Aus dem Deutschen Krebsforschungszentrum
(Vorstand: Prof. Dr. med. Michael Baumann, Ursula Weyrich)
Abteilung für Medizinische Bildverarbeitung
(Leiter: Prof. Dr. rer. nat. Klaus H. Maier-Hein)

Robustness of Medical Image Segmentation Algorithms in the Context of Federated Data

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Maximilian Armin Zenk
aus
Bad Soden-Salmünster

2025

Dekan: Prof. Dr. med. Michael Boutros
Doktorvater: Prof. Dr. rer. nat. Klaus H. Maier-Hein

Contents

Acronyms **i**

List of Figures **iii**

List of Tables **vii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Background . . . 4
1.2.1 Medical Image Segmentation . . . 4
1.2.2 Distribution Shifts and Medical Imaging . . . 9
1.2.3 Predictive Uncertainty Estimation . . . 11
1.2.4 Medical Image Analysis Competitions . . . 14
1.3 Related Work . . . 16
1.3.1 Generalization . . . 17
1.3.2 Failure Detection . . . 21
1.4 Objectives and Contributions . . . 27
1.4.1 Generalization . . . 27
1.4.2 Failure Detection . . . 28
1.5 Outline . . . 29

2 Materials and Methods **31**
2.1 Generalization . . . 31
2.1.1 Challenge Organization . . . 32
2.1.2 Datasets . . . 32
2.1.3 Annotation Quality Control . . . 39
2.1.4 Performance Assessment Methods . . . 39
2.1.5 Statistical Analyses . . . 41
2.1.6 Technical Infrastructure . . . 41

2.2 Failure Detection . . . 44
2.2.1 Task Definition . . . 44
2.2.2 Evaluation . . . 45
2.2.3 Datasets . . . 50
2.2.4 Segmentation Algorithm . . . 55
2.2.5 Failure Detection Methods . . . 58

3 Results **67**
3.1 Generalization . . . 67
3.1.1 Results of the FeTS challenge 2021 (Pilot Study) . . . 68
3.1.2 Results of the FeTS challenge 2022 . . . 72
3.2 Failure Detection . . . 88
3.2.1 Failure Detection Benchmark Results . . . 88
3.2.2 Extensions of the Benchmark . . . 99
3.2.3 Analysis of the Evaluation Protocol . . . 104

4 Discussion **109**
4.1 Generalization . . . 109
4.1.1 Interpretation of the Challenge Results . . . 110
4.1.2 Comparison to Related Work . . . 112
4.1.3 Limitations and Future Work . . . 114
4.2 Failure Detection . . . 115
4.2.1 Interpretation of the Benchmark Results . . . 115
4.2.2 Comparison to Related Work . . . 117
4.2.3 Limitations and Future Work . . . 119
4.3 Overall Conclusions . . . 121

5 Summary **123**

6 Zusammenfassung **125**

Bibliography **129**

Own Contributions and Publications **169**
Own share in data acquisition and data analysis . . . 169
Own Publications . . . 171

A Appendix **175**
A.1 Additional Results . . . 175
A.1.1 Generalization . . . 175

A.1.2 Failure Detection 175
A.2 Additional Image Samples from the Failure Detection Benchmark 175
A.3 Details on the FeTS Challenge Submissions (Algorithm Characteristics) . 189
A.4 Dokumentation der verwendeten KI-Hilfsmittel 191

Acknowledgments **193**

Eidesstattliche Versicherung **195**

Angaben zu verwendeten KI-basierten Elektronischen Hilfsmitteln **197**

Acronyms

MRI	magnetic resonance imaging
CT	computed tomography
T1	native T1-weighted
T2	T2-weighted
T1-Gd	contrast-enhanced T1-weighted
FLAIR	Fluid Attenuated Inversion Recovery
ET	enhancing tumor
TC	tumor core
WT	whole tumor
ED	edema
NCR	necrotic core/necrocyst
FeTS	Federated Tumor Segmentation
BraTS	Brain Tumor Segmentation
LGG	low-grade glioma
HGG	high-grade glioma
OCT	optical coherence tomography
US	ultrasound
DSC	Dice similarity coefficient
HD	Hausdorff distance
NSD	normalized surface dice
MAE	mean absolute error
AUROC	area under the receiver operating curve
AUROC$_f$	area under the receiver operating curve using binary failure labels

AURC	area under the risk-coverage curve
PC	Pearson correlation coefficient
SC	Spearman correlation coefficient
ReLU	rectified linear unit
BN	batch normalization
GPU	graphics processing unit
CE	cross-entropy
IQR	interquartile range
OOD	out-of-distribution
i.i.d.	independent and identically distributed
RQ	research question
MIC	Medical Image Computing
DKFZ	German Cancer Research Center
MICCAI	medical image computing and computer assisted intervention
CSF	confidence scoring function
CNN	convolutional neural network
NN	neural network
DNN	deep neural network
AI	artificial intelligence
DL	deep learning
SGD	stochastic gradient descent
PE	predictive entropy
MI	mutual information
RF	regression forest
ROI	region of interest
ML	Machine Learning
VAE	Variational autoencoder
MC	Monte Carlo
FD	failure detection
IDH	isocitrate dehydrogenase

List of Figures

1.1 Examples of segmentation algorithm predictions on brain tumor data from several origins. . . . 3
1.2 Example for semantic segmentation in medical images. . . . 5
1.3 Architecture visualization of the U-Net . . . 7
1.4 Probabilistic model and examples for distribution shifts encountered in medical image segmentation. . . . 11
1.5 Illustration of two sources of uncertainty in model predictions in a two-dimensional, artificial example . . . 13

2.1 Partitioning of training and test sets by geographically distinct institutions. 34
2.2 Geographical distribution of the participants, as well as training and testing institutions in the FeTS22 challenge. . . . 35
2.3 Patient population in terms of sex, IDH status, and age across 32 institutions that contributed test datasets to FeTS22. . . . 36
2.4 Training case example from the FeTS Challenge 2022. . . . 38
2.5 Diagram of the federated evaluation workflow used in the FeTS Challenge. 43
2.6 Example for the risk-coverage curve based on synthetic data. . . . 49
2.7 Examples from the CT and MRI datasets used for the benchmark. . . . 54
2.8 Examples from the non-CT/MRI datasets used for the benchmark. . . . 55
2.9 Schematic of the segmentation network backbone (U-Net). . . . 57
2.10 Overview of the components involved in failure detection. . . . 58

3.1 Distribution of metric values for each institution and participating team of the FeTS21 challenge. . . . 70
3.2 Ranking statistics for the FeTS21 challenge. . . . 71
3.3 Aggregated challenge results of the FeTS22 challenge. . . . 76
3.4 Detailed results of the FeTS22 challenge of the top-ranked submissions. . . 77
3.5 Examples of common segmentation errors in the FeTS Challenge. . . . 80

3.6 2D-Histograms of true region size versus segmentation metrics for all test cases in the FeTS22 challenge. 81
3.7 Ranking stability for each region and metric evaluated in the FeTS22 challenge. 82
3.8 Test set segmentation performance measured by DSC, of a single U-Net . 89
3.9 Performance difference (DSC) between prediction models on the test set. . 90
3.10 Comparison of aggregation methods in terms of AURC scores for all datasets. 91
3.11 Ranking distribution obtained through bootstrapping for aggregation methods. 92
3.12 Comparison of image-level failure detection methods in terms of AURC scores for all datasets 95
3.13 Scatter plot of confidence scores produced by mean pairwise DSC versus true DSC scores. 96
3.14 Qualitative analysis of ensemble predictions on all datasets. 98
3.15 Test set segmentation performance measured by DSC of a single U-Net on non-CT/MRI datasets. 100
3.16 Experimental results for varying number of training samples in the heart dataset. 103
3.17 Ranking distribution plots based on 1000 bootstrap samples, compared between different risk functions. 104
3.18 Example illustrating why AURC is most suitable as a failure detection metric 106
3.19 Comparison of OOD-AUROC scores for different failure detection methods 108

A.1 Aggregated challenge results of the FeTS22 challenge for each evaluated model and institution 176
A.2 Case-level challenge results of the top-ranked algorithm for each institution of the test set 177
A.3 Effect of annotation quality control on rankings. 178
A.4 Effect of annotation quality control on the mean Dice similarity coefficient (DSC) distributions for the best-performing model. 178
A.5 AURC scores for all datasets and methods in the failure detection benchmark. 179
A.6 Pearson correlation coefficient for all datasets and methods in the failure detection benchmark. 181
A.7 Spearman correlation coefficient for all datasets and methods in the failure detection benchmark. 182
A.8 Samples from the test set of the 2D brain (toy) dataset. 183
A.9 Samples from the test set of the brain tumor dataset. 184
A.10 Samples from the test set of the heart dataset. 185
A.11 Samples from the test set of the kidney tumor dataset. 186

A.12 Samples from the test set of the prostate dataset. 187
A.13 Samples from the test set of the Covid dataset. 188

List of Tables

1.1 Comparison of related work for evaluating generalization on unseen institutions. . . . 20
1.2 Comparison of related studies with benchmarking character. . . . 26

2.1 Statistics of the training, validation and test cases for the FeTS challenges . 33
2.2 Comparison of geographical diversity between the test sets of the BraTS 2021 and FeTS challenges. . . . 35
2.3 Metric candidates for segmentation failure detection. . . . 48
2.4 Summary of datasets used in the failure detection benchmark. . . . 51
2.5 Hyperparameters for the segmentation and failure detection methods used in the failure detection benchmark. . . . 59
2.6 Overview of failure detection methods included in the benchmark 65

3.1 Extended ranking and algorithm characteristics of all models evaluated in the FeTS Challenge 2022 83
3.2 AURC scores on the test sets for different pixel-level uncertainty measures and aggregation methods 93
3.3 Mean AURC scores on the test sets for different failure detection methods. 94
3.4 AURC scores on the test sets of the non-CT/MRI datasets for confidence aggregation methods. . . . 100
3.5 Comparison of the AURC scores of failure detection methods with different segmentation backbones. . . . 101
3.6 Multi-metric comparison of failure detection methods. . . . 107

A.1 AURC scores (×100) on the test sets for all compared failure detection methods 180
A.2 Mapping from model ID to scientific publication for the subset of BraTS 2021 models evaluated within FeTS22 190

1 Introduction

1.1 Motivation

Imaging plays an important role in today's medicine, as it can provide rich information about the patient with noninvasive technology. Consequently, the demand for radiological examinations is rising, with the combined number of computed tomography (CT) and magnetic resonance imaging (MRI) scans acquired per year increasing by 29.8 % in England within the past five years (NHS 2023). Data from the Royal College of Radiologists for the United Kingdom suggests that growth in the workforce cannot keep pace, estimating that 1962 additional clinical radiologists would have been required in 2023, corresponding to a gap of 30 % (RCR 2023). This development does not only lead to high workloads and stress for radiologists, but it can ultimately also cause harm to patients through delayed or inaccurate radiological readings (Alexander et al. 2022). Technological assistance through artificial intelligence (AI) tools is a promising strategy to alleviate the aforementioned challenges, as they can automate time-consuming or tedious tasks (Kalidindi and Gandhi 2023).

One such task is semantic segmentation, which is the focus of this thesis and consists in annotating for each pixel in an image whether it belongs to a region of interest, such as an organ or a lesion. The resulting segmentation masks are required as an intermediate step for many medical image analysis pipelines, enabling, for instance, volumetric assessment of tumor burden (Kickingereder et al. 2019), organs-at-risk delineation for radiotherapy planning (Nikolov et al. 2021), and extraction of imaging phenotypes for large-cohort association studies (Aerts et al. 2014; Bai et al. 2020). Unfortunately, manual segmentation is laborious and introduces variability between or within annotators (Harari et al. 2010; Nelms et al. 2012; Menze et al. 2015; Joskowicz et al. 2019). This motivates research in automatic segmentation algorithms, as they have the potential to improve the efficiency of the clinical workflow while also making the segmentations more reproducible (Kickingereder et al. 2019; Nikolov et al. 2021). Advancements in deep learning (DL) methodology (LeCun et al. 2015) and hardware accelerators have revolutionized the field of computer vision

and medical image analysis in the past decade. Segmentation algorithms were devised that learn to generate segmentations on unseen images, which can match the quality of human annotations if given enough example pairs of images and segmentation masks as a training dataset. State-of-the-art methods can even adapt automatically to diverse medical image segmentation tasks (Isensee et al. 2021a).

