmentation with an accuracy below a certain threshold is dropped out after each iteration. This iterative process is stopped once the consensus fusion converges. After repeating the processes for each label, a final segmentation is obtained.

3. RESULTS

The broad availability of Python, Electron.js, and Docker allows us to support all major operating systems with an easy installation process. Users can choose to process data using the command line

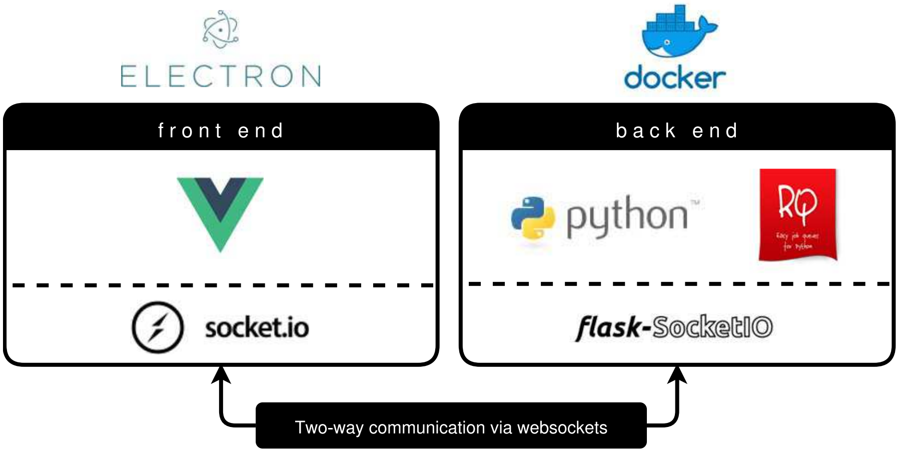


FIGURE 3 | BraTS Preprocessor software architecture (GUI variant). The front end is implemented by a Vue.js web application packaged via Electron.js. To ensure a constant runtime environment the Python based back end resides in a Docker container (Merkel, 2014). Redis Queue allows for load balancing and parallelization of the processing. The architecture enables two-way communication between front end and back end by implementing Socket.IO on the former and Flask-Socket.IO on the latter. In contrast to this the python package's front end is implemented using python-socketio.

(CLI) or through the user friendly graphical user interface (GUI). Depending on the available hardware, multiple threads are run to efficiently use the system's resources.

3.1. Practicality in Clinical and Scientific Practice

To test the practicality of BraTS Toolkit we conducted a brain tumor segmentation experiment on 191 patients of the BraTS 2016 dataset. As a first step we generated candidate tumor segmentations. BraTS Segmentor allowed us to rapidly obtain tumor delineations from ten different algorithms of the BraTS algorithmic repository (Bakas et al., 2018). The standardized user interface of BraTS Segmentor abstracts all the required background knowledge regarding docker and the particularities of the algorithms. In the next step we used BraTS Fusionator to fuse the generated segmentations by consensus voting. **Figure 4** shows that fusion by iterative SIMPLE and class-wise majority voting had a slight advantage over single algorithms. This effect was particularly driven by removal of false positives as illustrated for an exemplary patient in **Figure 5**. BraTS Toolkit enabled us to conduct the experiment in a user-friendly way. With only a few lines of Python code we were able to obtain segmentation results in a fully-automated fashion. This impression was confirmed by experiments on further in house data-sets where we also deployed the CLI and GUI variants of all three BraTS Toolkit components with great feedback from clinical and scientific practitioners. Users especially appreciated the increased robustness and precision of consensus segmentations compared to existing single algorithm solutions.

4. DISCUSSION

Overall, the BraTS Toolkit is a step toward the democratization of automatic brain tumor segmentation. By lowering resource and knowledge barriers, users can effectively disseminate dockerized brain tumor segmentation algorithms collected through the BraTS challenge. Thus, it makes objective brain tumor volumetry, which has been demonstrated to be superior to traditional image assessment (Kickingereder et al., 2019), readily available for scientific and clinical use.

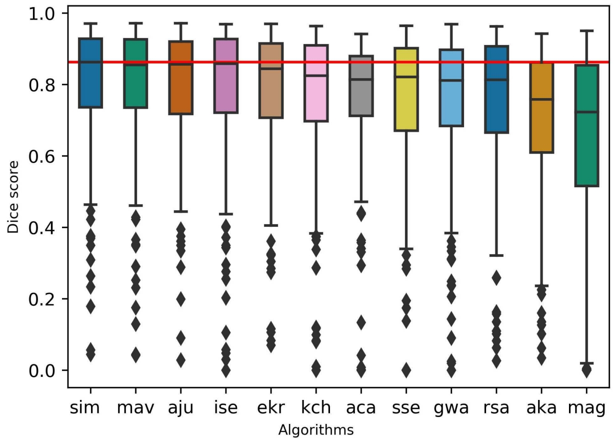


FIGURE 4 | Evaluation of the segmentation results on the BraTS 2016 data set for whole tumor labels on $n = 191$ evaluated test cases. We generated candidate segmentations with ten different algorithms. Segmentation methods are sorted in descending order by mean dice score. The two fusion methods, iterative SIMPLE (sim) and class-wise majority voting displayed on the left, outperformed individual algorithms depicted further right. The red horizontal line shows the SIMPLE median dice score ($M = 0.863$) for better comparison.

Currently, BraTS segmentation algorithms and therefore BraTS Segmentor require each of T1, T1c, T2, and FLAIR sequences to be present. In practice, this can become a limiting factor due to errors in data acquisition or incomplete protocols leading to missing modalities. Recent efforts try to bridge this

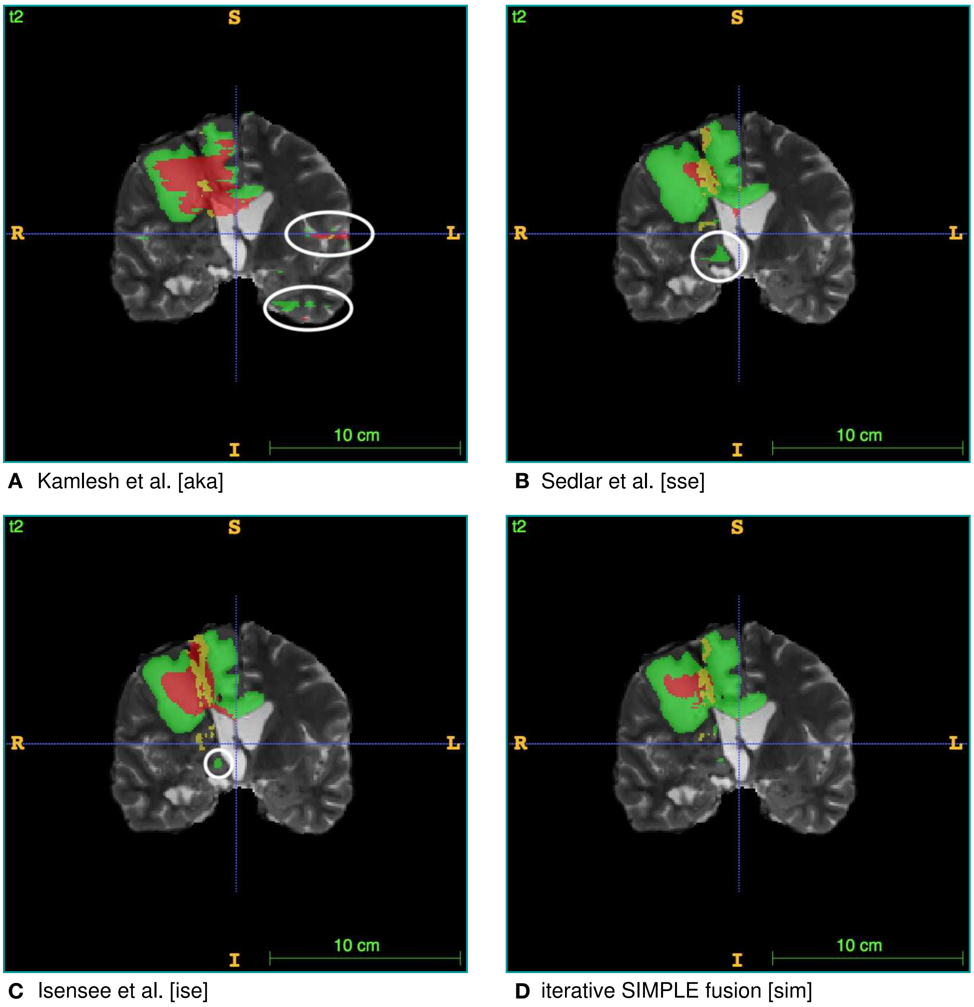


FIGURE 5 | Single algorithm vs. iterative SIMPLE consensus segmentation. T2 scans with segmented labels by exemplary candidate algorithms from **(A)** Pawar et al. (2018), **(B)** Sedlar (2018), and **(C)** Isensee et al. (2017) (Green: edema; Red: necrotic region/non-enhancing tumor; Yellow: enhancing tumor). **(D)** Shows a consensus segmentation obtained using the iterative SIMPLE fusion. Notice the false positives marked with white circles on the candidate segmentations. These outliers are effectively reduced in the fusion segmentation shown in **(D)**.

gap by using machine learning techniques to reconstruct missing image modalities (e.g., Dorent et al., 2019; Li et al., 2019).

Other crucial aspects of data preprocessing are the lack of standards for pulse sequences across different scanners and manufacturers, and absence of data acquisition protocols' harmonization in general. For the moment, we address this only with primitive image standardization strategies as described in **Figure 2**. However, in clinical and scientific practice, we already found our application to be very robust across different data sources. Brain extraction with HD-BET also proved to be sound for patients from multiple institutions with different pathologies (Isensee et al., 2019).

These limitations are in fact some of the key motivations for our initiative. We strive to provide researchers with tools to build comprehensive databases which capture more of the data variability in magnetic resonance imaging. In the longterm this will enable the development of more precise algorithms. With BraTS Toolkit clinicians can actively contribute to this process.

Through well-defined interfaces, the resulting output from our software can be integrated seamlessly with further downstream software to create new scientific and medical applications such as but not limited to, fully-automatic MR reporting[4] or tumor growth modeling (Ezhov et al., 2019; Lipková et al., 2019). Another promising future direction is to focus on integration with the local PACS to enable streamlined processing of imaging data directly from the radiologist's workplace.

[4] Our Kraken web service can be seen as an an exemplary prototype for this (for the moment it is not for clinical use, but for research and entertainment purposes only). The Kraken is able to send automatically generated segmentation and volumetry reports to the user's email address: https://neuronflow.github.io/kraken/.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

FK conceptualized the BraTS Toolkit, programmed the BraTS Preprocessor and contributed to paper writing. CB programmed and conceptualized the BraTS Fusionator and BraTS Segmentor and contributed to paper writing. DW, JL, IE, and JK conceptualized the BraTS Preprocessor and contributed to paper writing. GT conceptualized the BraTS Preprocessor software architecture and contributed to paper writing. CZ conceptualized the BraTS Preprocessor and provided feedback on the BraTS Fusionator. BW and BM conceptualized the BraTS Toolkit and contributed to paper writing.

FUNDING

BM, BW, and FK are supported through the SFB 824, subproject B12. Supported by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. With the support of the Technical University of Munich–Institute for Advanced Study, funded by the German Excellence Initiative.

REFERENCES

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *NeuroImage* 54, 2033–2044. doi: 10.1016/j.neuroimage.2010.09.025

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.

Cheplygina, V., de Bruijne, M., and Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Analysis* 54, 280–296. doi: 10.1016/j.media.2019.03.009

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and 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), 424–432.

Dorent, R., Joutard, S., Modat, M., Ourselin, S., and Vercauteren, T. (2019). "Hetero-modal variational encoder-decoder for joint modality completion and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019* eds D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan (Cham: Springer International Publishing), 74–82.

Dvorak, P., and Menze, B. (2015). "Local structure prediction with convolutional neural networks for multimodal brain tumor segmentation", in *Medical Computer Vision: Algorithms for Big Data*, (Cham: Springer International Publishing), 59–71.

Ezhov, I., Lipkova, J., Shit, S., Kofler, F., Collomb, N., Lemasson, B., et al. (2019). "Neural parameters estimation for brain tumor growth modeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer International Publishing), 787–795.

Fischl, B. (2012). Freesurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Geremia, E., Menze, B. H., Ayache, N. (2012). "Spatial decision forests for glioma segmentation in multi-channel mr images," in *MICCAI Challenge on Multimodal Brain Tumor Segmentation*, (Citeseer), 34.

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. Med. Imaging*, 30, 1617–1634. doi: 10.1109/TMI.2011.2138152

Isensee, F., Kickingereder, P., Bonekamp, D., Bendszus, M., Wick, W., Schlemmer, H.-P., et al. (2017). "Brain tumor segmentation using large receptive field deep convolutional neural networks," in *Bildverarbeitung für die Medizin 2017* (Springer), 86–91.

Isensee, F., Schell, M., Tursunova, I., Brugnara, G., Bonekamp, D., Neuberger, U., et al. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, (Wiley Online Library), 40, 4952–964. doi: 10.1002/hbm.24750

Jena, R., and Awate, S. P. (2019). "A bayesian neural net to segment images with uncertainty estimates and good calibration," in *International Conference on Information Processing in Medical Imaging* (Springer), 3–15.

Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., et al. (2018). "On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer International Publishing), 682–690.

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., et al. (2019). Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 20, 728–740. doi: 10.1016/S1470-2045(19)30098-1

Kofler, F., Paetzold, J., Ezhov, I., Shit, S., Krahulec, D., Kirschke, J., et al. (2019). "A baseline for predicting glioblastoma patient survival time with classical statistical models and primitive features ignoring image information," in *International MICCAI Brainlesion Workshop* (Springer).

Langerak, T. R., van der Heide, U. A., Kotte, A. N., Viergever, M. A., Van Vulpen, M., Pluim, J. P., et al. (2010). Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *IEEE Trans. Med. Imaging* 29, 2000–2008. doi: 10.1109/TMI.2010.2057442

Li, H., Paetzold, J. C., Sekuboyina, A., Kofler, F., Zhang, J., Kirschke, J. S., et al. (2019). "Diamondgan: unified multi-modal generative adversarial networks for mri sequences synthesis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, eds D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan (Cham: Springer International Publishing), 795–803.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* 264, 47–56. doi: 10.1016/j.jneumeth.2016.03.001

Lipková, J., Angelikopoulos, P., Wu, S., Alberts, E., Wiestler, B., Diehl, C., et al. (2019). Personalized radiotherapy design for glioblastoma: integrating mathematical tumor models, multimodal scans, and bayesian inference. *IEEE Trans. Med. Imaging* 38, 1875–1884. doi: 10.1109/TMI.2019.2902044

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015a). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Menze, B. H., Van Leemput, K., Lashkari, D., Riklin-Raviv, T., Geremia, E., Alberts, E., et al. (2015b). A generative probabilistic model and discriminative extensions for brain lesion segmentation—with application to tumor and stroke. *IEEE Trans. Med. Imaging* 35, 933–946. doi: 10.1109/TMI.2015.2502596

Menze, B. H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., and Golland, P. (2010). "A generative model for brain tumor segmentation in multi-modal images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Berlin; Heidelberg: Springer Berlin Heidelberg), 151–159.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux J.* 2014:2.

Milletari, F., Navab, N., and 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), 565–571.

Pawar, K., Chen, Z., Shah, N. J., and Egan, G. (2018). "Residual encoder and convolutional decoder neural network for glioma segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 263–273.

Prastawa, M., Bullitt, E., Moon, N., Van Leemput, K., and Gerig, G. (2003). Automatic brain tumor segmentation by subject specific modification of atlas priors1. *Acad. Radiol.* 10, 1341–1348. doi: 10.1016/S1076-6332(03)00506-3

Ronneberger, O., Fischer, P., Brox, and Thomas. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Sedlar, S. (2018). "Brain tumor segmentation using a multi-path cnn based method," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 403–422.

Shah, M. P., Merchant, S., and Awate, S. P. (2018). "Ms-net: mixed-supervision fully-convolutional networks for full-resolution segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 379–387.

Shboul, Z. A., Alam, M., Vidyaratne, L., Pei, L., Elbakary, M. I., and Iftekharuddin, K. M. (2019). Feature-guided deep radiomics for glioblastoma patient survival prediction. *Front. Neurosci.* 13:966. doi: 10.3389/fnins.2019.00966

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

10 FAILURE PREDICTION

Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles

Authors: *Florian Kofler*, Ivan Ezhov, Lucas Fidon, Carolin M Pirkl, Johannes C Paetzold, Egon Burian, Sarthak Pati, Malek El Husseini, Fernando Navarro, Suprosanna Shit, Jan Kirschke, Spyridon Bakas, Claus Zimmer, Benedikt Wiestler, Bjoern H Menze

Venue: *Frontiers in neuroscience 2021: 15* [21]

Abstract: A multitude of image-based machine learning segmentation and classification algorithms has recently been proposed, offering diagnostic decision support for the identification and characterization of glioma, Covid-19 and many other diseases. Even though these algorithms often outperform human experts in segmentation tasks, their limited reliability, and in particular the inability to detect failure cases, has hindered translation into clinical practice. To address this major shortcoming, we propose an unsupervised quality estimation method for segmentation ensembles. Our primitive solution examines discord in binary segmentation maps to automatically flag segmentation results that are particularly error-prone and therefore require special assessment by human readers. We validate our method both on segmentation of brain glioma in multimodal magnetic resonance - and of lung lesions in computer tomography images. Additionally, our method provides an adaptive prioritization mechanism to maximize efficacy in use of human expert time by enabling radiologists to focus on the most difficult, yet important cases while maintaining full diagnostic autonomy. Our method offers an intuitive and reliable uncertainty estimation from segmentation ensembles and thereby closes an important gap toward successful translation of automatic segmentation into clinical routine.

Contribution: Project conception and coordination, experiment design and implementation, data analysis, manuscript preparation

ORIGINAL RESEARCH
published: 30 December 2021
doi: 10.3389/fnins.2021.752780

Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles

Florian Kofler[1,2,3*], *Ivan Ezhov*[1,3], *Lucas Fidon*[4], *Carolin M. Pirkl*[1], *Johannes C. Paetzold*[1,3], *Egon Burian*[2], *Sarthak Pati*[1,5,6,7], *Malek El Husseini*[1,2], *Fernando Navarro*[1,3,8], *Suprosanna Shit*[1,3], *Jan Kirschke*[2], *Spyridon Bakas*[5,6,7], *Claus Zimmer*[2], *Benedikt Wiestler*[2†] *and Bjoern H. Menze*[1,9†]

[1] Department of Informatics, Technical University Munich, Munich, Germany, [2] Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [3] TranslaTUM - Central Institute for Translational Cancer Research, Technical University of Munich, Munich, Germany, [4] School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom, [5] Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Pennsylvania, PA, United States, [6] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Pennsylvania, PA, United States, [7] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Pennsylvania, PA, United States, [8] Department of Radio Oncology and Radiation Therapy, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [9] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

OPEN ACCESS

Edited by:
Gabriele Lohmann,
Max Planck Institute for Biological Cybernetics, Germany

Reviewed by:
Xue Feng,
University of Virginia, United States
Qian Wang,
Shanghai Jiao Tong University, China

Correspondence:
Florian Kofler
florian.kofler@tum.de

[†]*These authors share senior authorship*

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *03 August 2021*
Accepted: *26 October 2021*
Published: *30 December 2021*

Citation:
Kofler F, Ezhov I, Fidon L, Pirkl CM, Paetzold JC, Burian E, Pati S, El Husseini M, Navarro F, Shit S, Kirschke J, Bakas S, Zimmer C, Wiestler B and Menze BH (2021) Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles.
Front. Neurosci. 15:752780.
doi: 10.3389/fnins.2021.752780

A multitude of image-based machine learning segmentation and classification algorithms has recently been proposed, offering diagnostic decision support for the identification and characterization of glioma, Covid-19 and many other diseases. Even though these algorithms often outperform human experts in segmentation tasks, their limited reliability, and in particular the inability to detect failure cases, has hindered translation into clinical practice. To address this major shortcoming, we propose an unsupervised quality estimation method for segmentation ensembles. Our primitive solution examines discord in binary segmentation maps to automatically flag segmentation results that are particularly error-prone and therefore require special assessment by human readers. We validate our method both on segmentation of brain glioma in multi-modal magnetic resonance - and of lung lesions in computer tomography images. Additionally, our method provides an adaptive prioritization mechanism to maximize efficacy in use of human expert time by enabling radiologists to focus on the most difficult, yet important cases while maintaining full diagnostic autonomy. Our method offers an intuitive and reliable uncertainty estimation from segmentation ensembles and thereby closes an important gap toward successful translation of automatic segmentation into clinical routine.

Keywords: quality estimation, failure prediction, anomaly detection, ensembling, fusion, OOD, CT, MR

1. INTRODUCTION

Advances in deep learning for segmentation have facilitated the automated assessment of a variety of anatomies and pathologies in medical imaging. In particular for glioma, automatic segmentation has shown great promise as a basis for objective assessment of tumor response (Kickingereder et al., 2019). In segmentation challenges such as BraTS (Menze et al., 2015), VerSe (Sekuboyina et al., 2021) and LiTS (Bilic and et al., 2019) virtually all top-performing solutions are based on ensembling. Recent efforts such as *HD-GLIO* (Kickingereder et al., 2019; Isensee et al., 2021), *GaNDLF* (Pati et al., 2021), and *BraTS Toolkit* (Kofler et al., 2020) have paved the way to apply state-of-the-art deep-learning ensembles in clinical practice. Even though algorithms often outperform human readers (Kofler et al., 2021), algorithmic reliability remains a major obstacle toward safe implementation of automated segmentation (and hence volumetry) into clinical routine (D'Amour et al., 2020). Researchers in the field of Out-of-Distribution (OOD) detection try to address this shortcoming by discovering systematic patterns within convolutional neural networks (CNN) (Schölkopf et al., 2001; Jungo et al., 2018; Mehrtash et al., 2020; Berger et al., 2021; Ruff et al., 2021). These sophisticated anomaly detection methods have the disadvantage of being limited to CNNs, often specific CNN architectures.