To make such models useful in clinical applications, they need to produce trustworthy results not only for research datasets but also when used at various hospitals not seen during model development. However, it is well known that the performance of DL models can decrease when applied to data with *distribution shifts* (AlBadawy et al. 2018; Zech et al. 2018; Badgeley et al. 2019; Beede et al. 2020; Campello et al. 2021), which means that characteristics of the data distribution differ between the model training and deployment stage. In the context of medical image segmentation, distribution shifts are expected when using data from hospitals not seen during training, for instance due to different scanners or patient populations (Castro et al. 2020). Examples for the task of brain tumor segmentation are shown in fig. 1.1. Including training data from every environment at which the model should be deployed is not practical in medical image analysis, because clinical data usually cannot be shared outside the institution where they were acquired, due to privacy concerns and data protection regulations. Hence, the amount of accessible multicentric data is still small for many segmentation tasks and additional methods that improve the robustness of segmentation models are necessary to make them reliable in real clinical applications.

This thesis investigates two complementary approaches to the robustness problem and addresses existing gaps in research for each of them from a comparative, benchmarking perspective:

1. **Generalization:** The objective of this approach is to produce high-quality segmentations even on images with distribution shifts, thus preventing segmentation failures in the first place. Although various methods have been proposed towards this goal (Zhou et al. 2023; Yoon et al. 2024), systematic evaluation efforts are limited. The standard for benchmarking medical image analysis methods are public biomedical competitions (Maier-Hein et al. 2018), as they provide fair and reproducible conditions for comparing algorithms. Such competitions have only recently begun to address generalization to distribution shifts between training and test data on a small scale, by collecting multicentric data and setting aside data from a few institutions for testing (Campello et al. 2021; Aubreville et al. 2023; Payette et al. 2024). However, it remains an open question how well state-of-the-art segmentation algorithms generalize when tested on large-scale data originating from many, geographically distributed institutions unseen during training, with a diversity close to applications "in the wild".

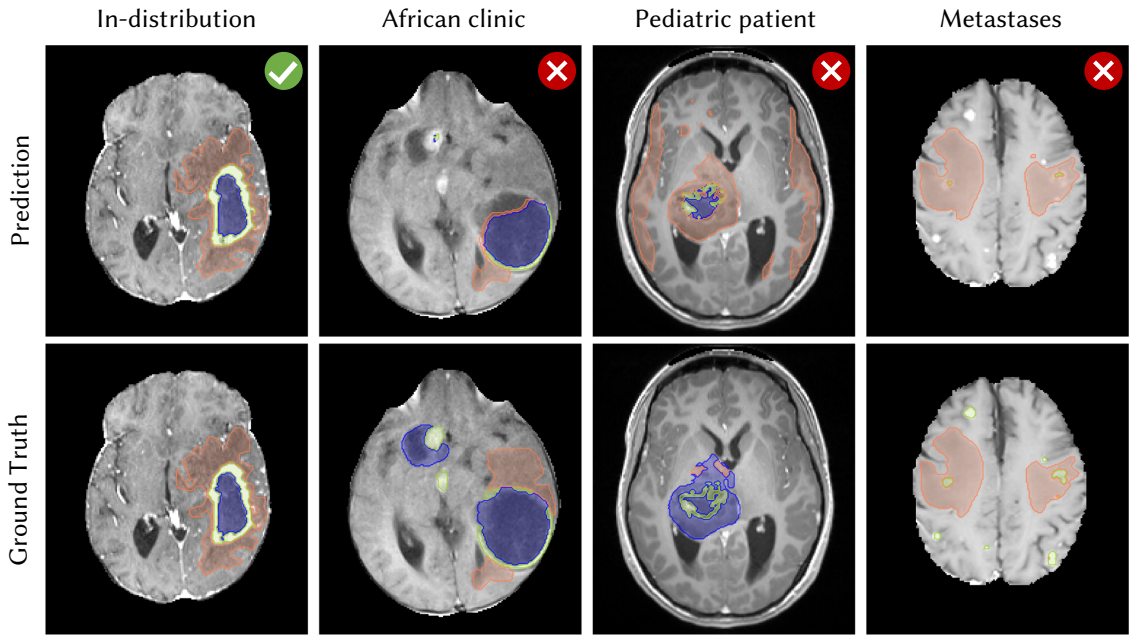

Figure 1.1: Examples of segmentation algorithm predictions (edema = orange, enhancing tumor = light green, non-enhancing tumor = blue) on data from several origins. Only the contrast-enhanced T1 MRI sequence is shown. A nnU-Net (Isensee et al. 2021a) trained on one fold of a multi-centric dataset with brain tumor (glioma) patients performed well on in-distribution data (adults mostly from North America and Europe). For other clinics, populations, or tumor types, however, failures can be observed, indicating a lack of robustness. Obvious errors were selected here for illustration purposes, but less severe errors that occur more frequently in practice can also cause problems in real-world applications. Images are from the BraTS 2021 and 2024 Challenge datasets (Baid et al. 2021; Adewole et al. 2023; Karargyris et al. 2023; Kazerooni et al. 2024; Moawad et al. 2024).

2. **Failure detection:** As errors cannot always be avoided, failure detection aims to automatically detect when segmentations might be inaccurate, so that they can be filtered or manually corrected before affecting downstream image analysis tasks. Three lines of research state failure detection as a main motivation: uncertainty estimation aims to complement the model's prediction with a general-purpose confidence score, usually on the pixel-level (for example Mehrtash et al. 2020). Out-of-distribution detection methods attempt to identify samples that are dissimilar to the training data, as these are likely to cause prediction errors (for example González et al. 2022). Lastly, segmentation quality estimation is an approach that predicts segmentation metrics values for a segmentation model's output without access to the ground truth (for example Valindria et al. 2017). Although failure detection is a common down-

stream task for all of these strategies, they are usually studied separately, since there are no standardized evaluation protocols compatible with all approaches (Jaeger et al. 2022). Furthermore, as previous works are limited to a single segmentation task or do not consider distribution shifts, they cannot answer how self-configuring segmentation methods like nnU-Net can profit from failure detection. Overall, there is currently no clear picture of which segmentation failure detection methods work reliably when applied to different datasets with distribution shifts.

1.2 Background

1.2.1 Medical Image Segmentation

Image segmentation algorithms partition the image into multiple regions that have some common characteristics. This thesis focuses on semantic segmentation (henceforth just called segmentation), the task of assigning a class label to each pixel in the image, which captures the meaning of each pixel in the context of a particular image analysis task. In medical images, for instance, the goal is often to annotate which parts of the image belong to different organs or pathologies, so possible classes for a segmentation task in the head could be "skull", "hippocampus", "tumor", "background". An example for brain tumor segmentation in MRI is shown in fig. 1.2. The segmented region of interests (ROIs) can be used for research on imaging biomarkers (Aerts et al. 2014), volumetric quantification of tumor burden (Kickingereder et al. 2019) or radiotherapy treatment planning (Nikolov et al. 2021), to name a few examples. As annotating images pixel-by-pixel* is a time-consuming effort, algorithms for automatic segmentation have been studied extensively. In the following, a brief overview of traditional segmentation approaches is given, after which the current state of the art based on deep learning algorithms is described. The section ends with a description of how segmentation algorithms are evaluated.

1.2.1.1 Segmentation Methods without Deep Learning

Early approaches applied thresholds on the image intensities, detected edges, or reformulated segmentation as a graph cut problem to segment images (Toennies 2017). While such techniques can work for medical applications in which image intensities are reliable predictors of the anatomical target structure, they have limitations in the presence of noise or for more complex segmentation tasks.

*Radiological imaging modalities like CT or MRI are inherently three-dimensional volumes consisting of voxels. For simplicity, in this section the basic elements of images are just called pixels, using the term voxels only if the 3D nature should be emphasized.

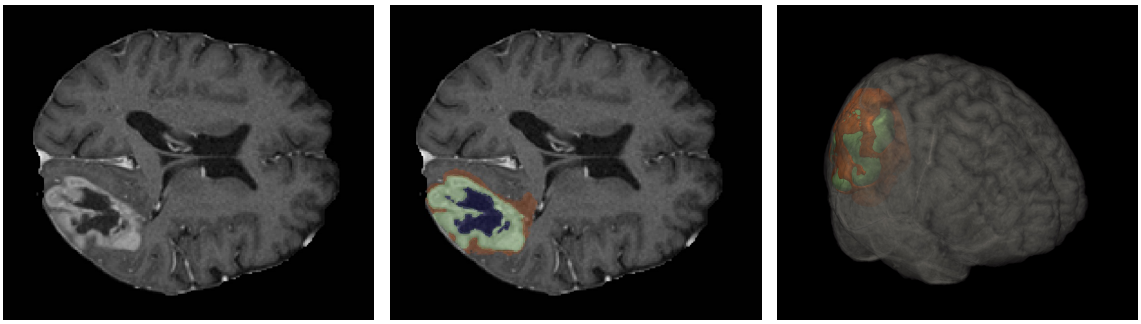

Figure 1.2: Example for semantic segmentation in medical images. One slice of a T1-weighted MRI sequence of the brain acquired with contrast agent is shown (left), next to the overlaid ground truth segmentation mask (middle) and a 3D rendering (right). Each pixel is assigned a segmentation class (indicated by a color overlay): peritumoral edema (orange), enhancing tumor (light green), necrosis (dark blue) and background (no color). The image originates from the training data published in (Zenk et al. 2025a).

A more modern approach is based on example images for which a manual reference segmentation was generated, so-called "atlases". As medical images usually depict similar anatomies, two images from different subjects can often be registered, which means that a spatial transform can be computed that aligns the anatomies of both subjects as well as possible while ensuring the transformation remains invertible. This technique can be used for medical image segmentation: To generate a segmentation of a new image, it is registered to a set of atlases and the atlases' annotations are transferred back to the new image with the identified spatial mapping. If there are multiple atlases, there are different strategies to merge their labels (Iglesias and Sabuncu 2015).

Medical images can also be segmented using statistical shape models, which rely on prior knowledge of the shape and image appearance of organs or other target structures. After choosing a shape representation, such as a collection of landmarks, the shape model captures the mean shape and variation in a training dataset with manually annotated landmarks. To fit a shape model to a new image, two additional components are required: an appearance model that captures the texture of the target structure and a search algorithm that allows finding the correct location of the shape (Heimann and Meinzer 2009).

1.2.1.2 Segmentation Methods based on Deep Learning

Advancements in the field of deep learning (LeCun et al. 2015) during the last decade led deep neural networks (DNNs) to become state of the art in a wide range of applications, including protein structure prediction (Jumper et al. 2021), natural language processing

(Brown et al. 2020) and computer vision (Kirillov et al. 2023). Based on the availability of large training datasets and hardware accelerators, most notably graphics processing units (GPUs), these models are able to learn which features to extract from the raw data during training to achieve low error in a supervised learning task.

Briefly summarizing LeCun et al. (2015), deep learning is based on artificial neural networks, which are usually organized into multiple layers of processing units (neurons). Each neuron receives input from neurons in the previous layer and applies a nonlinear function with tunable parameters. In the simplest case, each neuron computes a weighted sum of the outputs from the previous layer and feeds it through non-linearity like the sigmoid or rectified linear unit (ReLU) function, known as the activation function. If all operations in the network are differentiable, the backpropagation algorithm allows optimization of a task-specific objective function with respect to the network parameters on a given training set using stochastic gradient descent (SGD), enabling the network to learn a nonlinear relation between input and output. DNNs are extremely flexible in which building blocks to use and how to assemble them to a network, but for medical image analysis tasks, convolutional neural networks (CNNs) are currently the best-performing class (Isensee et al. 2024). Each layer in a CNN applies multiple learnable discrete convolutions to the output of the previous layer, which restricts the inputs of each neuron to a local neighborhood and makes neurons share parameters, thereby processing images efficiently.

The connections between neurons, and how information flows between them, determine the architecture of a neural network. For medical image segmentation, the U-Net architecture (Ronneberger et al. 2015) stands out, which was originally proposed for 2D images but quickly adapted to 3D images (Çiçek et al. 2016). Its layers are arranged in an encoder-decoder construct, with additional information flow between the two branches through skip connections (fig. 1.3). The encoder consists of multiple stages operating at successively decreasing spatial resolution. In each stage, convolutional layers extract and refine various image features, while the spatial dimension of the resulting feature maps is reduced with pooling layers before the next stage. The feature representation at the end of the encoder hence has low spatial resolution, but a high semantic depth per position. The decoder roughly mirrors the structure of the encoder, as it is designed to upsample the feature representation again, to convert it into a segmentation map. Transpose convolutions are used for learnable upsampling operations in the U-Net, helping to locate the learned information from the encoder in the original image space. Skip connections from each stage of the encoder to the corresponding stage of the decoder are added to recover finer information about the location of segments. Finally, convolutional layers are used in the decoder to combine and process the information received from the skip connections and upsampled feature maps.