In contrast, we present a primitive, and therefore more applicable, solution exploiting discord in binary segmentation maps to estimate segmentation quality in an unsupervised fashion. We evaluate our method on segmentation of brain glioma in multi-modal magnetic resonance (MR)—and of lung lesions in computer tomography (CT) images. Our method allows detecting error-prone segmentation results, which require special assessment by human readers. Working only on binary segmentation maps enables our method to analyze the segmentations of human readers, classical machine learning, and modern deep learning approaches interchangeably. As segmentations are the basis for objective disease assessment as well as subsequent image analysis, our method addresses an urgent need for improving the trustworthiness of automatic segmentation methods. Furthermore, by implementing our method healthcare providers can streamline efficient use of human workforce, arguably the most persistent and major bottleneck in healthcare service worldwide (Krengli et al., 2020; Starace et al., 2020).

2. METHODS

2.1. Unsupervised Quality Estimation

Figure 1 depicts the quality estimation procedure. By aggregating and comparing multiple candidate segmentations, cases with large discordance, therefore a high chance of failure, can be rapidly identified. In more detail, our method consists of the following steps:

1. We obtain candidate segmentations from all methods in an ensemble, and then compute a fusion from the candidate segmentations.
2. We calculate similarity metrics between the fused segmentation result and the individual candidate segmentations.
3. We obtain the threshold for setting an alarm value by subtracting the *median absolute deviation (mad)* of the similarity metric times the tunable parameter α from its *median* value. This happens individually for each candidate image. We prefer the *median* based statistics for their better robustness toward statistical outliers. For metrics that are negatively correlated with segmentation performance, such as Hausdorff distance, we propose to use the additive inverse.
4. We set an alarm flag if the individual similarity metric is below the computed threshold. For *infinite* (or *Nan*) values, which can for instance happen for distance-based metrics such as Hausdorff distance, alarm flags are raised too.
5. Finally, we accumulate the alarm flags to obtain risk scores and therefore quality estimation for each image.

The results of this procedure are illustrated in **Figure 4**. We hypothesize that a higher count of alarm flags is associated with worse segmentation quality, here measured by lower volumetric Dice performance.

2.2. MR Experiment: Multi-Modal Brain Tumor Segmentation

To test the validity of our approach we use BraTS Toolkit *(btk)* (Kofler et al., 2020) to create a segmentation ensemble for brain glioma in multi-modal magnetic resonance (MR) images. Therefore, we incorporate five segmentation algorithms (Feng et al., 2019; Isensee et al., 2019; McKinley et al., 2019, 2020; Zhao et al., 2019) developed within the scope of the BraTS challenge (Menze et al., 2015; Bakas et al., 2017a,b,c, 2018). We compute alarms according to the above procedure based on Dice similarity and Hausdorff distances.

2.2.1. Fusions and Segmentation Metrics

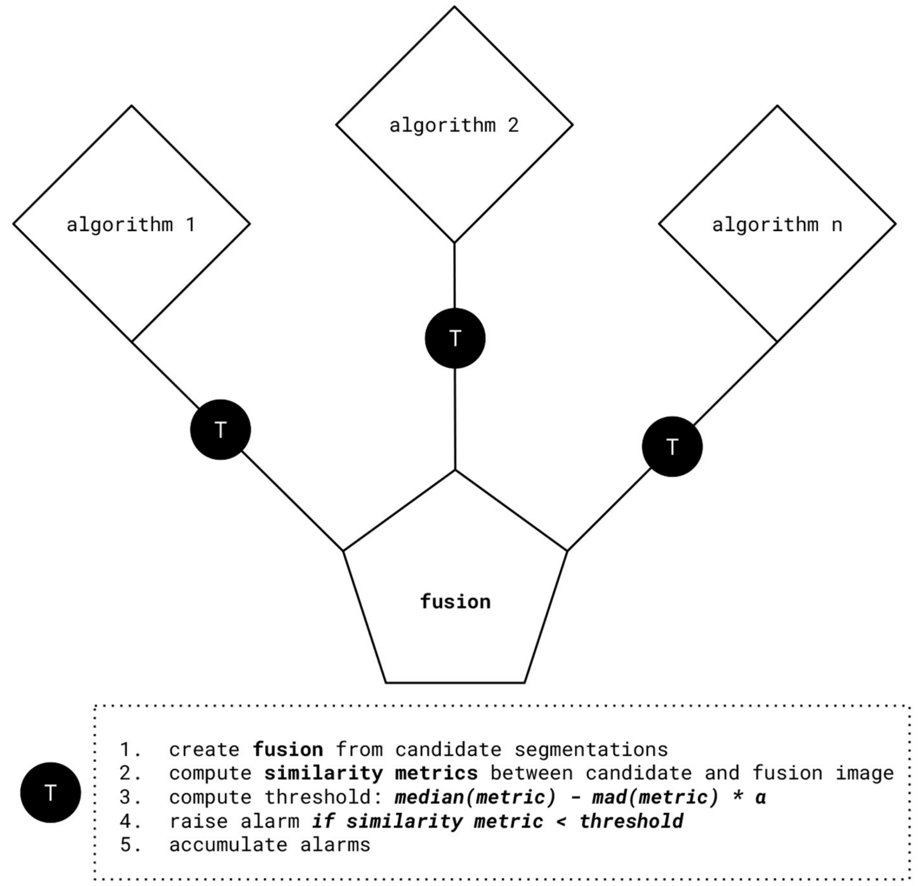
We fuse the segmentations with an equally weighted majority voting using *btk* (Kofler et al., 2020) and compute segmentation quality metrics with *pymia* (Jungo et al., 2021). **Figure 2** illustrates fusions and individual segmentations with an example exam.

2.2.2. Data

We evaluate on a dataset of 68 cases capturing the wide diversity in glioma imaging. Our dataset consists of 15 high-grade glioma (HGG) from the publicly available Rembrandt dataset (Gusev et al., 2018), as well as another 25 HGG from TUM university hospital (MRI TUM). Furthermore, we evaluate 13 low-grade glioma (LGG) from Rembrandt and 15 from MRI TUM. Two expert radiologists generated the ground truth segmentations using *ITK-SNAP* (Yushkevich et al., 2006) and corrected each other's tumor delineations.

2.3. CT Experiment: COVID-19 Lung CT Lesion Segmentation

For further validation, we compose an ensemble based on the MONAI challenge baseline (MONAI CORE Team, 2020)



```
1.  create fusion from candidate segmentations
2.  compute similarity metrics between candidate and fusion image
3.  compute threshold: median(metric) - mad(metric) * α
4.  raise alarm if similarity metric < threshold
5.  accumulate alarms
```

FIGURE 1 | Quality estimation procedure. After computing fusion from the candidate segmentations, similarity metrics between the fused and the candidate segmentations are evaluated. Using this information, we obtain threshold values by subtracting the median absolute deviation (mad) of similarity metrics times the tunable parameter α from their median value. We set an alarm flag if the individual similarity metric is below the computed threshold, for example: $median(Dice) - mad(Dice) * \alpha$.

developed for the *COVID-19 Lung CT Lesion Segmentation Challenge - 2020* (Clark et al., 2013). To segment lung lesions in computer tomography (CT) images, the code implements a 3d-Unet inspired by Falk et al. (2019). q2a1 We first train the original baseline for 500 epochs. Then we generate a small ensemble of three networks by warmstarting the training with the baseline's model weights and replacing the following parameters for the respective model for training another 500 epochs:

To obtain our first model (ADA) we swap the baseline's original Adam optimizer to *AdamW* (Loshchilov and Hutter, 2019). In a similar fashion, the second model (RAN) utilizes Ranger (Wright, 2019) to make use of Gradient Centralization (Yong et al., 2020). Our third model (AUG) adds an augmentation pipeline powered by batchgenerators (Isensee et al., 2020), torchio (Pérez-García et al., 2020), and native MONAI augmentations. In addition we switch the optimizer to stochastic gradient descent (*SGD*) with momentum (momentum = 0.95).

Our metric for training progress is the volumetric Dice coefficient. All networks are trained with an equally weighted Dice plus binary cross-entropy loss. The training is stopped once we observe no further improvements for the validation set. We conduct model selection by choosing the respective model with the best volume Dice score on the validation set. The code for the CNN trainings is publicly available via GitHub (***censored to maintain the double blind review process***).

2.3.1. Fusions and Segmentation Metrics

To unify the individual outputs of our ensembles' components to a segmentation mask we choose SIMPLE (Langerak et al., 2010) fusion. SIMPLE is an iterative fusion method introduced by Langerak et al., which tends to outperform generic majority voting across various segmentation problems. An example segmentation for one exam is illustrated in **Figure 3**. We generate SIMPLE fusions using BraTS Toolkit (Kofler et al., 2020) and generate alarms for Dice scores calculated with *pymia* (Jungo et al., 2021). Segmentation quality metrics, in particular volumetric Dice coefficient and Hausdorff distances, for the test set are obtained through the challenge portal (COVID Challenge Team, 2021).