State-of-the-art segmentation pipelines based on deep learning are complex algorithms

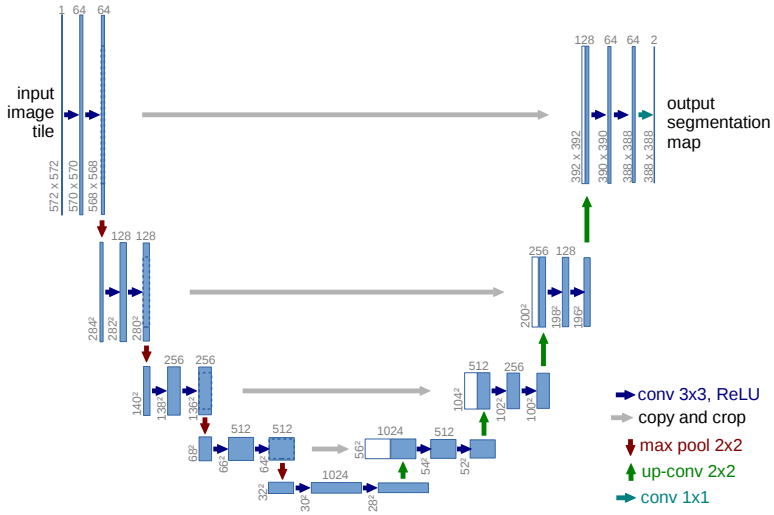

Figure 1.3: Architecture visualization of the U-Net (Ronneberger et al. 2015). Computational operations by the network are indicated by arrows, while feature representations are visualized as boxes with spatial dimensions annotated on the left and feature channel dimensions on the top. The kernel size is given for convolution operations (as in conv 3×3) and similarly the extent of local pooling operations. The decoder can produce a high-resolution segmentation map by aggregating the rich semantics from lower layers with additional localization information from the skip connection. Figure by Ronneberger et al. (2015) reproduced with permission from Springer Nature.

that depend not only on their network architecture but also on many other components, such as the loss function, optimizer, data loading pipeline, and pre-/post-processing steps. Configuring all hyperparameters associated with these components is a difficult task that needs to be performed for each application in which the segmentation network should be used. Isensee et al. (2021a) developed nnU-Net, an algorithm for medical image segmentation that automatically configures the segmentation algorithm based on dataset properties, which include the imaging modality as well as the distribution of voxel intensities and spacings. The resulting segmentation algorithm has three groups of parameters: Fixed parameters are dataset-agnostic, rule-based parameters are derived based on the dataset characteristics, and empirical parameters are set based on the results of different variants in cross-validation. The details of all configuration parameters are beyond the scope of this thesis and can be found in the original publication (Isensee et al. 2021a). One interesting finding of their study is, however, that a fixed, U-Net-like network architecture performs well on many datasets if the rest of the segmentation pipeline is configured appropriately. The automatic configuration by nnU-Net remains a strong baseline for medical image segmentation until today (Isensee et al. 2024) and is an ideal

foundation for developing new methods. As it can be applied to new datasets without user input, nnU-Net is also promising for studying generalization to distribution shifts in this thesis.

1.2.1.3 Evaluation of Segmentation Methods

Without a good measure of segmentation performance, developing a new method is impossible. Performance metrics for semantic segmentation compare two binary segmentation masks A and B, which can correspond to the reference segmentation and model prediction, for example. For multi-class segmentation problems, metrics are computed for each class separately and can be averaged to obtain an overall segmentation quality measure, if necessary. Following Reinke et al. (2021), segmentation metrics can be categorized into overlap-based or boundary-based metrics. The most popular overlap-based metric is the Dice similarity coefficient (DSC), mathematically defined as

$$DSC = \frac{2|A \cap B|}{|A| + |B|}, \tag{1.1}$$

where $|A|$ denotes the number of elements in A. DSC can assume values between 0 (no overlap) and 1 (equivalent to $A = B$). Two examples for boundary-based metrics are the Hausdorff distance (HD) and the normalized surface dice (NSD). HD measures the largest distance between the boundaries of two segmentation masks:

$$HD(A, B) = \max\left\{\max_{a \in A} d(a, B), \max_{b \in B} d(b, A)\right\}, \tag{1.2}$$

where $d(a, B) = \min_{a \in A} ||a - b||$ is the distance of point a to the set of points in B. A good segmentation minimizes the HD. The NSD (Nikolov et al. 2021) takes a different approach, introducing a tolerance parameter τ, which can be adapted to the application and determines how much deviation between the boundaries is acceptable. Denoting the boundary of a segmentation mask with $\mathcal{S}$, which is a surface in a 3D volume, and the border region with width τ around that surface with $\mathcal{B}^{(\tau)}$, the NSD is calculated as

$$NSD = \frac{|\mathcal{S}_A \cap \mathcal{B}_B^{(\tau)}| + |\mathcal{S}_B \cap \mathcal{B}_A^{(\tau)}|}{|\mathcal{S}_A| + |\mathcal{S}_B|}. \tag{1.3}$$

Here, $|\mathcal{S}_A \cap \mathcal{B}_B^{(\tau)}|$ is defined as the length of that part of the boundary $\mathcal{S}_A$ that is contained in the border region $\mathcal{B}_B^{(\tau)}$. Similar to DSC, this metric ranges from 0 to 1, but it focuses on boundary overlap instead of volume overlap.

Since all metrics have their particular pitfalls and measure different aspects of the segmentation performance, it is usually recommended selecting them for validation based

on the dataset and annotation characteristics in a particular application (Maier-Hein et al. 2024). If possible, evaluating both overlap-based metrics and boundary-based metrics allows a more comprehensive analysis.

1.2.2 Distribution Shifts and Medical Imaging

Supervised DL usually minimizes a loss function on the available training data, but the final goal is to achieve low error on unseen data as well. While these algorithms usually generalize to new samples from the same data distribution, often referred to as independent and identically distributed (i.i.d.) samples, they often exhibit higher error rates in real-world applications with data from a different distribution (Koh et al. 2021). The discrepancy between training and testing distribution is called distribution shift or dataset shift (Moreno-Torres et al. 2012; Quiñonero-Candela et al. 2022). This section describes the theoretical background and how distribution shifts practically arise in many medical imaging applications.

Mathematically, for a supervised learning task with random variables X as input features and Y as prediction targets, let the training data distribution $P_{\text{tr}}(X, Y)$ and the data distribution during testing $P_{\text{ts}}(X, Y)$. Sometimes these are called source domain and target domain, respectively. The situation $P_{\text{tr}}(X, Y) \neq P_{\text{ts}}(X, Y)$ corresponds to dataset shifts. In the general Machine Learning (ML) literature, a few special shifts can be identified when factorizing the joint distribution in one of two ways: $P(X, Y) = P(Y|X)P(X)$ or $P(X, Y) = P(X|Y)P(Y)$. Following the nomenclature by Moreno-Torres et al. (2012), these elementary shifts are:

- Covariate shift: It refers to a situation where the feature distribution changes but the mapping from x to y remains identical: $P_{\text{ts}}(X) \neq P_{\text{tr}}(X)$ and $P_{\text{ts}}(Y|X) = P_{\text{tr}}(Y|X)$.
- Prior shift: Here, only the output distribution is affected while the conditional distribution remains unchanged: $P_{\text{ts}}(Y) \neq P_{\text{tr}}(Y)$ and $P_{\text{ts}}(X|Y) = P_{\text{tr}}(X|Y)$.
- Concept shift: This is the challenging setting in which the conditional distributions between training and testing differ while the marginals stay the same: $P_{\text{ts}}(X) = P_{\text{tr}}(X)$ and $P_{\text{ts}}(Y|X) \neq P_{\text{tr}}(Y|X)$ or $P_{\text{ts}}(Y) = P_{\text{tr}}(Y)$ and $P_{\text{ts}}(X|Y) \neq P_{\text{tr}}(X|Y)$.

In practice, combinations of these shifts can make generalization even harder.

Towards a more practical description and a better understanding of these distribution shifts in the context of medical imaging, Castro et al. (2020) took a causality perspective and introduced an unobserved variable Z, which represents the physical reality of a subject's anatomy. An image X is then a noisy measurement of specific anatomical properties. The dependency structure between image X, label Y and anatomy Z can be visualized

as a directed graph, which reflects the factorization of the joint distribution. For image segmentation, the causal dependency is usually $Z \rightarrow X \rightarrow Y$, corresponding to a joint distribution of

$$P(X, Y, Z) = P(Z)P(X|Z)P(Y|X) \tag{1.4}$$

Analogously to the elementary shifts above, each factor can differ between training and testing domain D, resulting in three basic types of shifts for medical image segmentation (fig. 1.4). These correspond roughly to the prior, covariate, and concept shifts above, and are mathematically defined as:

- **Population shift** ($P_{\text{tr}}(Z) \neq P_{\text{ts}}(Z)$), which means that there are anatomical differences between the patient populations during training and deployment, for example in terms of age, sex or genetics.

- **Acquisition shift** ($P_{\text{tr}}(X|Z) \neq P_{\text{ts}}(X|Z)$), which is commonly encountered when the patient population is identical but the imaging acquisition process changes between training and testing, for example due to differences in scanners, protocols, or modalities.

- **Annotation shift** ($P_{\text{tr}}(Y|X) \neq P_{\text{ts}}(Y|X)$), which occurs if the annotation on the testing set is performed differently than on the training data. It can be caused by changes in the annotation protocol or differences between individual human annotators.

As before, in practice often multiple shifts occur simultaneously. For other tasks than image segmentation, the causal dependency structure may look different and the interpretation of the shifts changes, but since this thesis focuses on segmentation, the reader is referred to Castro et al. (2020) for more details on this topic. Categorizing distribution shifts in the way above helps to design datasets for studying individual shifts and develop approaches for tackling them.

How do such shifts arise in practice? Two frequent reasons are sample selection bias and non-stationary environments (Moreno-Torres et al. 2012). Sample selection bias means that the training data represents only a subset of the target population because some cases are not selected for training, for example due to image/annotation quality control or inclusion criteria in a clinical study. If these selection procedures are not replicated during deployment, the resulting data distribution can be shifted. Non-stationary environments describe the situation of the distribution changing over time, for instance due to new imaging equipment, change of labeling guidelines, or epidemics. Both selection bias and non-stationary environments can easily occur when models are developed on data from some hospitals and deployed in other hospitals, which is the federated data scenario investigated in this thesis.

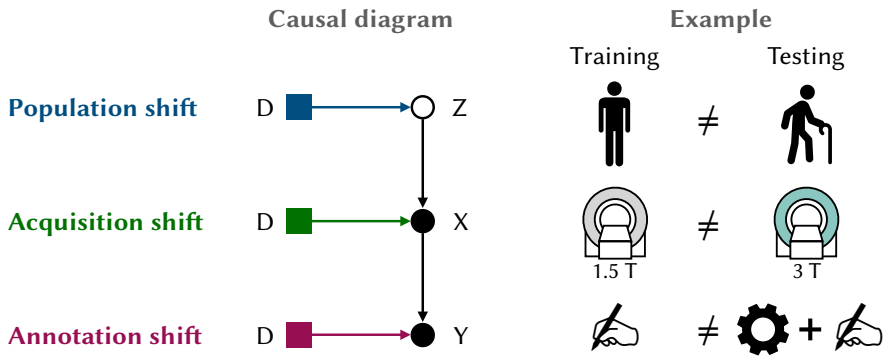

Figure 1.4: Probabilistic model and examples for distribution shifts encountered in medical image segmentation. The graphical model has random variables for the unobserved patient anatomy (Z), the observed image (X), and observed label (Y). Shifts between the training and testing domain are introduced by the domain indicator D. For population shift, only the patient population changes, for example through age differences. Acquisition shifts can be caused by changes in scanners or acquisition protocols. Annotation shifts can occur, for instance, if training images are annotated manually, whereas a semi-automatic workflow is used for the data the model is tested on.

1.2.3 Predictive Uncertainty Estimation

Shifts in the data distribution between the training and testing phases can lead to mistakes when segmenting images, both for humans and algorithms. However, while humans often notice such shifts as an anomaly and are able to express uncertainty in their segmentation (for example saying "I am unsure if this part belongs to the ROI because of an artifact in the scan"), deep learning models do not possess the inherent capability to estimate or explain their confidence. This section briefly describes how the notion of uncertainty can be formalized for general supervised learning problems and how uncertainty estimates can be used in practice. Individual methods for uncertainty quantification are described in the related work section 1.3.2.