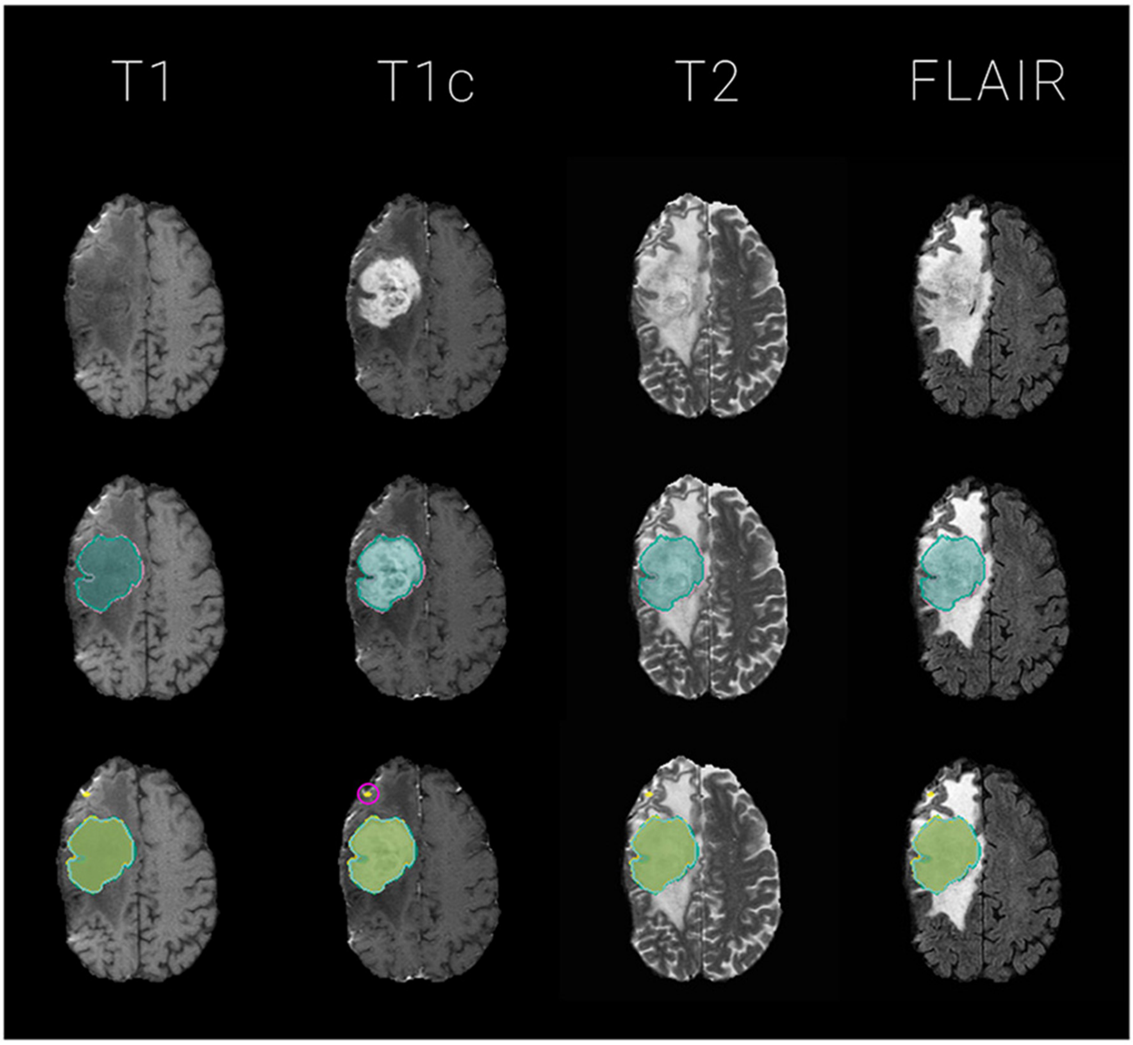


FIGURE 2 | Exemplary glioma segmentation exam with multi-modal MR. Segmentations are overlayed on T1, T1c, T2, FLAIR images for the tumor's center of mass, defined by the *tumor core* (*necrosis* and *enhancing tumor*) of the ground truth label. The segmentation outlines represent the *tumor core* labels, meaning the sum of *enhancing tumor* and *necrosis* labels. **Top**: the four input images without segmentation overlay; **Middle**: ground truth segmentation (*GT*) in *reddish purple* vs. majority voting fusion (*mav*) in *bluish green*; **Bottom**: *mav* fusion in *bluish green* vs. individual segmentation algorithms in various colors. Notice the small outliers encircled in pink on the frontal lobe which probably contribute to the raise of 3 Dice - and 4 Hausdorff distance based alarms for this particular exam with a mediocre volumetric Dice similarity coefficient with the *ground truth* data of *0.66*.

2.3.2. Data

We run our experiments on the public dataset of the COVID-19 Lung CT Lesion Segmentation Challenge - 2020 (COVID Challenge Team, 2021), supported by the Cancer Imaging Archive (TCIA) (Clark et al., 2013).

2.4. Calibration of Alpha (α)

The α parameter can be fine-tuned to account for different optimization targets and adjusted dynamically depending on workload, e.g., in an extreme triage scenario, an alarm flag could only be raised for the strongest outliers, hence a high α should be chosen. Once the situation has been amended, α can be reset to a smaller value, resulting in a more sensitive failure prediction.

With the default value $\alpha = 0$ the threshold is set to the median. Therefore, approximately half of the cases will trigger an alarm for each metric. Alternatively, alpha can be automatically adjusted to maximize the Pearson correlation coefficient with a segmentation quality metric or entropy, or combinations thereof. **Tables 1**, **2** illustrate how the distributions of alarm counts correlate with Dice performance and the resulting entropy in response to variations in α.

Note that α can also be adjusted for each segmentation target class, as well as, each of the ensemble's components, and for each similarity metric on an individual basis to fine-tune the quality estimation toward specific needs. For instance, hence the *enhancing tumor* label is of higher clinical relevance for glioma

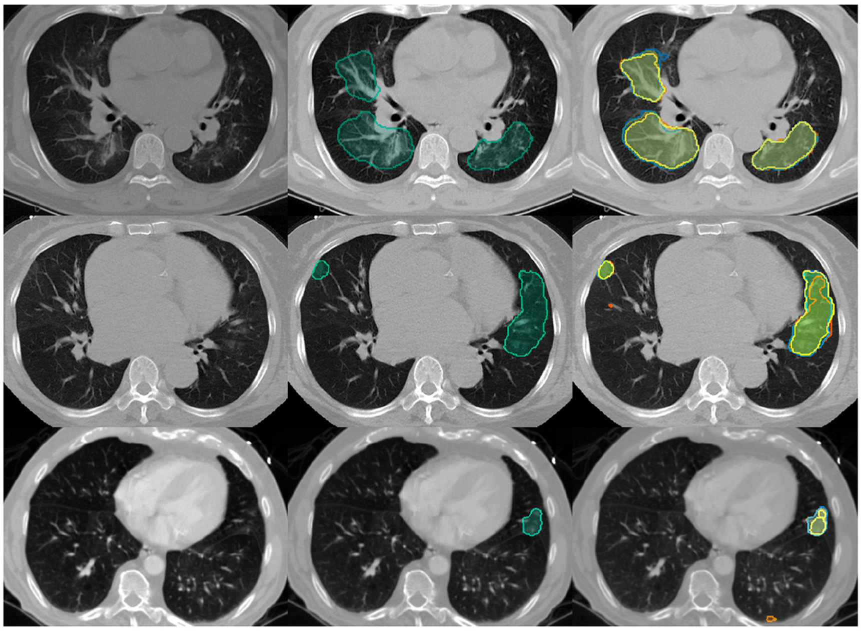


FIGURE 3 | Example Covid-19 lung lesion segmentation exams with CT images. Segmentations are overlayed for the lesions' center of mass, defined by the slice with most lesion voxels: **Left**: the empty input images; **Middle**: SIMPLE segmentation fusion (simple) in *bluish green*; **Right**: SIMPLE fusion in *bluish green* vs. individual segmentation algorithms in various colors. The volumetric Dice similarity coefficients with the *ground truth* and respective alarm counts are as following: Top row: *0.81, 0*; Middle row: *0.58, 2*; Last row: *0.14, 3*.

TABLE 1 | Distribution of alarm counts depending on α for the MR experiment: The table illustrates the number of images classified in the individual alarm count categories *(a)* from *0* to *10*; for different values of α.

Alpha	Entropy	r:dice	r:hd	0a	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a
−3.00	−0.00	NA	NA	0	0	0	0	0	0	0	0	0	0	68
−2.00	0.22	NA	0.04	0	0	0	0	0	0	0	0	0	4	64
−1.00	1.28	−0.27	−0.2	0	0	0	1	0	2	2	5	5	13	40
−0.75	1.80	−0.55	−0.27	0	0	1	3	4	2	10	5	4	14	25
−0.50	2.02	−0.63	−0.3	0	6	1	3	5	4	11	1	10	8	19
−0.25	2.33	−0.7	−0.38	3	5	4	5	7	4	6	7	8	7	12
−0.10	2.37	−0.73	−0.41	7	4	4	6	4	7	7	8	7	6	8
0.00	2.35	−0.76	−0.45	9	5	7	4	4	6	6	8	8	3	8
0.10	2.30	−0.77	−0.46	9	6	10	3	6	7	2	9	5	3	8
0.25	2.28	−0.77	−0.51	11	7	12	3	2	7	3	8	5	5	5
0.50	2.23	−0.78	−0.59	15	11	8	3	2	4	5	8	4	4	4
0.75	2.06	−0.73	−0.59	18	13	7	3	1	5	6	7	2	6	0
1.00	1.97	−0.72	−0.58	23	12	3	3	2	6	8	6	3	2	0
2.00	1.71	−0.66	−0.55	30	10	6	4	3	8	2	5	0	0	0
3.00	1.40	−0.65	−0.52	37	11	4	1	3	10	1	1	0	0	0

Additionally, we depict the Pearson correlation coefficients for the Dice (r:dice) *- and Hausdorff distance* (r:hd) *based alarm counts with volumetric Dice segmentation performance, as well as the respective alarm count distribution's entropy. The selected value for* α *of 0.1 is highlighted in pink The resulting computed thresholds are depicted in* ***Table 3****.*

(Weller et al., 2014), one might consider setting the associated thresholds to more conservative values using a smaller *alpha*.

For simplicity, we set parameter α to *0.1* for each class, component and metric in our analysis. This results in a slightly less conservative failure prediction compared to the default.

3. RESULTS

Our method accurately predicts the segmentation performance in both experiments and is able to capture segmentation failures. Even though our code is not optimized for speed, the

TABLE 2 | Distribution of alarm counts depending on α for the CT experiment: The table illustrates the number of images classified in the individual alarm count categories *(a)* from *0* to *3*; for different values of α.

Alpha	Entropy	r:dice	0a	1a	2a	3a
−3.00	−0.00	NA	0	0	0	46
−2.00	−0.00	NA	0	0	0	46
−1.00	0.58	−0.45	0	3	5	38
−0.75	0.88	−0.56	5	2	6	33
−0.50	1.19	−0.67	6	7	8	25
−0.25	1.32	−0.64	10	7	10	19
−0.10	1.36	−0.73	12	8	11	15
0.00	1.37	−0.7	13	8	14	11
0.10	1.37	−0.7	15	10	11	10
0.25	1.33	−0.62	18	9	11	8
0.50	1.20	−0.61	23	6	12	5
0.75	1.17	−0.69	25	9	8	4
1.00	1.13	−0.71	26	10	6	4
2.00	0.86	−0.67	33	8	2	3
3.00	0.66	−0.62	37	6	1	2

Additionally, we depict the Pearson correlation coefficients for the Dice (r:dice) *based alarm counts with volumetric Dice segmentation performance, as well as the respective alarm count distribution's entropy. The selected value for α of 0.1 is highlighted in pink. The resulting computed Dice similarity thresholds are as following: ADA: 0.9489; RAN: 0.9446; AUG: 0.9024.*

computation of the fused segmentation masks, similarity metrics and resulting alarm counts is a matter of seconds. Quantitative metrics for the MR and CT experiment are summarized in **Figure 4**.

3.1. MR Experiment

Setting α to *0.1* leads to an even distribution across alarm count groups, (see **Tables 1**, **3**). **Figure 4A** plots the average Dice coefficients across the tumors labels: *enhancing tumor, necrosis and edema* against the alarm count. We observe a strong negative correlation between segmentation performance and increasing alarm count: *Pearson's* $r = -0.72$, $p = 3.874e\text{-}12$. This is also reflected in the Hausdorff distance, (see **Figure 4B**).

3.2. CT Experiment

Choosing an α of *0.1* leads to an even distribution across alarm count groups, (see **Table 2**). **Figure 4C** plots Dice coefficients[1] on the challenge test set against alarm count. As for the MR experiment, we find a strong negative correlation between segmentation performance and increasing alarm count: *Pearson's* $r = \text{-}0.70$, *p-value* $= 4.785e\text{-}08$. As observed before, this effect is mirrored by the Hausdorff distance, (see **Figure 4D**).