In Bayesian statistics, a model for the data is specified as a likelihood $p(y|x, \theta)$, which defines the data-generating process as a function of the model parameters θ, and a prior distribution $p(\theta)$, which expresses the initial belief about which parameters are probable before observing any data (Gal 2016). Given a set of training data X with associated ground truths Y, this prior belief is updated, resulting in the posterior distribution

$$p(\theta|X, Y) = \frac{p(Y|X, \theta)p(\theta)}{p(Y|X)}. \tag{1.5}$$

Here, the numerator contains the likelihood evaluated over the whole dataset, while the denominator is called the model evidence and requires integration over all parameters:

$$p(Y|X) = \int p(Y|X, \theta)p(\theta)\, d\theta \tag{1.6}$$

In Bayesian inference, a prediction on a new data point x^* is performed by integrating over all possible parameter values, weighting their contribution by how likely they are according to the posterior distribution:

$$p(y|x^*, X, Y) = \int p(y|x^*, \theta)p(\theta|X, Y)\, d\theta \tag{1.7}$$

The final prediction is usually chosen as $\hat{y} = \arg\max p(y|x^*, X, Y)$ and its uncertainty is quantified by the spread of the distribution. Interestingly, eq. (1.7) also provides an intuition about the sources of uncertainty. Aleatoric uncertainty is captured by the first term in the integral and describes inherent randomness in the data (from the perspective of the model). In semantic segmentation, for example, most images contain ambiguous regions near the boundary of a structure of interest, which will be labeled slightly differently by independent annotators. The second factor in the integral of eq. (1.7), in contrast, stems from the uncertainty of the model parameters that were estimated from the available training data, often called epistemic uncertainty. Collecting more data can in principle reduce this uncertainty, as it restricts the posterior to a smaller region of probable parameters. An intuitive explanation of aleatoric and epistemic uncertainty is provided in fig. 1.5 for a two-dimensional classification example. Applying Bayesian inference in practice is not trivial, because solving the integrals in the predictive distribution and model evidence from eqs. (1.6) and (1.7) analytically is only possible for simple models. Variational inference and Monte Carlo (MC) sampling can help to find approximations (Gal 2016), but they are not discussed here. Instead, it is assumed that an approximate predictive distribution in the form of the average of M MC samples is available:

$$\overline{p}(y|x^*) = \frac{1}{M} \sum_{k=1}^{M} p(y|x^*, \theta^{(k)}) \tag{1.8}$$

While the predictive distribution is the most complete representation of uncertainty, often scalar uncertainty measures are useful in practice. Among the many possible measures, three are used in the experiments for this thesis: The softmax response, predictive entropy, and mutual information.

- The softmax response (also known as maximum softmax) quantifies how much probability mass is spread around the mode of the predictive distribution. It is

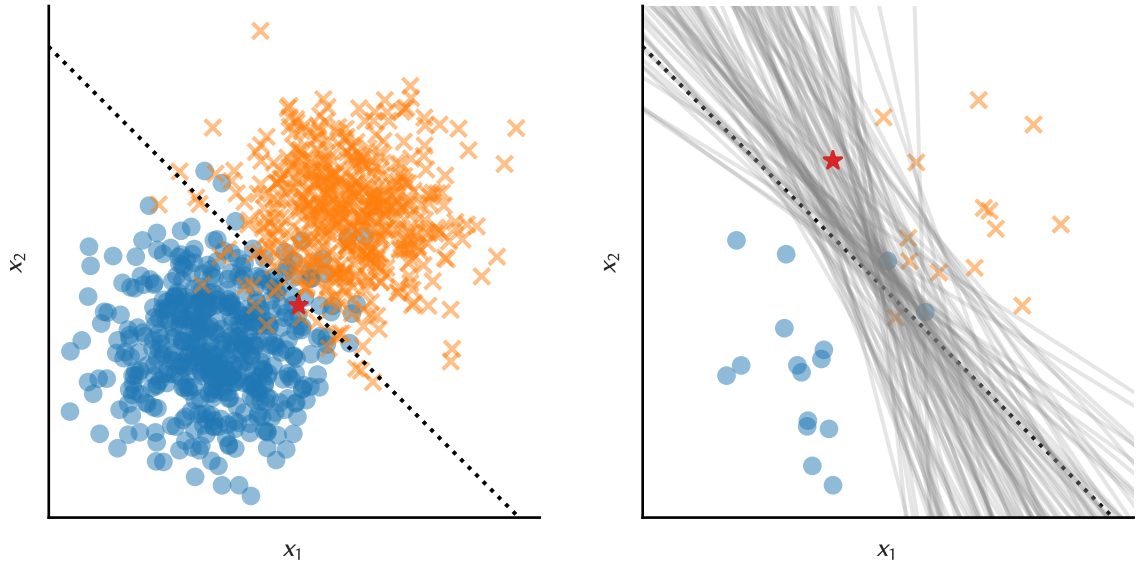

Figure 1.5: Illustration of two sources of uncertainty in model predictions in a two-dimensional, artificial example for a binary classification task, where blue circles belong to the negative class and orange crosses to the positive class. The dashed black line is the optimal decision boundary based on the true data distributions. Left: Aleatoric uncertainty is inherent in the data generation process, so any model should have lower confidence for data points like the red star, which are near the class boundary. Right: Epistemic uncertainty is rooted in the limited knowledge about the best model parameters, and can be reduced with more data. To illustrate this, 100 logistic regression models were fitted using scikit-learn (Pedregosa et al. 2011) based on the scarce training data. As their decision boundaries (gray lines) vary, the classification of a new sample at the location of the red star has high uncertainty.

defined via the maximum probability across classes c:

$$U_{SR} = 1 - \max_c \overline{p}(y = c|x^*) \tag{1.9}$$

- Predictive entropy is a measure from information theory that describes how close the distribution is to a uniform distribution:

$$U_{PE} = -\sum_c \overline{p}(y = c|x^*) \log \overline{p}(y = c|x^*) \tag{1.10}$$

- Mutual information was adapted for uncertainty estimation from information theory as well, capturing the shared information between prediction and model parameter

posterior. It is defined as

$$U_{MI} = U_{PE} - \frac{1}{M} \sum_{k=1}^{M} \sum_{c} p(y = c|x^*, \theta^{(k)}) \log p(y = c|x^*, \theta^{(k)}) \,. \tag{1.11}$$

This measure has been associated with epistemic uncertainty (Gal 2016).

The uncertainty measures above were proposed for image classification, but for a segmentation model they can also be computed per pixel. In the rest of this thesis, uncertainty and confidence are often used as antonyms, so all the uncertainty scores above can be inverted to get a confidence score.

Uncertainty estimates can be applied in the context of segmentation in several ways: Pixel-wise uncertainty maps can be used to interpret the model's output, which requires that the predictive probabilities are close to the empirical frequencies in the data (also known as calibration). For other downstream tasks, confidence scores are needed on the image level, so that pixel-level uncertainty estimation methods have to be extended with adequate aggregation methods. Prominent image-level applications include:

- Out-of-distribution (OOD) detection, which aims to identify samples that are outside the support of the training data distribution. Uncertainty quantification methods are used for OOD detection under the assumption that uncertainty is higher for unusual samples than for in-distribution data.
- Failure detection, which is tightly bound to a specific supervised task (like segmentation), with the goal of identifying erroneous predictions. If uncertainty scores are higher for low-quality predictions than for accurate ones, they can be filtered out.
- Active learning, a process in which data annotation and learning alternate, to make the labeling more efficient. Typically, an initial model is trained on a few labeled examples, and more samples are iteratively annotated thereafter. Selecting the most informative samples for annotation can be based on uncertainty scores.

These downstream applications can be used to evaluate the quality of uncertainty estimates, but each requires a specific evaluation methodology and may prefer different uncertainty methods (Kahl et al. 2024). This thesis focuses on failure detection, for which detailed evaluation methodology is presented in section 2.2.2.

1.2.4 Medical Image Analysis Competitions

In the field of medical image analysis, cutting-edge segmentation algorithms are often compared in international competitions, also known as *challenges*. All participants in

a challenge receive the same training data and try to develop the best algorithm for a specific medical image analysis task chosen by the organizing team. Testing data is not released to avoid overfitting. Challenges are an established method to fairly compare algorithms, with image segmentation of different structures as the most common task (Maier-Hein et al. 2018). Since one part of this thesis represents an analysis of the results for such a challenge (section 3.1), detailed background information is provided here on how these competitions are organized. While Maier-Hein et al. (2020) presented a template for comprehensive, structured reporting of essential design components, the concepts of challenges are described here more concisely from a practical viewpoint.

Before the competition takes place, an organizing research team has to prepare the challenge, which includes defining which image analysis task should be solved and what the required inputs and outputs of challenge submissions are. The organizers then collect imaging data and coordinate their annotations. The final challenge dataset consists of several cases, which comprise imaging and non-imaging data that serve for training or evaluating the submissions. For example, in a brain tumor segmentation challenge one or more MRI scans together with the ground truth segmentation make up one case. During the challenge preparation, it is also decided how to evaluate algorithms and how to compare them. Details on these assessment methods will be given further below when describing the challenge results analysis.

After the preparations, the actual competition begins, which is subdivided into a development phase and a testing phase. During the development phase, participants optimize their algorithmic solutions to the challenge task based on the training set and competition rules provided by the organizers. Sometimes evaluation on a separate public validation set is offered, so that participants obtain feedback on how well they do compared to others on a preliminary, public leaderboard.

At the end of the development phase, all participants submit their algorithms to the organizers. The successful submissions are then evaluated on a previously unseen test set. The essential components for assessing submissions are metrics, which quantify the performance of individual algorithms on the test set, and a ranking strategy, which compares the performance of different teams and determines the winners. For semantic segmentation, metrics are usually computed per case in the test set, but for other tasks like object detection or classification, some metrics take in the results on the whole dataset at once. Examples of segmentation metrics are the DSC and the NSD, which were introduced in section 1.2.1.3. One or more metrics can be computed, to evaluate different performance aspects relevant for the challenge task. Based on the table of metric values for each submission and test case, which represent the challenge results data, there are different ways how to arrive at a ranking. If the challenge results contain values for multiple metrics, rankings are usually computed for each metric first and later combined. A detailed

description of ranking methods can be found in (Maier-Hein et al. 2018; Wiesenfarth et al. 2021); here only the two most common strategies for computing a ranking of a single metric are briefly summarized:

1. *Aggregate-then-rank* approach: Metric values are aggregated across test cases for each submission first, for example through averaging. Teams are then ranked according to their aggregated values.

2. *Rank-then-aggregate* approach: First, all submissions are ranked on each test case individually. The resulting per-case ranks are then aggregated to arrive at a final ranking score.

While aggregate-then-rank methods are most intuitive, the rank-then-aggregate procedure allows handling cases more easily for which individual submissions did not produce an output. In general, however, both methods are valid choices and potential ranking differences resulting from them should be discussed in the challenge results analysis. In particular, ranking stability should be investigated (Maier-Hein et al. 2018) to assess how much randomness in the test data selection affects the ranking. Apart from determining the challenge winner, challenge organizers usually also perform analyses that characterize and compare submissions, deriving general insights into which solutions worked and, if possible, why.

In summary, challenges are a large joint effort by organizing and participating teams. As a challenge attracts independent participants developing diverse methods for the same task and fairly evaluates their submissions, the challenge results at the end of the competition represent a comprehensive picture of the current state of the art.

1.3 Related Work

As DL-based segmentation algorithms improve, they become more attractive for real-world clinical deployments, in which robustness is essential for a trustworthy operation. Hence, a growing number of studies focus on evaluating and improving the robustness of segmentation models when confronted with distribution shifts, which is also the context of this thesis. This section summarizes the related work, starting in section 1.3.1 with a description of methods and benchmarks that focus on the generalization task. After that, section 1.3.2 provides an overview of the previous work on uncertainty estimation in the field of medical image analysis, failure detection as a downstream task, and the current state of benchmarks.

Disclosure

Section 1.3.1 is based on the manuscript summarizing the FeTS challenge, which has been accepted for publication (Zenk et al. 2025a), so portions of the text resemble the original manuscript text.
Section 1.3.2 is derived from a previously published article (Zenk et al. 2025b), so portions of the text resemble the original manuscript text, in accordance with the publisher's license.
If parts of the text replicate sections from the corresponding manuscripts, this is explicitly stated beforehand.

1.3.1 Generalization

This section starts with a description of the problem settings in which generalization to distribution shifts is usually studied, and reviews relevant recent approaches. After that, previous efforts for benchmarking model generalization and applications of federated evaluation are described, with a focus on their relation to the benchmark developed in this thesis.