4. DISCUSSION

It is important to note that, the validity of our method is closely tied to the chosen evaluation metrics' representation of segmentation performance (Kofler et al., 2021). For our experiments, we evaluate the volumetric Dice score and Hausdorff distance. Based on this fundamental assumption, we provide an unsupervised quality estimation for segmentation ensembles that does not perform any background diagnostic decisions and fully maintains the radiologists' diagnostic autonomy.

We demonstrate efficacy for two different use cases, namely multi-modal glioma segmentation in brain MR and Covid-19 lesion segmentation in lung CT images. The sensitivity of our method can be fine-tuned to specific requirements by adjusting α for ensemble components, classes, and segmentation quality metrics. Additionally, the low computational requirements make it easy to integrate into existing pipelines as computing the alarms takes only seconds and creates very little overhead.

Even though there are various efforts, such as the *BraTS algorithmic repository*[2], to facilitate clinical translation of state-of-the-art segmentation algorithms, quality estimation mechanisms represent a currently unmet, yet important milestone on the road toward reliably deploying deep learning segmentation pipelines in clinical practice. The proposed solution can assist clinicians in navigating the plethora of exams, which have to be reviewed daily. It provides a neat prioritization mechanism, maximizing the efficient use of human expert time, by enabling focus on the most difficult, yet important cases.

It is important to note further limitations of our method. First of all, it can only be applied to model ensembles and not to single algorithms. However, as most top-performing segmentation solutions employ ensembling techniques there is a broad field of potential application. Second, the computation of alarms relies on discordance in the ensemble. If all components of the ensemble converge to predicting the same errors they cannot be detected. Notably, we did not observe such a case in our experiments, even though our CT segmentation ensemble featured only three models employing the same architecture and little variation in training parameters. As our method profits from bigger ensembles and more variations in the network training, one could argue that our experiment is probably more difficult than most real-world scenarios. Along these lines, Roy et al. (2019) activated dropout during inference and Fort et al. (2020) demonstrated that it might be enough to choose different random initialization to achieve variance in network outputs. Third, even though the default value of α, *0* and *0.1*, which we chose for demonstration purposes, performed well in our experiments, there might be segmentation problems for which α needs to be manually fine-tuned.

Future research could investigate whether α how global thresholding, instead of the proposed individual thresholding per algorithm, affects the results. It should also be explored whether the methodology can be improved by including further

[1] Our basic ensemble reaches a median volumetric Dice score of *0.67*. We observe a wide performance distribution with a minimum of *0*, a maximum of *0.93* and a standard deviation of *0.25* around a mean of *0.61*, as displayed in **Figure 4C**. With regard to volumetric Dice coefficients mainly low-performing outliers separate our method from the top-performing methods in the challenge.

[2] https://www.med.upenn.edu/sbia/brats2017/algorithms.html

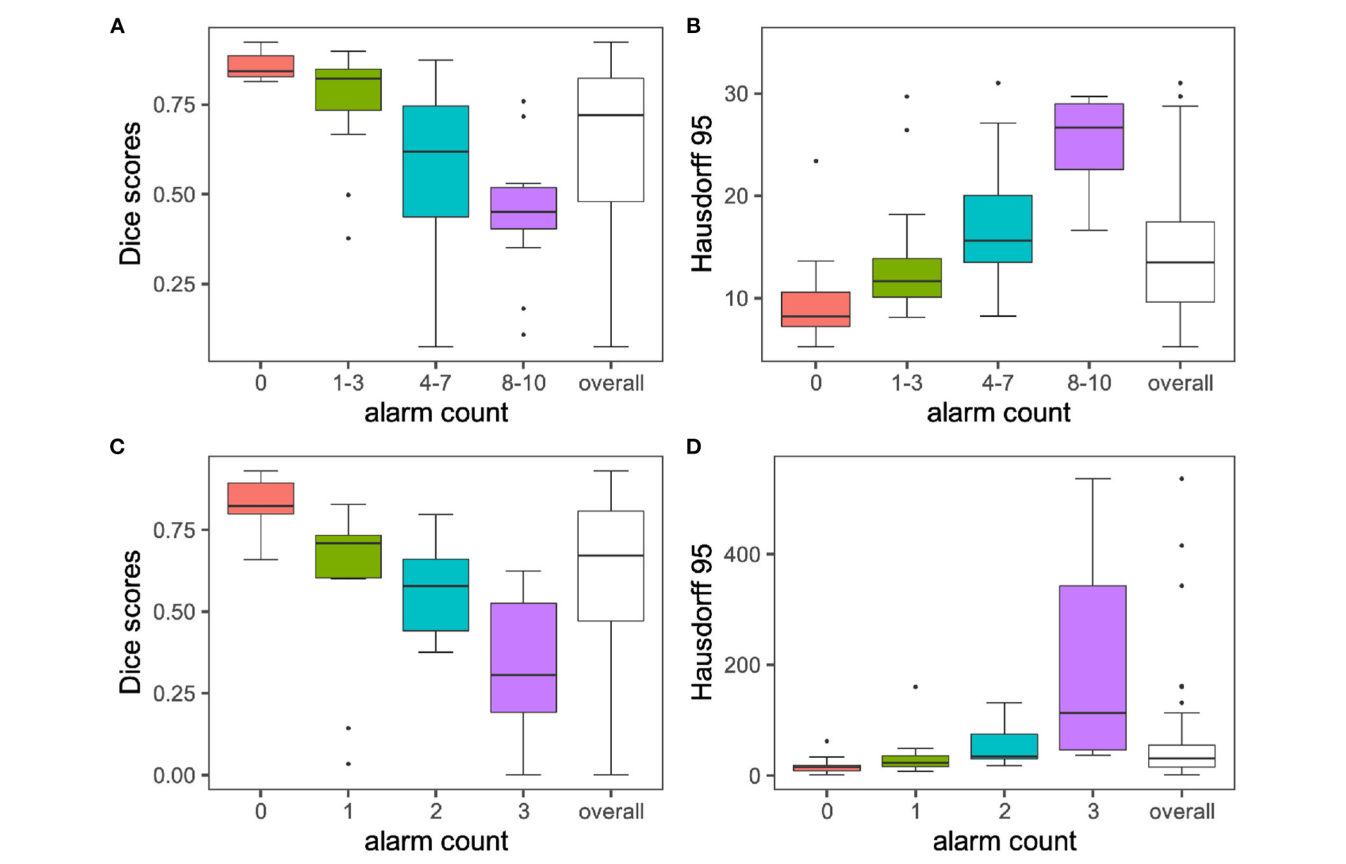


FIGURE 4 | Segmentation performances vs. alarm counts. The group means are illustrated with horizontal black lines. For display purposes only the 0–95 percent quantile is displayed for Hausdorff distances on the y-axis. In line with the performance of the volumetric Dice coefficient, Hausdorff distances increase with increasing alarm count. Infinite values for Hausdorff distances, which can happen when ground truth or prediction are empty, are excluded from the plot. Subplots **(A)** + **(B)** illustrate findings for the MR experiment, while subplots **(C)** + **(D)** depict results for the CT experiment.

TABLE 3 | Thresholds computed with $\alpha = 0.1$ for the MR experiment per algorithm: The columns *Dice* and *Hausdorff* depict, the respective volumetric Dice and Hausdorff distance based thresholds for the alarm computation for each of the segmentation algorithms.

Algorithm	Citation	Dice	Hausdorff
micdkfz	Isensee et al., 2019	0.9055	10.2277
xfeng	Feng et al., 2019	0.9092	8.9835
scan2019	McKinley et al., 2020	0.9147	8.8292
scan	McKinley et al., 2019	0.9084	10.4850
zyx	Zhao et al., 2019	0.9293	8.4451

segmentation metrics and to which extend it generalizes to other segmentation problems.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The CT data can be found here: https://covid-segmentation.grand-challenge.org/data/. The MR data will be published at: https://neuronflow.github.io/btk_evaluation/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

FK, IE, and LF contributed to conception and design of the study. FK, IE, CP, and JP wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The annotation of the dataset was made possible through the joint work of Children's National Hospital, NVIDIA, and

National Institutes of Health for the COVID-19-20 Lung CT Lesion Segmentation Grand Challenge. BM, BW, and FK are supported through the SFB 824, subproject B12. Supported by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. LF, SS, and IE are supported by the Translational Brain Imaging Training Network (TRABIT) under the European Union's Horizon 2020 research & innovation program (Grant agreement ID: 765148). With the support of the Technical University of Munich - Institute for Advanced Study, funded by the German Excellence Initiative. JP and SS are supported by the Graduate School of Bioengineering, Technical University of Munich. Research reported in this publication was partly supported by the National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH), under award numbers NCI:U01CA242871 and NINDS:R01NS042645. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH. Research reported in this publication was partly supported by AIME GPU cloud services.

ACKNOWLEDGMENTS

We would like to thank the YouTube channel Primitive Technology for inspiration regarding the manuscript's title and methodology.

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM Collection*. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG Collection*. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*.

Berger, C., Paschali, M., Glocker, B., and Kamnitsas, K. (2021). Confidence-based out-of-distribution detection: a comparative study and analysis. *arXiv preprint arXiv:2107.02568*. doi: 10.1007/978-3-030-87735-4_12

Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., et al. (2019). The liver tumor segmentation benchmark (LiTS). *arXiv preprint arXiv:1901.04056*.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

COVID Challenge Team (2021). *COVID Challenge*. Available online at: https://covid-segmentation.grand-challenge.org/Data/

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., et al. (2019). U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* 16, 67–70. doi: 10.1038/s41592-018-0261-2

Feng, X., Tustison, N., and Meyer, C. (2019). "Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer), 279–288. doi: 10.1007/978-3-030-11726-9_25

Fort, S., Hu, H., and Lakshminarayanan, B. (2020). Deep ensembles: a loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J.-C., Fine, H., and Madhavan, S. (2018). The rembrandt study, a large collection of genomic data from brain cancer patients. *Sci. Data* 5:180158. doi: 10.1038/sdata.2018.158

Isensee, F., and et al. (2019). "No new-net," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 234–244. doi: 10.1007/978-3-030-11726-9_21

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Isensee, F., Jager, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., et al. (2020). *Batchgenerators - A Python Framework for Data Augmentation*. doi: 10.5281/zenodo.3632567

Jungo, A., Meier, R., Ermis, E., Herrmann, E., and Reyes, M. (2018). Uncertainty-driven sanity check: application to postoperative brain tumor cavity segmentation. *arXiv preprint arXiv:1806.03106*.

Jungo, A., Scheidegger, O., Reyes, M., and Balsiger, F. (2021). pymia: a python package for data handling and evaluation in deep learning-based medical image analysis. *Comput. Methods Prog. Biomed.* 198:105796. doi: 10.1016/j.cmpb.2020.105796

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., et al. (2019). Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 20, 728–740. doi: 10.1016/S1470-2045(19)30098-1

Kofler, F., Berger, C., Waldmannstetter, D., Lipkova, J., Ezhov, I., Tetteh, G., et al. (2020). Brats toolkit: translating brats brain tumor segmentation algorithms into clinical and scientific practice. *Front. Neurosci.* 14:125. doi: 10.3389/fnins.2020.00125

Kofler, F., Ezhov, I., Isensee, F., Balsiger, F., Berger, C., Koerner, M., et al. (2021). Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for cnn training beyond rolling the dice coefficient. *arXiv preprint arXiv:2103.06205*.

Krengli, M., Ferrara, E., Mastroleo, F., Brambilla, M., and Ricardi, U. (2020). Running a radiation oncology department at the time of coronavirus: an Italian experience. *Adv. Radiat. Oncol.* 5, 527–530. doi: 10.1016/j.adro.2020.03.003

Langerak, T. R., van der Heide, U. A., Kotte, A. N., Viergever, M. A., Van Vulpen, M., and Pluim, J. P. (2010). Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Trans. Med. Imaging* 29, 2000–2008. doi: 10.1109/TMI.2010.2057442

Loshchilov, I., and Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

McKinley, R., Meier, R., and Wiest, R. (2019). "Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer), 456–465. doi: 10.1007/978-3-030-11726-9_40

McKinley, R., Rebsamen, M., Meier, R., and Wiest, R. (2020). "Triplanar ensemble of 3D-to-2D CNNs with label-uncertainty for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi and S. Bakas (Cham: Springer International Publishing), 379–387. doi: 10.1007/978-3-030-46640-4_36

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., and Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* 39, 3868–3878. doi: 10.1109/TMI.2020.3006437

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

MONAI CORE Team (2020). *MONAI*. doi: 10.5281/zenodo.4323059

Pati, S., Thakur, S. P., Bhalerao, M., Baid, U., Grenko, C., Edwards, B., et al. (2021). Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv preprint arXiv:2103.01006*.

Pérez-García, F., Sparks, R., and Ourselin, S. (2020). TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv preprint arXiv:2003.04696*. doi: 10.1016/j.cmpb.2021.106236

Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C. (2019). Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *Neuroimage* 195, 11–22. doi: 10.1016/j.neuroimage.2019.03.042

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., et al. (2021). A unifying review of deep and shallow anomaly detection. *Proc. IEEE* 109, 756–795. doi: 10.1109/JPROC.2021.3052449

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.* 13, 1443–1471. doi: 10.1162/089976601750264965

Sekuboyina, A., Husseini, M. E., Bayat, A., Loffler, M., Liebl, H., Li, H., et al. (2021). Verse: a vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Anal.* 2021:102166. doi: 10.1016/j.media.2021.102166

Starace, V., Brambati, M., Battista, M., Capone, L., Gorgoni, F., Cavalleri, M., et al. (2020). A lesson not to be forgotten. Ophthalmologists in Northern Italy become internists during the SARS-CoV-2 pandemic. *Am. J. Ophthalmol.* 220, 219–220. doi: 10.1016/j.ajo.2020.04.044

Weller, M., van den Bent, M., Hopkins, K., Tonn, J. C., Stupp, R., Falini, A., et al. (2014). Eano guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *Lancet Oncol.* 15, e395-e403. doi: 10.1016/S1470-2045(14)70011-7

Wright, L. (2019). Ranger - a synergistic optimizer. *GitHub Repos*. Available online at: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer

Yong, H., Huang, J., Hua, X., and Zhang, L. (2020). Gradient centralization: a new optimization technique for deep neural networks. *arXiv preprint arXiv:2004.01461*. doi: 10.1007/978-3-030-58452-8_37

Yushkevich, P. A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015

Zhao, Y.-X., Zhang, Y. M., Song, M., and Liu, C. L. (2019). "Multi-view semi-supervised 3D whole brain segmentation with a self-ensemble network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 256–265. doi: 10.1007/978-3-030-32248-9_29

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

11 PRIMITIVE SURVIVAL PREDICTION

A Baseline for Predicting Glioblastoma Patient Survival Time with Classical Statistical Models and Primitive Features Ignoring Image Information

Authors: *Florian Kofler*, Johannes C Paetzold, Ivan Ezhov, Suprosanna Shit, Daniel Krahulec, Jan S Kirschke, Claus Zimmer, Benedikt Wiestler, and Bjoern H Menze

Venue: The manuscript was published in the proceedings of *International MICCAI Brainlesion Workshop. Springer, Cham, 2019.* [23].

To avoid violation of copyrights the final manuscript submitted to *Springer* for typesetting is reproduced on the following pages.

Abstract: Gliomas are the most prevalent primary malignant brain tumors in adults. Until now an accurate and reliable method to predict patient survival time based on medical imaging and meta-information has not been developed. Therefore, the survival time prediction task was introduced to the Multimodal Brain Tumor Segmentation Challenge (BraTS) to facilitate research in survival time prediction.

Here we present our submissions to the BraTS survival challenge based on classical statistical models to which we feed the provided metadata as features. We intentionally ignore the available image information to explore how patient survival can be predicted purely by metadata. We achieve our best accuracy on the validation set using a simple median regression model taking only patient age into account. We suggest using our model as a baseline to benchmark the added predictive value of sophisticated features for survival time prediction.

Contribution: Project conception and coordination, method implementation, data analysis, manuscript preparation

A baseline for predicting glioblastoma patient survival time with *classical statistical* models and *primitive* features ignoring image information.

Florian Kofler[1], Johannes C. Paetzold[1], Ivan Ezhov[1], Suprosanna Shit[1], Daniel Krahulec[2], Jan S. Kirschke[1], Claus Zimmer[1], Benedikt Wiestler[1]⋆, and Bjoern H. Menze[1]⋆

[1] Technical University Munich, 81675 Munich, Germany
[2] Philips Healthcare, MR R&D Clinical Science, 5684 PC Best, Netherlands
{florian.kofler, johannes.paetzold}@tum.de

Abstract. Gliomas are the most prevalent primary malignant brain tumors in adults. Until now an accurate and reliable method to predict patient survival time based on medical imaging and meta-information has not been developed [4]. Therefore, the survival time prediction task was introduced to the Multimodal Brain Tumor Segmentation Challenge (BraTS) to facilitate research in survival time prediction.
Here we present our submissions to the BraTS survival challenge based on classical statistical models to which we feed the provided metadata as features. We intentionally ignore the available image information to explore how patient survival can be predicted purely by metadata. We achieve our best accuracy on the validation set using a simple median regression model taking only patient age into account. We suggest using our model as a baseline to benchmark the added predictive value of sophisticated features for survival time prediction.

Keywords: survival time prediction · glioma · glioblastoma · HGG · LGG · brain tumor · benchmark · BraTS · medical imaging · MRI

1 Introduction

Accurate estimation of a patient's prognosis is at the heart of clinical decision-making, both for clinical trials as well as daily clinical care.

Survival time prediction and statistics are frequently requested not only by terminally ill patients, but also by the general public. Survival time prognosis is considered as one of the most important factors in palliative medicine for three major reasons [23]:

⋆ contributed equally as senior authors

1. Necessity for medical law and insurance decisions, e.g. in the United States of America two independent doctors have to agree on a survival prognosis to decide on hospice eligibility.
2. Survival prognosis is critical for medical decision making, which weights the risks of medical procedures against expected benefits. For instance, in pain management, it can be beneficial to deliver addictive and potentially harmful doses of antidepressants and neurolytic agents to patients with short life expectancy.
3. Lifetime prognosis enables doctors to assist patients in making critical life decisions [19].

Considering the relevance of reliable survival time predictions, it is particularly striking how statistics reveal that clinicians are often unsuccessful in predicting patient survival [7]. Many studies have shown this issue, e.g. [18] found that about 50 percent of survival predictions for patients with lung cancer are erroneous. Specifically, the patients did not survive half of the predicted time frame or survived more than double the predicted time. Most clinicians' predictions of survival time are overly optimistic [7]. An important finding is that, the longer the patient-doctor relationship exists, the larger the optimism bias is within the doctor's survival prognosis [5]. This indicates that human subjectivity is a major source of error, besides the difficulties for clinicians to integrate prognostic information from multiple sources (e.g. demographic, genomic or imaging information).

These inconsistencies and relevant bias in clinicians survival prediction demand more quantitative approaches such as statistical or learning based models to assist in creating more realistic survival predictions. This has been empirically studied in the literature for various terminal diseases, e.g. by Henderson et al. for patients with lung cell cancer, using statistical models [9]. Recently, learning based methods exploring image information have proven to outperform medical doctors in survival time predictions for a multitude of diseases [14,11].

The Brain Tumor Segmentation Challenge (BraTS) focuses on a specific type of brain neoplasms called gliomas. Gliomas are one of the most prevalent brain tumors in adults and can be roughly distinguished in two major classes: aggressive high-grade gliomas, and low-grade gliomas. The life expectancy of a patient with a high-grade glioma has a median remaining life span of fewer than two years, while for low-grade gliomas, it is more than five years [17]. The survival prediction task was introduced to the BraTS challenge to crowd-source the development of an accurate and generalizable prediction model [1,2,3,4,16].

The BraTS dataset was acquired at multiple clinical centers, therefore presenting several real-world challenges. For instance, scans are often acquired using different imaging protocols, and follow-up scans are acquired at varying time points. These inconsistencies, among others, pose severe problems to clinicians as well as automated diagnostic approaches.

In the BraTS survival challenge, the images as well as corresponding metadata are given to the participants to predict patient survival. Most contributions to the challenge explore image information using learning models, e.g. U-Net [4]. However, in the BraTS2018 survival challenge, a simple linear regression considering only patient age and simple tumor region sizes as features achieved third place [21,20]. This could be attributed to a lack of larger and diverse datasets, which could be resolved in future challenges by extending the data across clinics or using recently successful generative approaches [8,15]. Another methodological reason could be the insufficient structure of extracted imaging features and contradicting feature interpretation.

Inspired by Weninger et al. [20], we systematically explored how far one can get using only metadata for survival time prediction. We intentionally disregard image information and instead explore a multitude of classical statistical models and metadata based features.

2 Methods

2.1 Models

As a baseline, we implemented simple ordinary least squares (OLS) linear models [6]. Additionally, we fitted linear model with three orthogonal polynomials [10] and quantile regression models [12]. We computed p-values and confidence intervals for the model coefficients and evaluated the goodness of fit of the models by adjusted $\boldsymbol{R^2}$ for the linear models and the quantile models by $\boldsymbol{V}$, as suggested by Koenker respectively [13].

2.2 Features

We deliberately ignored image information and instead focused on primitive features extracted from the patients' metadata. Besides the patients' age we included resection status [22] and the clinical institution (extracted from the patient ID e.g. "CBICA") as predictors for our models. The clinical institution feature differentiates regional factors such as access to healthcare, different population etc. that might affect survival time.

2.3 Dataset

As the test set includes only patients with gross total resection (GTR), we evaluated our models' performance on the GTR subset. Additionally, we also took patients with only partial tumor resections into account to find out whether we can retrieve additional information from these cases.