1.3.1.1 Problem Setting and Approaches

Generalizing to data distributions that differ from the training distribution but remain conceptually similar enough to perform image classification or segmentation, is a critical challenge in both medical image analysis and the broader field of computer vision. These data distributions are often referred to as domains. Background information on which distribution shifts are expected between domains in medical imaging was given in section 1.2.2. In a typical setup, the training data originates from one or more source domains, while the testing data belongs to distinct target domains. Various formulations with different constraints have been explored to solve this challenging generalization problem, as summarized by Zhou et al. (2023). One such setting is *domain generalization*, where a prediction model is trained solely on source domain data with the goal of achieving low prediction error on unseen target domains, ensuring effective generalization. Another approach is *test-time adaptation*, which assumes access to a limited amount of unlabeled data from the target domain during model deployment. This data can be used to tune the model on the target data distribution. A related setting, *unsupervised domain adaptation*, also utilizes unlabeled data from the target domain but assumes that training data from the source domain remains available during adaptation. However, in the context of deploying pretrained models in medical applications, sharing training data along with the model is

often unrealistic. Consequently, unsupervised domain adaptation is not further considered in this thesis. Lastly, *transfer learning* is an approach that adapts the model using a small amount of labeled data from the target domain. While transfer learning is promising for many applications, obtaining annotated target domain samples for segmentation tasks is a labor-intensive process. Therefore, this thesis rather focuses on generalization techniques like domain generalization and test-time adaptation, which do not rely on labeled target domain data.

Comprehensive surveys on popular approaches to domain generalization in natural image classification can be found in the articles by Gulrajani and Lopez-Paz (2021) and Zhou et al. (2023). Here, only general strategies from the surveys are repeated instead of referencing individual methods. Algorithms from the medical domain are described in the next paragraph. A frequent goal is to ensure that neural networks learn robust, domain-invariant features. This is achieved by introducing self-supervised auxiliary training objectives or by aligning feature distributions across source domains. Such methods encourage the model to focus on generic features rather than relying on domain-specific shortcuts, thereby improving its ability to classify samples from unseen domains. Another widely used technique is data augmentation, which mitigates overfitting and, to some extent, simulates potential domain shifts during training by applying diverse and realistic image transformations in the data-loading pipeline.

In the medical image computing domain, many works developed advanced data augmentation techniques. For instance, increasing the number or magnitude of augmentations applied to training samples has been shown to enhance generalization performance in semantic segmentation (Full et al. 2020; Zhang et al. 2020). More complex augmentation strategies have been proposed as well, such as leveraging transformations in the latent space (Chen et al. 2021) or simulating causal interventions with strong photometric transformations (Ouyang et al. 2023). The latter method also explicitly encourages the model to learn domain-invariant features by applying two such augmentations to the same image and aligning the two resulting predictive distributions of the model. Test-time adaptation is another common strategy in medical imaging. For example, Karani et al. (2021) adapt a normalizer block that precedes the main segmentation model. The parameters of this block are fine-tuned such that the main segmentation model's output becomes plausible under an implicit prior learned during training via a denoising auto-encoder. This method effectively transforms target domain images to resemble those from the source domain. Similarly, He et al. (2021) introduced shallow adaptor blocks at each stage of a U-Net architecture, which are adjusted at test time using auto-encoders that capture the training distribution across various feature levels. A more detailed overview of relevant methods for domain generalization and test-time adaptation in medical imaging is provided in Yoon et al. (2024).

1.3.1.2 Robustness Benchmarking Efforts

New generalization methods require careful validation of established benchmarks. This section describes previous benchmarking efforts and highlights research gaps, which are relevant for this thesis because it also presents a benchmarking study on generalization to unseen institutions.

In the general ML community, domain generalization methods were tested in various settings, introducing distribution shifts through artificial corruptions (Hendrycks and Dietterich 2019), independent dataset collections (Gulrajani and Lopez-Paz 2021) or realistic shifts occurring in real-world applications (Koh et al. 2021). Synthetic image transformations that introduce distribution shifts are also used in medical image segmentation to study generalization capabilities (Chen et al. 2021; Boone et al. 2023) or uncertainty estimation on OOD samples (González et al. 2022; Ng et al. 2023). Furthermore, distribution shifts are present in multi-centric datasets, which are often constructed by combining multiple public datasets from different sources (Liu et al. 2020; Wang et al. 2020b; Ogier du Terrail et al. 2022; Korevaar et al. 2023).

While these works use publicly available datasets and define standardized train-test splits, another group of studies implements benchmarks in the framework of international competitions, also known as challenges, which invite participants to develop their own algorithm and use standardized datasets, often keeping the testing data private for a fair comparison (details in section 1.2.4). Brain tumor segmentation algorithms have been studied in the challenge setting for almost a decade, resulting in large, multi-centric datasets today (Menze et al. 2015; Bakas et al. 2019; Baid et al. 2021), but the evaluation did not focus on the generalization to unseen medical centers before the work performed in this thesis. This kind of robustness was explicitly considered in other competitions, using images from different geographical sites and scanners for cardiac segmentation, mitosis detection, and fetal brain annotation (Campello et al. 2021; Aubreville et al. 2023; Payette et al. 2024), for example. However, the diversity in their testing data was limited to two unseen domains, arguably due to the difficulty of collecting large, multi-centric datasets. The results of the three above-cited competitions identified data augmentation and ensembling as success factors for generalizing to new institutions or scanners.

1.3.1.3 Federated Evaluation for Medical Image Analysis

Research can benefit from sharing data and medical images between institutions, but there are also high hurdles associated with it, such as patient privacy and a lack of standardization (Bell and Shimron 2024). Federated workflows can help to overcome some of them, by keeping the data decentralized at each owner's institution and sending around algorithms instead. For validating models on large-scale data, the collaborative, multi-site evaluation

Table 1.1: Comparison of related work for evaluating generalization on unseen institutions. The subset of all previous work that is most relevant to this thesis is shown. Existing works either perform a federated evaluation of individual algorithms after federated learning (above the gray line) or evaluate on centralized data for benchmarking competitions. The benchmark presented in this thesis (marked with *) is the first to employ federated evaluation in a competition, which allows scaling up the number of cases from unseen domains/institutions significantly, to test the robustness of segmentation models "in the wild".

Paper	Task	# unseen domains	# cases unseen domains	Competition?	Fed. evaluation?
Dayan et al. (2021)	Classification	3	1503	✗	✓
du Terrail et al. (2023)	Classification	2	157	✗	✓
Dou et al. (2021)	Obj. Detection	3	55	✗	✓
Pati et al. (2022a)	Segmentation	6	590	✗	✓
Aubreville et al. (2023)	Obj. Detection	2	51	✓	✗
Campello et al. (2021)	Segmentation	2	64	✓	✗
Malinin et al. (2022)	Segmentation	2	99	✓	✗
Payette et al. (2024)	Segmentation	2	80	✓	✗
Zenk et al. (2025a)*	Segmentation	26	2253	✓	✓

of models (federated evaluation) is especially interesting. In the context of mobile devices, Wang et al. (2019b) and Paulik et al. (2021) describe experiments and a technical system, respectively, in which federated evaluation serves to determine the best hyperparameters for personalizing ML models to devices of individual users. Studies on federated learning for medical image analysis applications routinely perform federated evaluation, often also setting aside individual institutions for external validation of generalizability (Dayan et al. 2021; Dou et al. 2021; Pati et al. 2022a; Ogier du Terrail et al. 2023). While practical hurdles often prevent federated learning from being established across a large consortium of real medical centers (Bujotzek et al. 2024), federated evaluation alone can be significantly simpler to implement and offers opportunities for benchmarking real-world robustness independently of federated learning (Karargyris et al. 2023).

Conclusion Cross-site generalization is an important research problem in medical image segmentation and has gained attention in recent years. Although many methods aim to tackle it, they have so far only been benchmarked on datasets from a few institutions, as shown in table 1.1. Federated evaluation workflows have the potential to overcome this

limitation, but have not been used in biomedical competitions before the research reported in this thesis.

1.3.2 Failure Detection

Various approaches share the common goal of identifying cases where model predictions are likely to be erroneous, including uncertainty estimation, segmentation quality estimation, and distribution shift or OOD detection. Typically, uncertainty estimation operates at the pixel level, whereas quality estimation and distribution shift detection are focused on the image level. Transitioning from pixel-level to image-level metrics can be achieved through confidence aggregation techniques. This section is organized accordingly, beginning with pixel-level methods, followed by aggregation strategies, and concluding with image-level approaches. The latter methods are further sub-categorized into methods for segmentation quality estimation and distribution shift detection. Finally, a section about benchmarking efforts for failure detection is included, to point out how the benchmark presented in this thesis fills a gap in the previous work.

1.3.2.1 Pixel-level Confidence Methods

In a recent survey, Lambert et al. (2024) summarized existing work on uncertainty quantification methods used in medical image analysis. From those, the most popular methods that are applicable to segmentation networks are described here.

As a fundamental baseline, confidence maps can be computed directly from the softmax probabilities of segmentation models based on deep learning, for example by using the probability of the predicted class or by calculating the entropy of the predictive distribution (as described in section 1.2.3). Guo et al. (2017) found, however, that neural network predictions are often overconfident when being wrong, and proposed temperature scaling of the probabilities as a simple solution. This initiated extensive research on how to improve the networks' calibration, which is the property that the probability of being correct and the predicted probability p are close on average across a large data sample.

More advanced methods rely on Bayesian deep learning, which is based on the idea that uncertainty can be better estimated when taking into account the full posterior of the model parameters given the training data (see section 1.2.3 for details), instead of relying on a point estimate, as done by deterministic neural networks. In Bayesian neural networks (NNs), parameters do not have fixed values but are assigned a probability distribution, for example a normal distribution with learnable mean and variance (Blundell et al. 2015) or a distribution on a rank-1 subspace (Dusenberry et al. 2020). Since exact Bayesian inference is intractable with large neural networks, approximations like variational inference are usually employed. Gal and Ghahramani (MC-Dropout 2016) proposed

one of the most popular uncertainty estimation methods so far, finding that training networks with Dropout layers (Srivastava et al. 2014) approximates Bayesian inference. Their method uses MC samples obtained by keeping Dropout activated during prediction to estimate the uncertainty. Due to its simplicity and flexibility, adaptations to medical image segmentations are abundant (Roy et al. 2019; Hoebel et al. 2020; Jungo et al. 2020; Kwon et al. 2020; Mehrtash et al. 2020; Nair et al. 2020).

A common alternative to Bayesian NNs are ensembles (Lakshminarayanan et al. 2017). After training multiple models with different random initialization, their averaged predictions on test samples can improve the accuracy and simultaneously provide pixel uncertainties from their prediction probabilities via entropy or mutual information, for example. Although multiple networks have to be trained for an ensemble, the method is simple to implement and its confidence estimates have been shown more reliable than those of other methods even in settings with distribution shifts (Ovadia et al. 2019). Therefore, ensemble uncertainty was also applied to medical image segmentation, for example by Mehrtash et al. (2020) and Hoebel et al. (2022).

Another approach to pixel-level uncertainty estimation utilizes test-time augmentation (Wang et al. 2019a). From each test sample, this method generates multiple versions using image transformations. By feeding each version into the segmentation model and reversing the applied transformations, multiple predictions are obtained. Similar to an ensemble, the consensus prediction can be more accurate than individual predictions, and pixel-level uncertainty can be estimated from the predictive distributions. Kahl et al. (2024) found that test-time augmentation primarily models epistemic uncertainty.

While the methods above were based on discriminative NN, generative segmentation models have been explored for estimating aleatoric uncertainty. Kohl et al. (2018) proposed to combine a U-Net with a conditional Variational autoencoder (VAE), which allows sampling from a distribution of plausible segmentations that reflect the ambiguity in the ground truth and variation between different raters. Towards the same goal, Monteiro et al. (2020) suggested learning a low-rank multivariate normal distribution over the logit space instead, which benefits efficiency.

Evidential deep learning takes yet another approach to uncertainty estimation by introducing distributional uncertainty as an additional component besides aleatoric and epistemic uncertainty. These methods model the predictive distribution explicitly with a Dirichlet distribution. By parametrizing the Dirichlet with a neural network, they can produce uncertainty estimates in a single forward pass (Malinin and Gales 2018; Sensoy et al. 2018). This approach was mostly applied to natural image classification so far, but also transferred to brain tumor segmentation by Zou et al. (2022).

1.3.2.2 Pixel Confidence Aggregation Methods

The aggregation of pixel-level uncertainties into an image-level score is necessary for detecting whether the segmentation as a whole is erroneous, yet it has received limited attention in the literature. A simple baseline is the mean confidence over the entire image or only the foreground (Mehrtash et al. 2020; González et al. 2022), which is sometimes replaced with the sum or log-sum (Nair et al. 2020; Czolbe et al. 2021).

One study on brain tumor segmentation (Jungo et al. 2020) compared various aggregation techniques. Apart from a simple averaging baseline, the authors developed methods based on prior knowledge about typical uncertainty maps and also investigated how automatically extracted features of the confidence maps can improve aggregation. These features describe the texture and shape (among others) of a region of interest and are also called radiomics features (Griethuysen et al. 2017). The results demonstrated that the latter methods, learning from radiomics features, improved failure detection performance.

In the comprehensive evaluation of uncertainty methods by Kahl et al. (2024), three aggregation methods were used: (i) a simple sum of uncertainties, (ii) patch-based aggregation, which sums uncertainties for each patch in a sliding window fashion and computes the maximum uncertainty across all patches, (iii) threshold-based aggregation, which averages all pixel uncertainties above a threshold that is tuned on the validation set. The last two methods are meant to address the inherent bias of mean uncertainty towards images with large foreground regions.

1.3.2.3 Image-level Failure Detection Methods

Segmentation Quality Estimation For segmentation algorithms in clinical use, quality control is essential for a safe and trustworthy operation. As the true segmentation is usually not available in such settings, many methods aim to estimate the segmentation quality in terms of metrics such as DSC without access to the ground truth.

Formulated as a regression task, various ML methods can be applied to solve it based on suitable training data. Early approaches trained support vector machine regressors on handcrafted features of the segmented region (Kohlberger et al. 2012). More recently, features were learned directly from the raw data while training deep neural networks to regress target quality measures like the DSC scores (Robinson et al. 2018). As failure cases with low segmentation quality are rare, this work also balanced the training distribution of target DSC scores to avoid biases towards high values. These regression networks were further refined with a decoder network and a secondary training objective that aim at predicting the pixel-wise segmentation error map (Qiu et al. 2023), for example.

Reverse classification accuracy (RCA Valindria et al. 2017) involves fitting a "reverse" segmentation model based on a single predicted segmentation, which is evaluated on

a reference database of images with ground truth. The authors hypothesize that the accuracy of the predicted segmentation is correlated with the best accuracy achieved on the reference database by the reverse model because a good predicted segmentation should transfer at least to some samples. The reverse segmentation model can be freely chosen, but registration-based approaches worked best in the experiments by Valindria et al. (2017). As this method requires registering each prediction to the complete set of reference images, it is computationally expensive.

Generative models have also found application in the context of segmentation quality estimation. Wang et al. (2020c) proposed a method that leverages a VAE (Kingma and Welling 2013) trained on pairs of images and their corresponding ground truth masks, thus learning a model of good-quality segmentations. For a test sample, the method encodes an image-mask pair and optimizes their latent representation. Decoding the optimized representation yields a surrogate image and mask from the manifold of good-quality segmentations. The true segmentation metrics are then approximated with the metrics calculated between original and surrogate mask. Xia et al. (2020) and Li et al. (2022) combine a generative model with a comparison/difference module that is similar to a regression network. The generator is trained to synthesize images based on ground truth segmentation masks. For the difference module, two heads for image-level and pixel-level quality estimation are then trained on the features of a Siamese network that encodes both the original and synthetic images. During test time, predicted masks are fed into the generator, which can result in artifacts in the synthetic image if the mask is inaccurate. The difference module can map these inconsistencies to quality estimates.

The methods described above are model-agnostic, meaning they can be applied to segmentation masks generated by any algorithm, including those produced manually by humans. While this versatility is advantageous, it also has potential limitations: The quality estimation method itself may fail due to biases in its training data, and a model-specific quality estimator may achieve superior performance. An alternative approach by DeVries and Taylor (2018) tailors a quality regression network to a segmentation model that provides pixel confidence maps by incorporating the confidence map as an additional input. This method could also be interpreted as a confidence aggregation technique. Another example method builds on MC-Dropout, which generates multiple predictions for each test sample. Roy et al. (2019) propose calculating various uncertainty measures based on these Monte Carlo samples. For example, segmentation metrics like the DSC can be computed between each pair of predictions, with the average across all pairs serving as a quality estimate.

Distribution Shift Detection As explained earlier, deep neural networks tend to fail more often when the testing data does not originate from the same data distribution as

the training data. Therefore, detecting distribution shifts has been explored as another approach to failure detection, based on the assumption that predictions on OOD samples cannot be trusted. Numerous works on OOD detection have been published in the ML community (Salehi et al. 2022).

In the field of medical image segmentation, distribution shift detection is often implemented as density estimation of the training data distribution. Confidence scores can be obtained by evaluating the likelihood of a test sample under the probabilistic model. The approaches differ primarily in the choice of features for density estimation and the probabilistic models employed. For example, Liu et al. (2019) train a VAE on ground truth segmentations and the likelihood approximated by the VAE loss serves as confidence score. They also fit a linear model that maps these scores to segmentation metrics to make this method applicable for segmentation quality estimation, but this is not strictly required for OOD detection. Alternatively, Graham et al. (2022) proposed to extract image features with a VQ-GAN (Esser et al. 2021) and estimate probability density with a transformer network. Another method leverages the latent representations produced by the segmentation network for the training set. A multivariate Gaussian is fitted to these latent features, and uncertainty is quantified as the Mahalanobis distance of a test sample from the training distribution (González et al. 2022).

1.3.2.4 Benchmarking Efforts for Failure Detection

This thesis does not focus on individual failure detection methods but rather compares several recent methods on multiple datasets, covering different anatomical regions and target structures. This section illustrates the need for such a benchmark, which is also summarized in table 1.2.

In the field of image classification, Jaeger et al. (2022) developed the first comprehensive benchmark for failure detection on natural images, recommending a risk-coverage analysis as a unifying evaluation protocol. For medical images, studies on failure detection on multiple datasets with and without distribution shifts (Bernhardt et al. 2022; Bungert et al. 2023) showed that no confidence estimation method could reliably outperform a simple softmax baseline.

For medical image segmentation, existing benchmarking studies are limited to a specific anatomical region or a subset of methods. Jungo et al. (2020) evaluate pixel confidence methods in the tasks of uncertainty calibration, error localization, and failure detection (after confidence aggregation) on a single brain tumor dataset without distribution shifts. Mehrtash et al. (2020) considered distribution shifts in two of three datasets but did not include image-level methods, as they concentrated on the calibration of pixel confidence methods, investigating failure detection only as a secondary task. Hoebel et al. (2022)

Table 1.2: Comparison of related studies with benchmarking character. The benchmark presented in this thesis (marked with *) is the only one that evaluates both image-level methods and a range of pixel-level + aggregation combinations for failure detection on multiple datasets with distribution shifts. Only medical datasets are counted and datasets with shift may also include some cases without shift in the test set. (✓) in the methods columns means that only a single method was investigated. The evaluated tasks are failure detection (FD), out-of-distribution (OOD) detection and calibration (calib.). Img/cls/pxl refers to image-/class-/pixel-level for FD.

Paper	**Methods**			**# Datasets**		**Task**		
	Image-level	Pixel-level	Aggreg.	no shift	with shift	FD	OOD	Calib.
Jungo et al. (2020)	✗	✓	✓	1	0	im	✗	✓
Mehrtash et al. (2020)	✗	✓	(✓)	1	2	cls	✗	✓
Mehta et al. (2022)	✗	✓	✗	1	0	px	✗	✗
Hoebel et al. (2022)	(✓)	✓	(✓)	0	2	im	✓	✗
Malinin et al. (2022)	-	-	-	0	1	im	✗	✗
Li et al. (2022)	✓	✓	✗	0	1	im, px	✗	✗
Ng et al. (2023)	✗	✗	(✓)	0	1	cls	✗	✓
Adams et al. (2023)	✓	✗	(✓)	1	1	cls	✗	✗
Vasiliuk et al. (2023)	✓	✓	(✓)	0	2	✗	✓	✗
Kahl et al. (2024)	✗	✓	✓	0	1	im	✓	✓
Zenk et al. (2025b)*	✓	✓	✓	1	5	im	✓	✗

experimented with two datasets that had different pathology characteristics in the test split and compared only a single image-level and aggregation method. Segmentation quality methods for the single application of cardiac segmentation were evaluated separately on the pixel- and image-level by Li et al. (2022), but no aggregation methods were considered. Another benchmark for the heart region had dataset shifts in the test set and did not consider image-level methods (Ng et al. 2023). Recently, Adams and Elhabian (2023) evaluated failure detection methods on two organ segmentation tasks. However, only one of the datasets had distribution shifts and the method selection was restricted to pixel confidence methods with a single aggregation (sum). Vasiliuk et al. (2023) evaluated a similar task as OOD detection, incorporating two datasets with distribution shifts as OOD characteristics. Their evaluation does not allow in-distribution failure detection benchmarking, though. Finally, the comprehensive evaluation of uncertainty estimation methods by Kahl et al. (2024) examined multiple downstream tasks—among them failure detection—on one medical and one nonmedical dataset. They focused on the subset of methods with pixel-level uncertainty and aggregation, if necessary for the downstream

task.

Beyond the research studies above there are also two international competitions that are related to segmentation failure detection. The BraTS challenge 2020 (Mehta et al. 2022) evaluated uncertainty estimation methods for brain tumor segmentation, but their metric was compatible only with pixel-level methods and the challenge dataset did not include distribution shifts. The Shifts Challenge 2022 (Malinin et al. 2022) addressed distribution shifts in the test data, evaluating robustness and failure detection for lesion segmentation in the context of Multiple Sclerosis and another nonmedical dataset. A meta-analysis of the submitted approaches has not yet been published so far.

1.4 Objectives and Contributions

The motivation and related work sections (sections 1.1 and 1.3) showed that robustness in medical image segmentation can be approached through methods for generalization or failure detection, and that existing evaluation efforts for both categories usually compare algorithms on small research datasets in individual segmentation tasks, covering only a fraction of realistic distribution shifts. The overall goal of this thesis is to build large-scale benchmarks and to develop evaluation methodology which allows comparing state-of-the-art methods in generalization and failure detection. The objectives and contributions below are described separately for these two approaches to robustness.

1.4.1 Generalization

Validating the robustness of medical image segmentation methods to realistic distribution shifts is essential for clinical translation, yet public benchmarking competitions have so far been restricted to small-scale evaluations, involving one or two institutions/scanners not seen during training (Campello et al. 2021; Aubreville et al. 2023; Payette et al. 2024). Therefore, the objective of this thesis' part on generalization was to develop a large-scale, multi-institutional benchmark that assesses how well segmentation algorithms generalize to a diverse data distribution of images acquired in clinical routine at many sites, thus estimating performance "in the wild".

Contributions Work for this part of the thesis focused on the task of brain tumor segmentation in multi-modal MRI, which is difficult due to the heterogeneous shape and appearance of these tumors. An international competition was conducted, in which segmentation algorithms developed by external participants were evaluated on a large test dataset from 32 institutions, by implementing a federated evaluation workflow. The idea behind federated evaluation is to send segmentation algorithms to many collaborating

institutions that agreed to evaluate these models on their local data, which circumvents the high hurdles for sharing medical images and allows leveraging large, decentralized multi-centric datasets. Research performed for this thesis represents the first benchmarking competition to evaluate models in a federated environment, thus establishing a large-scale evaluation paradigm that has the potential to close the gap between research and clinical validation.

Research questions (RQs) investigated in the federated evaluation study:

RQ 1.1: Do current brain tumor segmentation algorithms generalize "in the wild"?

RQ 1.2: Which algorithm and dataset characteristics affect generalization?

RQ 1.3: Which practical hurdles are associated with federated evaluation?

1.4.2 Failure Detection

Models can also become more robust and trustworthy if they are able to indicate potential errors in their predictions. The task of detecting failures has motivated various methods related to uncertainty and quality control, but incompatible evaluation protocols have prevented a comprehensive comparison so far (Jaeger et al. 2022). The second part of this thesis aimed to fix this shortcoming, by developing a benchmark that compares diverse approaches towards the shared goal of segmentation failure detection on a wide range of medical imaging datasets with realistic distribution shifts.

Contributions In a first step, existing, unstandardized evaluation protocols were compared and essential requirements for failure detection evaluation were derived. Based on these, a unifying protocol was proposed that served as the basis for a large-scale benchmark of failure detection approaches. This study is the first to jointly evaluate methods from three major fields: uncertainty estimation, distribution shift detection, and segmentation quality estimation. A special focus of the benchmark was on confidence aggregation methods, which aggregate pixel-level confidence maps into image-level confidence scores. Although the aggregation strategy is an important component of many failure detection methods, it received little attention so far. To cover a broad spectrum of segmentation tasks with realistic failure sources, five publicly available CT and MRI datasets for organ and

lesion segmentation were employed in the benchmark. Distribution shifts were introduced in the test set by using data from different institutions, scanners or with different tumor type prevalence.

Research questions (RQs) investigated in the failure detection study:

RQ 2.1: What are best practices and pitfalls related to the evaluation of segmentation failure detection?

RQ 2.2: Which failure detection algorithms are reliable across multiple datasets?

RQ 2.3: How to aggregate pixel-level confidence into image-level scores for failure detection?

1.5 Outline

Each of chapters 2 to 4 is split into two subsections, which describe the work performed for this thesis on the topics of generalization and failure detection, respectively.

The generalization part starts with section 2.1, describing the materials and methods for the international competition organized in the context of this thesis. This includes the datasets, infrastructure, and evaluation methodology used for the study, as well as statistical analyses. The results of the competition and summarizing analyses are reported in section 3.1. Practical experiences from implementing a federated evaluation workflow are also part of the results. Section 4.1 discusses the generalization capabilities of algorithms submitted to the benchmark, as well as limitations and possible improvements of the federated study design.