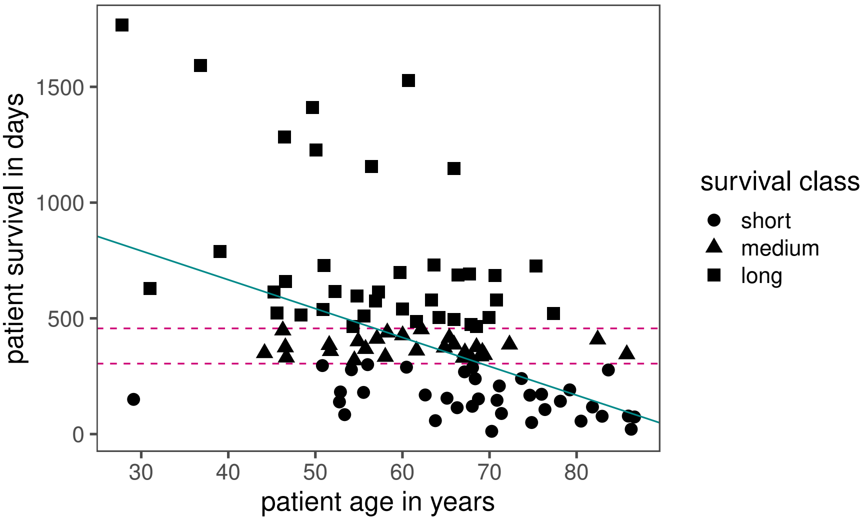


Fig. 1. Scatterplot of patient age versus survival time (*Pearson r*: -0.486), representing the space we fit our models to in the spirit of this XKCD comic [https://xkcd.com/2048/]. The dashed lines represent the thresholds to distinguish between short, medium and long-term survivors. The solid cyan line illustrates predictions of our proposed median regression model based on patient age.

3 Results

We designed our models on the training dataset considering measures of goodness of fit and p-values for model coefficients. Promising model configurations were evaluated on the validation dataset via the CBICA's Image Processing Portal (IPP).

3.1 Evaluation on the training set

The simple ordinary least squares (OLS) model outperformed the polynomial models on the training set, as reflected by higher values of adjusted R^2, see Table 1. Resection status and the clinical institution failed to add significant predictive value. These findings were also reflected in the analysis for the quantile models. For all models we achieved a much better fit on the subset of patients with gross total resection.

3.2 Evaluation on the validation set

Next, we evaluated on the validation set of 29 patients using the CBICA's Image Processing Portal (IPP). For the BraTS survival challenge, the predicted survival times are mainly evaluated by the accuracy of the survival prediction and

Table 1. Result table comparing the goodness of fit on the training set. We calculate the adjusted coefficient of determination R^2 for the OLS and polynomial models and V for the median model. The quantile model cannot be directly compared to the other models as the considered coefficients of determination (R^2 and V) are not the same as introduced by [13].

Model		$\mathbf{R^2}$ all data	$\mathbf{R^2}$ GTR only	**V** all data	**V** GTR only
OLS model	age	0.129	0.229	-	-
	age, resec., inst.	0.147	-	-	-
	age, inst.	0.129	0.243	-	-
Polyn. model	age	0.122	0.220	-	-
Median model	age	-	-	0.068	0.114
	age, resec., inst.	-	-	0.099	0.121
	age, inst.	-	-	0.072	0.114

secondarily by the metrics denoted in Table 2. Accuracy is defined as classifying patients correctly in one of three survival time bins. Three bins are defined as short-term survivors with a remaining survival time of fewer than ten months, mid-term survivors with a remaining survival time between ten and 15 months and long-term survivors with more than 15 months of remaining survival time. A glance at the scatterplot 1 reveals that these bins cannot be derived intuitively from the data and the accuracy-based challenge scoring might potentially lead to the paradox situation where a better fitting model performs worse in the classification-based challenge.

On the validation set, we find that the quantile models using only age as predictors achieve the best accuracy (0.552). We attribute this to the median models' decreased susceptibility to outliers, especially given the low number of patients in the training and validation dataset. However, the metrics for the survival time predictions in days are differing, for example, the polynomial model using age only as a predictor has the lowest mean squared error and the Spearman R is identical for five different solutions, see Table 2.

Given that we achieved a much better fit on the GTR subset for the training set and because features other than age fail to add predictive value reliably, we selected a median model trained solely on the GTR subset and taking only age as an input for evaluation on the test set. A positive side effect of this approach is the simple deployment in clinical and scientific practice. The predictions of this median model are illustrated in scatterplot 1.

Comparison to other challenge participants. During the course of the challenge, we also compared our best performing model to the other participants on the validation set, knowing that most participating teams also consider image information. We monitored the leader board during the validation phase and

Table 2. Result table for the performance of our models on the validation set. Scores as calculated in the BRATS survival challenge leader board. Here the features used are encoded as age, resection status (resec.) and institution (inst.). The evaluation metrics for each submission (Subm.) are the accuracy, the mean squared error (MSE), median squared error (medianSE), standard squared error (stdSE) and SpearmanR.

Model		Accuracy	MSE	medianSE	stdSE	SpearmanR
OLS model	age; GTR only	0.448	90127.4	36773.6	123765.8	0.265
	age, inst.; GTR only	0.345	111571.2	40332.8	175070.8	0.165
	age, inst., resec.; all	0.310	105081.9	35523.3	161929.7	0.155
Polyn. model	age ; all	0.448	90383.3	30953.2	131065.2	0.265
	age ; GTR only	0.448	88113.3	32745.4	136508.8	0.265
Median model	age; all	**0.552**	101877.8	26958.2	116475.5	0.265
	age; GTR only	**0.552**	93572.3	30927.6	139847.1	0.265
	age, inst., resec.; all	0.483	96845.3	44466.3	155227.5	0.263
	age, inst.; GTR only	0.276	118450.0	54195.3	188132.4	0.184

found that our approach with an accuracy of 0.552 is within the best third of submissions. When comparing the metrics for fitting days of survival time, e.g. MSE, to the other submissions with equal accuracy, we found that our model shows solid performance. Overall, the total accuracy of our survival time predictions, but also of the best performing survival prediction, leaves much room for improvement. Even the best performing algorithms fail in more than one third of predictions. For perspective it is interesting to consider that last year the top performing algorithm used a U-Net to extract advanced image features. This shows that even the state-of-the-art in machine learning applied to this problem does not achieve a reliable survival prediction [4].

3.3 Evaluation on the test set

Finally, we consider our scores on the test set of 107 patients. We find a slight drop in performance with an accuracy of 0.486, a MSE of 419660.8, a medianSE 53177.5, a stdSE 1255102.9 and a SpearmanR of 0.358. Accuracy and medianSE are similar to our performance on the validation set. While the accuracy and medianSE scores remain comparable to the performance on the training and validation set, the outlier sensitive MSE and stdSE are substantially worse. This suggests that our drop in performance is mostly driven by statistical outliers.

4 Conclusion

We implemented simple OLS models, polynomial models and median regression models and experimented with different metadata-based predictor variables intentionally disregarding all image features. A simple median regression using only patient age as an input performed best to predict survival time for glioblastoma patients with gross total resection. Our model can serve as a baseline to evaluate the predictive value of sophisticated features.

5 Acknowledgments

Bjoern Menze, Benedikt Wiestler and Florian Kofler are supported through the SFB 824, subproject B12.
Supported by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81.
With the support of the Technical University of Munich Institute for Advanced Study, funded by the German Excellence Initiative.
Johannes C. Paetzold is supported by the Graduate School of Bioengineering, Technical University of Munich.
Daniel Krahulec is supported by MR R&D Clinical Science, Philips Healthcare.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive (2017) (2017)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. The Cancer Imaging Archive **286** (2017)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**, 170117 (2017)
4. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
5. Christakis, N.A., Smith, J.L., Parkes, C.M., Lamont, E.B.: Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort studycommentary: Why do doctors overestimate? commentary: Prognoses should be based on proved indices not intuition. Bmj **320**(7233), 469–473 (2000)
6. Everitt, B.: Book reviews : Chambers jm, hastie tj eds 1992: Statisti cal models in s. california: Wadsworth and brooks/cole. isbn 0 534 16765-9. Statistical Methods in Medical Research **1**(2), 220–221 (1992). https://doi.org/10.1177/096228029200100208
7. Glare, P., Virik, K., Jones, M., Hudson, M., Eychmuller, S., Simes, J., Christakis, N.: A systematic review of physicians' survival predictions in terminally ill cancer patients. Bmj **327**(7408), 195 (2003)

8. Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., Nakayama, H.: Gan-based synthetic brain mr image generation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 734–738. IEEE (2018)
9. Henderson, R., Keiding, N.: Individual survival time prediction using statistical models. Journal of Medical Ethics **31**(12), 703–706 (2005). https://doi.org/10.1136/jme.2005.012427, `https://jme.bmj.com/content/31/12/703`
10. Kennedy, W.J., Gentle, J.E.: Statistical computing. Routledge (2018)
11. Kim, D.W., Lee, S., Kwon, S., Nam, W., Cha, I.H., Kim, H.J.: Deep learning-based survival prediction of oral cancer patients. Scientific reports **9**(1), 6994 (2019)
12. Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica: journal of the Econometric Society pp. 33–50 (1978)
13. Koenker, R., Machado, J.A.F.: Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association **94**(448), 1296–1310 (1999). https://doi.org/10.1080/01621459.1999.10473882
14. Lao, J., Chen, Y., Li, Z.C., Li, Q., Zhang, J., Liu, J., Zhai, G.: A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Scientific reports **7**(1), 10353 (2017)
15. Li, H., Paetzold, J.C., Sekuboyina, A., Kofler, F., Zhang, J., Kirschke, J.S., Wiestler, B., Menze, B.: Diamondgan: Unified multi-modal generative adversarial networks for mri sequences synthesis. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 795–803. Springer International Publishing, Cham (2019)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, ., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (Oct 2015). https://doi.org/10.1109/TMI.2014.2377694
17. Ohgaki, H., Kleihues, P.: Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. Journal of Neuropathology & Experimental Neurology **64**(6), 479–489 (2005)
18. Parkes, C.M.: Commentary: prognoses should be based on proved indices not intuition. British Medical Journal **320**, 473–473 (2000)
19. Steinhauser, K.E., Clipp, E.C., McNeilly, M., Christakis, N.A., McIntyre, L.M., Tulsky, J.A.: In search of a good death: observations of patients, families, and providers. Annals of internal medicine **132**(10), 825–832 (2000)
20. Weninger, L., Haarburger, C., Merhof, D.: Robustness of radiomics for survival prediction of brain tumor patients depending on resection status. Frontiers in Computational Neuroscience **13**, 73 (2019)

21. Weninger, L., Rippel, O., Koppers, S., Merhof, D.: Segmentation of brain tumors and patient survival prediction: Methods for the brats 2018 challenge. In: International MICCAI Brainlesion Workshop. pp. 3–12. Springer (2018)
22. Yang, K., Nath, S., Koziarz, A., Badhiwala, J.H., Ghayur, H., Sourour, M., Catana, D., Nassiri, F., Alotaibi, M.B., Kameda-Smith, M., et al.: Biopsy versus subtotal versus gross total resection in patients with low-grade glioma: A systematic review and meta-analysis. World neurosurgery **120**, e762–e775 (2018)
23. Youngner, S.J., Arnold, R.M.: The Oxford handbook of ethics at the end of life. Oxford University Press (2016)

Part V

CONCLUDING REMARKS AND OUTLOOK

12 CONCLUDING REMARKS AND OUTLOOK

This *Hitchhiker's Guide* meant to provide a gentle introduction for diving into the topic of *Machine Learning for Biomedical Image Analysis*. Even though the *guide's* focus is on semantic segmentation, readers should find that many of the presented concepts easily translate to other ML problems. The *guide* covers the whole workflow from dataset curation, model training, evaluation, and interpretation of results to refining the model. It is important to realize this not as a single *waterfall process*, but as an *iterative loop*.