Moving on to the failure detection part, section 2.2 contains detailed information on the proposed evaluation protocol, alongside a description of the datasets and methods used for the segmentation failure detection benchmark. Section 3.2 reports experimental results for the benchmark, showing that segmentation failures occur particularly in datasets with distribution shifts and that some methods compared in the benchmark are able to detect failures robustly. These results are interpreted and compared to findings from related work in section 4.2. Limitations of the benchmark and alternative evaluation frameworks are discussed, too.

Returning to the overarching topic of segmentation model robustness, the overall conclusions of the generalization and failure detection studies are described in section 4.3. Finally, a summary of this thesis is provided in chapter 5.

2 Materials and Methods

This chapter describes the methodology for the benchmarking studies on the generalization of segmentation methods and failure detection. In section 2.1, the methodology of the Federated Tumor Segmentation (FeTS) Challenge is presented, addressing generalization, and in section 2.2, the approach for comparing state-of-the-art failure detection methods.

Disclosure

Section 2.1 is based on the manuscript summarizing the FeTS Challenges, which has been accepted for publication (Zenk et al. 2025a) and is partially based on a preprint by the same authors (Pati et al. 2021). Therefore, portions of the text in this section resemble these original manuscript texts.
Section 2.2 is derived from a previously published article (Zenk et al. 2025b), so portions of the text resemble the original manuscript text, in accordance with the publisher's license.
If parts of the text replicate sections from the corresponding manuscripts, this is explicitly stated beforehand.

2.1 Generalization

As argued in the introduction, it is usually unknown how state-of-the-art algorithms based on deep learning generalize to data from medical institutions that did not contribute to the training data. The approach taken in this thesis to benchmarking the generalization capability of segmentation models is an international competition, also known as *challenge*, which is the standard of fair and reproducible method comparison in medical image analysis (Maier-Hein et al. 2018). General background information on challenges is provided in section 1.2.4. In the Federated Tumor Segmentation (FeTS) challenge reported here, teams from around the globe could use standardized training data to develop their methods, which were subsequently evaluated on federated testing data. Brain tumor segmentation

was chosen as the specific medical segmentation task for the challenge, because it is a complex but also well-studied segmentation problem, meaning that the performance of existing methods is already close to inter-rater agreement when evaluated on data from institutions that also provided training data (Bakas et al. 2019). The FeTS challenge aimed to determine how these models fare on larger datasets from many institutions unseen during training. This section provides details on the challenge organization, datasets, evaluation methodology and technical infrastructure.

2.1.1 Challenge Organization

The FeTS Challenge took place in two consecutive years: 2021 and 2022, referred to as FeTS21 and FeTS22, respectively, in the following. In both years, the challenge consisted of two tasks: Task 1 focused on federated learning while Task 2 focused on generalization in the wild. As I was the main organizer of Task 2 while others were responsible for Task 1, this thesis exclusively describes work on Task 2. The challenge designs of FeTS21 and FeTS22 (Bakas et al. 2021; Bakas et al. 2022) were accepted as official challenges at the medical image computing and computer assisted intervention (MICCAI) conference after a peer-review process, to make sure best practices are followed (Maier-Hein et al. 2020). Both challenges were divided into three phases: In the development phase (about 3 months), participants received multi-centric data for model training and validation. In the submission phase, concurrent with the last month of the development phase, participants obtained instructions on how to prepare their algorithm for test set evaluation and summarized their approach in a scientific paper. Finally, in the evaluation phase (about 2 months) the organizers ran the federated evaluation on the official, decentralized test data.

2.1.2 Datasets

2.1.2.1 Data Sources and Characteristics

The data used in the FeTS challenges originated from two sources: The Brain Tumor Segmentation (BraTS) challenge (Menze et al. 2015; Bakas et al. 2017; Bakas et al. 2019; Baid et al. 2021) and the FeTS federated learning initiative (Pati et al. 2022a). A retrospective, multi-institutional cohort was provided by these data sources, containing patients diagnosed with primary brain tumors (gliomas). Each *case* corresponds to a single anonymized patient and comprises four multi-parametric magnetic resonance imaging (MRI) scans:

- Native T1-weighted (T1)
- Contrast-enhanced T1-weighted (T1-Gd), with gadolinium-based contrast agents
- T2-weighted (T2)

Table 2.1: Statistics of the training, validation and test cases for the FeTS challenges 2021 and 2022. Source refers to the context in which the data was originally collected, namely the BraTS challenges and the FeTS-initiative (FeTS-I). Information on the number of institutions in the validation set could not be shared by the BraTS organizers. Institutions in the test set were unseen during training, except 385 test cases from 6 institutions in FeTS22 that contributed to both training and test set. The number of cases increased significantly from 2021 to 2022, and so did the diversity of test cases, in terms of the number of contributing institutions. Abbreviations: img = image, ref = reference segmentation.

		Source	# cases	# institutions	Accessible by
FeTS21	Training	BraTS20	341	17	public (img, ref)
	Validation	BraTS20	112	n/a	public (img), organizers (ref)
	Testing	FeTS-I	796	22	data owners (img, ref)
FeTS22	Training	BraTS21	1251	23	public (img, ref)
	Validation	BraTS21	219	n/a	public (img), organizers (ref)
	Testing	FeTS-I	2625	32	data owners (img, ref)

- Fluid Attenuated Inversion Recovery (FLAIR)

The patients were scanned in standard clinical practice before surgery. An overview of the number of cases used for training, validation, and testing can be found in table 2.1. As the image analysis task of the challenge is segmentation, a reference segmentation of tumor subregions was available for each case. Details on the annotations can be found in section 2.1.2.3. For each training case, challenge participants additionally received meta-information about the originating institution in the form of an anonymized institution identifier, resulting in the institution partitioning of the training data shown in fig. 2.1.

The goal of the FeTS challenge was to test brain tumor segmentation algorithms "in the wild", so diversity is an essential characteristic of the test dataset and a prerequisite for evaluating algorithmic robustness. Diversity is quantified here first and foremost in geographical terms. In FeTS21, a total of 22 institutions contributed to the test set, representing four different continents. The vast majority of cases were collected in institutions in North America and Europe. The FeTS22 challenge further increased size and heterogeneity, featuring test data from 32 institutions around the globe (fig. 2.2). Compared to the BraTS 2021 challenge (Baid et al. 2021), which is one of the largest multi-centric datasets available for medical image segmentation, the test dataset of FeTS22 adds 24 federated institutions and 2116 test cases, which correspond to four times more cases than BraTS21 (table 2.2). This underlines the contribution towards real-world generalization benchmarking of the FeTS challenge.

For the FeTS22 test set, additional meta-data was shared by most contributing institu-

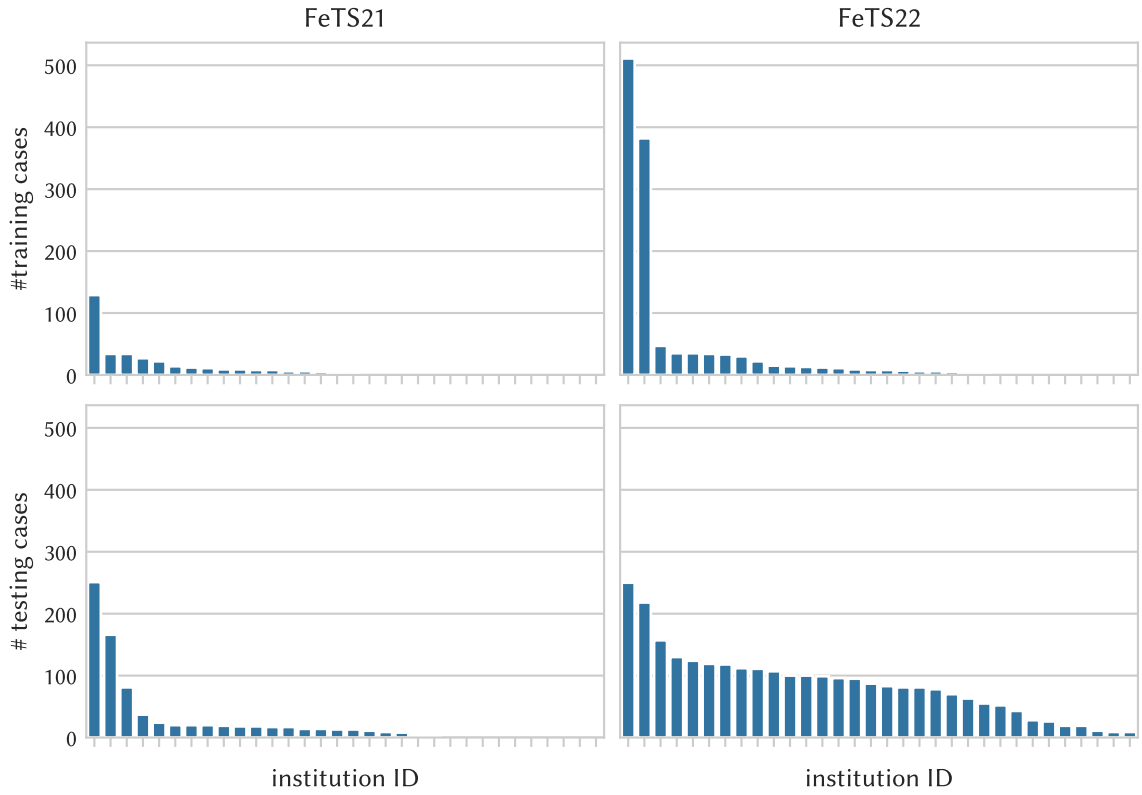

Figure 2.1: Partitioning of training and test sets by geographically distinct institutions. As different institutions contributed to the training and test sets, institution IDs are not consistent between subplots and are therefore not printed. While training cases (upper row) originated mostly from one or two institutions, the test cases (lower row) were distributed among a diverse federation of institutions. The overall number of cases increased significantly from FeTS21 (left) to FeTS22 (right).

tions with the organizers. This included information about the patient population (fraction of male/female patients, fraction of patients with isocitrate dehydrogenase (IDH) mutations, and age statistics) as well as information about the acquisition settings (scanner model, field strength, acquisition plane, MRI coil). From this institution-level meta-data gathered along with the imaging data, diversity can also be expressed in terms of population and acquisition characteristics. Patient population differences are visualized in fig. 2.3. Overall, the fraction of female and male subjects in the test set was 36% and 55%, respectively, and for 9% of patients no information about sex/gender was available. The population age varied between 7 and 94 years, with an average of 59 years. Mutations in genes that encode IDH play a role in the genesis and treatment of gliomas (Han et al. 2020). In the FeTS22 test set, information on the IDH status was available for 67% of

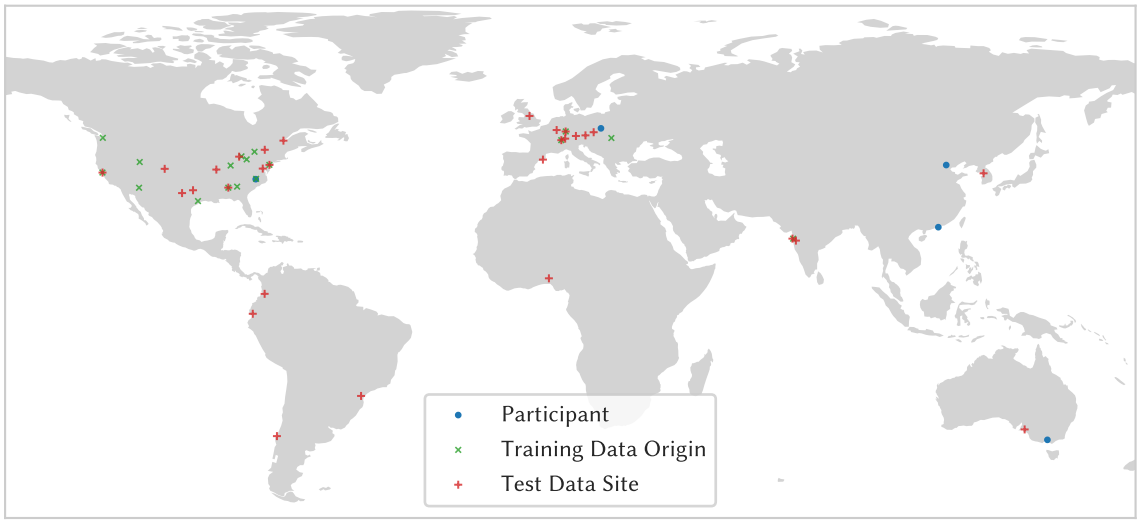

Figure 2.2: Geographical distribution of the participants, as well as training and testing institutions in the FeTS22 challenge. Training data was collected and shared with the participants, while the testing data remained distributed during the challenge. From the perspective of geographical diversity, FeTS22 represents a significant step towards evaluation "in the wild". Figure adapted from (Zenk et al. 2025a).