Even though CNNs are usually trained with *gradient descent*, it does not mean that all training parameters need to be explored in a pure *trial and error* fashion. Therefore, heuristics and pointers to derive design decisions in a *theory-driven* way are provided.

Hence, the first (i) part deals with identifying a good set of hyperparameters for model training. Besides general aspects, it covers particularities when dealing with MR and CT image data.

The second (ii) part covers the interpretation of network outputs. It reveals that a pure analysis of similarity metrics is not sufficient due to the limitations of human annotations. To circumnavigate this, it introduces methods to collect feedback from clinical practitioners.

Finally, the third (iii) part comprises strategies for optimizing model performance. Besides curating a better training set, researchers are advised to consider ensembling and to derive informed decisions regarding the network architecture and loss function.

Outlook: In the (biomedical) ML community, it is common practice to market a technical innovation by claiming an improvement in segmentation quality by demonstrating a small improvement in segmentation quality metrics such as DSC. Over the years, many segmentation challenges, such as BraTS, LiTS, KiTS, etc., emerged that decorate winners based on tiny improvements in similarity metrics with human annotations. It remains an open research question whether such improvements translate to real-world benefits for the application of ML in clinical workflows. As the radiologists in our experiments prefer network-generated over human-annotated segmentations, the presented findings indicate that this assumption might not hold up. For a successful translation of ML algorithms towards clinical practice, it is imperative to gain physicians' trust in the model predictions. Therefore, it seems necessary to iteratively involve physicians in the model training process. There is hope that the findings presented will stimulate debate and that researchers begin to question these established practices in *(biomedical)* ML.

BIBLIOGRAPHY

[1] Brian B Avants, Nick Tustison, Gang Song, et al. "Advanced normalization tools (ANTS)." In: *Insight j* 2.365 (2009), pp. 1–35.

[2] S Bakas, H Akbari, A Sotiras, M Bilello, M Rozycki, J Kirby, J Freymann, K Farahani, and C Davatzikos. *Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017)*. 2017.

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features." In: *Scientific data* 4 (2017), p. 170117.

[4] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection." In: *The Cancer Imaging Archive* 286 (2017).

[5] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge." In: *arXiv preprint arXiv:1811.02629* (2018).

[6] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. "VoxelMorph: a learning framework for deformable medical image registration." In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1788–1800.

[7] Yoshua Bengio. "Practical recommendations for gradient-based training of deep architectures." In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 437–478.

[8] David H Brainard. "The psychophysics toolbox." In: ().

[9] Kathryn Mary Broadhouse. "The physics of MRI and how we use it to reveal the mysteries of the mind." In: *Front. Young Minds* 7 (2019), p. 23.

[10] Joshua R De Leeuw and Benjamin A Motz. "Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task." In: *Behavior Research Methods* 48.1 (2016), pp. 1–12.

[11] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. "Neural Architecture Search: A Survey." In: (2018). DOI: 10.48550/ARXIV.1808.05377. URL: https://arxiv.org/abs/1808.05377.

[12] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. "Deep ensembles: A loss landscape perspective." In: *arXiv preprint arXiv:1912.02757* (2019).

[13] KM Harris, H Adams, DCF Lloyd, and DJ Harvey. "The effect on apparent size of simulated pulmonary nodules of using three standard CT window settings." In: *Clinical radiology* 47.4 (1993), pp. 241–244.

[14] Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P. Prabhu, Simon K. Warfield, and Ali Gholipour. "Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection." In: *IEEE Access* 7 (2019), pp. 1721–1735. DOI: 10.1109/access.2018.2886371. URL: https://doi.org/10.1109%2Faccess.2018.2886371.

[15] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. "Robust brain extraction across datasets and comparison with publicly available methods." In: *IEEE transactions on medical imaging* 30.9 (2011), pp. 1617–1634.

[16] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." In: *Nature methods* 18.2 (2021), pp. 203–211.

[17] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. "Automated brain extraction of multisequence MRI using artificial neural networks." In: *Human brain mapping* 40.17 (2019), pp. 4952–4964.

[18] Doris Kaltenecker, Rami Al-Maskari, Moritz Negwer, Luciano Hoeher, Florian Kofler, Shan Zhao, Mihail Todorov Todorov, Johannes Christian Paetzold, Benedikt Wiestler, Julia Geppert, et al. "Virtual reality empowered deep learning analysis of brain activity." In: *bioRxiv* (2023), pp. 2023–05.

[19] Florian Kofler, Christoph Berger, Diana Waldmannstetter, Jana Lipkova, Ivan Ezhov, Giles Tetteh, Jan Kirschke, Claus Zimmer, Benedikt Wiestler, and Bjoern H Menze. "BraTS Toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice." In: *Frontiers in neuroscience* (2020), p. 125.

[20] Florian Kofler, Ivan Ezhov, Lucas Fidon, Carolin M Pirkl, Johannes C Paetzold, Egon Burian, Sarthak Pati, Malek El Husseini, Fernando Navarro, Suprosanna Shit, et al. "Robust, Prim-

itive, and Unsupervised Quality Estimation for Segmentation Ensembles.” In: *Frontiers in Neuroscience* 15 (2021).

[21] Florian Kofler, Ivan Ezhov, Lucas Fidon, Carolin M Pirkl, Johannes C Paetzold, Egon Burian, Sarthak Pati, Malek El Husseini, Fernando Navarro, Suprosanna Shit, et al. “Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles.” In: *Frontiers in Neuroscience* 15 (2021).

[22] Florian Kofler, Ivan Ezhov, Fabian Isensee, Fabian Balsiger, Christoph Berger, Maximilian Koerner, Johannes Paetzold, Hongwei Li, Suprosanna Shit, Richard McKinley, et al. “Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient.” In: *arXiv preprint arXiv:2103.06205* (2021).

[23] Florian Kofler, Johannes C Paetzold, Ivan Ezhov, Suprosanna Shit, Daniel Krahulec, Jan S Kirschke, Claus Zimmer, Benedikt Wiestler, and Bjoern H Menze. “A baseline for predicting glioblastoma patient survival time with classical statistical models and primitive features ignoring image information.” In: *International MICCAI Brainlesion Workshop*. Springer. 2019, pp. 254–261.

[24] Florian Kofler et al. *blob loss: instance imbalance aware loss functions for semantic segmentation*. 2022. DOI: `10.48550/ARXIV.2205.08209`. URL: `https://arxiv.org/abs/2205.08209`.

[25] Florian Kofler et al. *Approaching Peak Ground Truth*. 2023. arXiv: `2301.00243 [cs.LG]`.

[26] Kristian Lange, Simone Kühn, and Elisa Filevich. “" Just Another Tool for Online Studies”(JATOS): An easy solution for setup and management of web servers supporting online studies.” In: *PloS one* 10.6 (2015), e0130834.

[27] Thomas Robin Langerak, Uulke A van der Heide, Alexis NTJ Kotte, Max A Viergever, Marco Van Vulpen, and Josien PW Pluim. “Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE).” In: *IEEE transactions on medical imaging* 29.12 (2010), pp. 2000–2008.

[28] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. “Loss odyssey in medical image segmentation.” In: *Medical Image Analysis* 71 (2021), p. 102035.

[29] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. “Why rankings of biomedical image analysis competitions should be interpreted with care.” In: *Nature communications* 9.1 (2018), pp. 1–13.

[30] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. “OpenSesame: An open-source, graphical experiment builder for the social sciences.” In: *Behavior research methods* 44.2 (2012), pp. 314–324.

[31] B. H. Menze et al. “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS).” In: *IEEE Transactions on Medical Imaging* 34.10 (2015), pp. 1993–2024. ISSN: 0278-0062. DOI: 10.1109/TMI.2014.2377694.

[32] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy.” In: *arXiv preprint arXiv:1809.04430* (2018).

[33] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. “TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning.” In: *Computer Methods and Programs in Biomedicine* (2021), p. 106236. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2021.106236. URL: https://www.sciencedirect.com/science/article/pii/S0169260721003102.

[34] Annika Reinke, Matthias Eisenmann, Minu D Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, et al. “Common limitations of image processing metrics: A picture story.” In: *arXiv preprint arXiv:2104.05642* (2021).

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[36] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. “Nonrigid registration using free-form deformations: application to breast MR images.” In: *IEEE transactions on medical imaging* 18.8 (1999), pp. 712–721.

[37] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. “Tversky loss function for image segmentation using 3D fully convolutional deep networks.” In: *International workshop on machine learning in medical imaging*. Springer. 2017, pp. 379–387.

[38] Euclid Seeram. *Computed Tomography-E-Book: Physical Principles, Clinical Applications, and Quality Control*. Elsevier Health Sciences, 2015.

[39] Euclid Seeram. “Computed tomography: A technical review.” In: *Radiologic technology* 89.3 (2018), 279CT–302CT.

[40] Boris Shirokikh, Alexey Shevtsov, Anvar Kurmukov, Alexandra Dalechina, Egor Krivov, Valery Kostjuchenko, Andrey Golanov, and Mikhail Belyaev. "Universal loss reweighting to balance lesion size inequality in 3d medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 523–532.

[41] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. "clDice-a novel topology-preserving loss function for tubular structure segmentation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16560–16569.

[42] Gijsbert Stoet. "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments." In: *Teaching of Psychology* 44.1 (2017), pp. 24–31.

[43] Carole H. Sudre et al. *Where is VALDO? VAscular Lesions Detection and segmentatiOn challenge at MICCAI 2021*. 2022. arXiv: 2208.07167 [cs.CV].

[44] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool." In: *BMC medical imaging* 15.1 (2015), pp. 1–28.

[45] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. "N4ITK: improved N3 bias correction." In: *IEEE transactions on medical imaging* 29.6 (2010), pp. 1310–1320.

[46] Paul A Yushkevich, John Pluta, Hongzhi Wang, Laura EM Wisse, Sandhitsu Das, and David Wolk. "IC-P-174: Fast Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe Subregions In 3 Tesla and 7 Tesla T2-Weighted MRI." In: *Alzheimer's & Dementia* 12 (2016), P126–P127.

[47] Hang Zhang, Jinwei Zhang, Chao Li, Elizabeth M Sweeney, Pascal Spincemaille, Thanh D Nguyen, Susan A Gauthier, Yi Wang, and Melanie Marcille. "ALL-Net: Anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation." In: *NeuroImage: Clinical* 32 (2021), p. 102854.

[48] Min Zhou, Chaoshi Niu, Li Jia, and Hu He. "The value of MGMT promote methylation and IDH-1 mutation on diagnosis of pseudoprogression in patients with high-grade glioma: a meta-analysis." In: *Medicine* 98.50 (2019).