Table 2.2: Comparison of geographical diversity between the test sets of the BraTS 2021 (Baid et al. 2021) and FeTS challenges. While BraTS21 already accumulated a large, multi-centric dataset, FeTS21 added many new institutions and test cases. FeTS22 further scaled up the size considerably and also boosted diversity by including new continents. Note that BraTS21 is a subset of the FeTS22 test set, so the difference between the corresponding column pairs represents the increase through the use of a federation.

	Number of test cases			Number of institutions		
Continent	BraTS21	FeTS21	FeTS22	BraTS21	FeTS21	FeTS22
Africa	0	0	17	0	0	1
America (North)	319	508	1053	5	12	14
America (South)	0	15	166	0	3	4
Asia	47	0	339	1	0	3
Australia	0	20	96	0	1	1
Europe	143	253	954	2	6	9
Total	509	796	2625	8	22	32

patients, with most of them having the wild type (60%) and mutations being present only in a minority (7%). Finally, the heterogeneity of MRI scanning equipment can be quantified as the number of scanner models utilized: The 19 institutions who reported the corresponding information used 21 different scanner models in total, from four different vendors (Siemens, GE, Philips, Hitachi). Other differences in the acquisition settings included magnetic field strength (1.5 T or 3 T), acquisition plane (axial or sagittal) and use of different MRI coils.

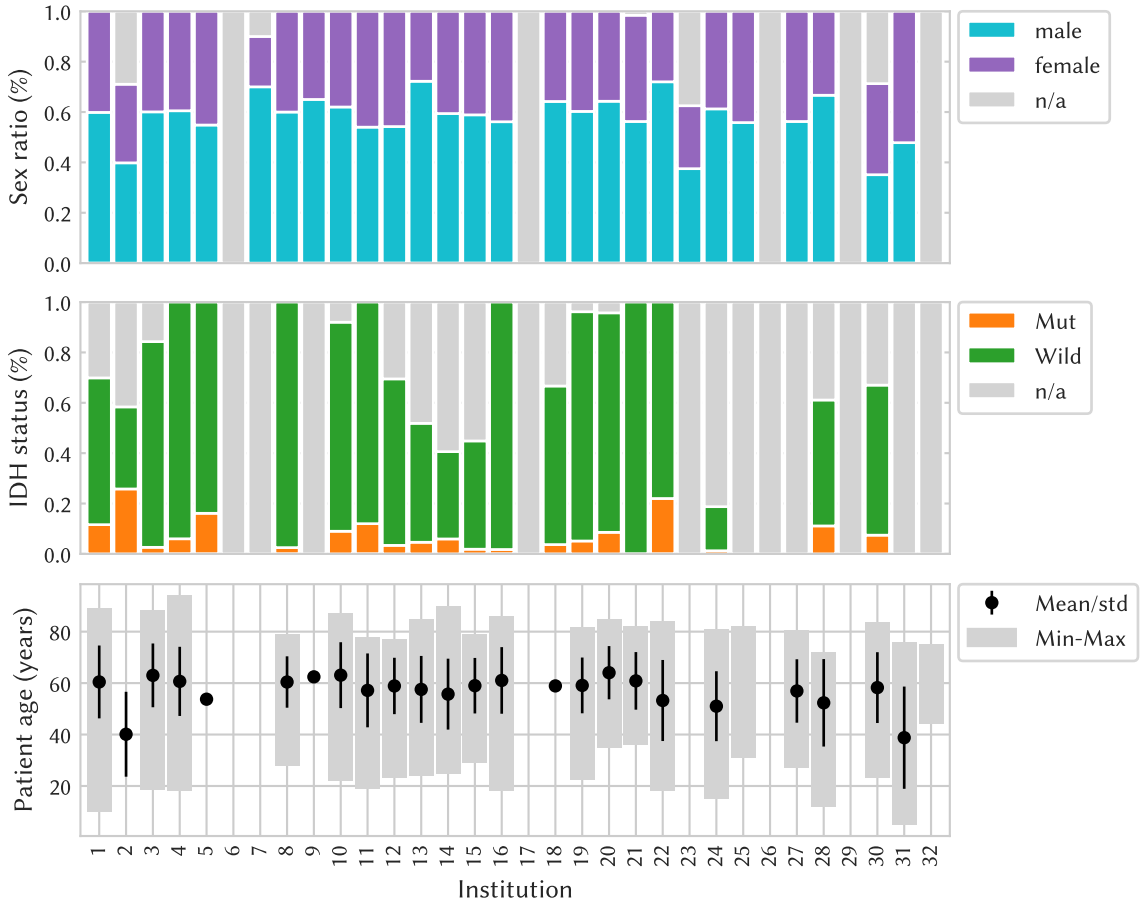

Figure 2.3: Patient population in terms of sex, IDH status, and age across 32 institutions that contributed test datasets to FeTS22. Some institutions could not provide any metadata and some only incomplete metadata, which leads to missing values in the diagrams. Overall, the populations are similar with respect to age, sex, and IDH status, but exceptions exist, for example, institution 2 with a high proportion of mutated IDH, and on average younger patients. Figure adapted from (Zenk et al. 2025a).

2.1.2.2 Preprocessing

All images used for this challenge were preprocessed during the BraTS challenges and the FeTS initiative, using the same preprocessing pipeline previously published through the Cancer Imaging Phenomics Toolkit (CaPTk, Davatzikos et al. 2018) and the FeTS tool (Pati et al. 2022b). It consisted of the following steps: First, all input MRI scans are rigidly registered to the SRI-24 anatomical atlas (Rohlfing et al. 2010) using the greedy diffeomorphic registration algorithm (Yushkevich et al. 2016). The scans were also resampled to a common spatial resolution of 1 mm^3. After that, a skull-stripping method based on deep learning (Thakur et al. 2020) was employed to extract the brain from the scans, which removes irrelevant signal and prevents facial re-identification of patients.

2.1.2.3 Annotations

The segmentation targets of the FeTS challenge are represented as masks, which assign each voxel to one of four classes, originally defined by Bakas et al. (2019) and Baid et al. (2021):

1. Enhancing tumor (ET) delineates the hyperintense signal of the T1-Gd sequence compared to T1. Neighboring vessels and sulci are not included. Class label: 4

2. Edema (ED) describes the peritumoral edematous and invaded tissue. It is hyperintense in the T2 and FLAIR sequences. Class label: 2

3. Necrotic core/necrocyst (NCR) consists of necrotic or cystic structures that are often located within the enhancing rim for high-grade gliomas. They appear hypointense in T1-Gd. Class label: 1

4. Background (BG) is defined as any tissue or surroundings that do not belong to the labels above. Class label: 0

In the first BraTS challenge, class label 3 was defined as the non-enhancing tumor, but it was later merged with NCR (Bakas et al. 2019). Figure 2.4 shows an example case with segmentation.

Annotations of these classes were performed on the preprocessed scans, following a clinically approved protocol (Bakas et al. 2019; Baid et al. 2021). Annotators received detailed guidelines on the radiologic appearance of each tumor substructure based on specific MRI sequences. They could choose their preferred annotation tool and perform the annotation either manually or semi-automatically, by combining automated segmentation with manual refinement.

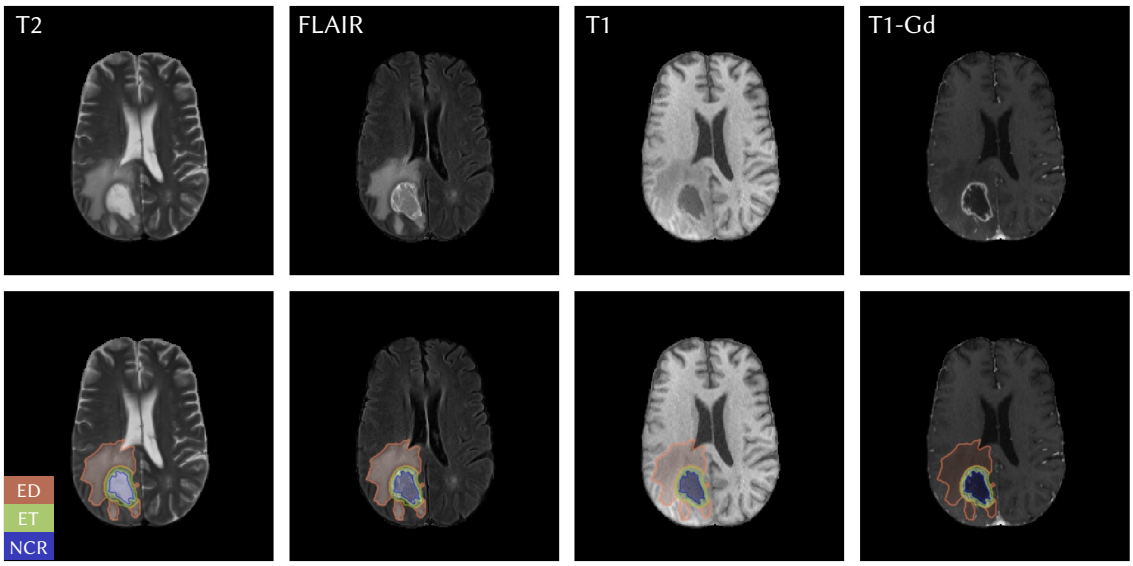

Figure 2.4: Training case example from FeTS22 along with its reference segmentation. One case consists of the four MRI sequences shown in the upper row. The lower row overlays the reference segmentation mask on the images. ED is best seen in the T2/FLAIR modalities, while ET and NCR are visible as contrast enhancements and hypointense regions, respectively, when comparing T1 and T1-Gd. The case shown here was chosen to illustrate the tumor labels and is easy to segment, but the variable appearance of gliomas makes this task more difficult in other cases.

The annotators differed between the BraTS and FeTS subsets of data (subsets are defined by "source" in table 2.1). For the BraTS data, each case was assigned to an annotator-approver pair. Annotators varied in experience, while approvers were experienced neuroradiologists with over 13 years of experience with glioma. Approvers reviewed and iteratively refined the annotations with the help of annotators until they were deemed of satisfactory quality for public release. For data from the FeTS federation, the annotation was done during the FeTS initiative (Pati et al. 2022a). Neuroradiology experts at each site created the reference segmentations with the tool of their choice following the BraTS annotation protocol (Bakas et al. 2019). However, the strong recommendation was to use a semi-automatic approach with predictions from three state-of-the-art brain tumor segmentation models, namely DeepMedic (Kamnitsas et al. 2017), nnU-Net (Isensee et al. 2021a), and DeepScan (McKinley et al. 2018), which were ensembled using the STAPLE algorithm (Warfield et al. 2004). These models were included in the toolkit provided to the data contributors, to facilitate the semi-automatic workflow (Pati et al. 2022a).

2.1.3 Annotation Quality Control

The FeTS challenge annotations were created by a diverse set of annotators using different tools, due to the federated nature of the challenge. Hence, it is important to quantify the quality and heterogeneity in the reference annotations. Although (Pati et al. 2022a) performed basic checks on the integrity of the annotations (making sure there were no unexpected label values) and identified problematic clients during federated learning based on their validation scores, they did not perform a systematic quality control in the federated data, because it was infeasible in the federated setting.

After the FeTS22 challenge, however, a part of the test data (1201 patients from 16 institutions) was shared with the organizers, so that a thorough quality control could be performed. Due to the lack of automated tools, I reviewed every available test case visually by looking for issues with the reference segmentation (such as inaccuracies, inconsistent labeling style, missing labels) or the underlying images (such as corruptions or artifacts). Based on this inspection, the quality of each case was classified as acceptable, borderline, or insufficient. Individual examples were discussed with a clinician scientist and a neuroradiologist. If the image was not corrupted and the reference segmentation accurate, the corresponding case received the "acceptable" quality label. Samples with "insufficient" quality contain major annotation errors or strong image artifacts that make segmentation impossible. The "borderline" label was introduced for cases in which the annotation has a few minor inaccuracies or in which it could not be decided without an official BraTS annotator whether the segmentation is consistent with the annotation protocol. Cases with insufficient quality were excluded from the challenge evaluation, resulting in a reduction of the total number of test cases from 2750 to 2625, which is the number reported in table 2.1.

2.1.4 Performance Assessment Methods

During the official challenge evaluation, the predicted segmentations of every submission were compared to the reference segmentation for performance assessment. Following the evaluation protocol of previous BraTS challenges (Bakas et al. 2019; Baid et al. 2021), the tumor structure labels from section 2.1.2 were converted into nested tumor subregions before evaluation because they reflect the clinical application task better (tumor volumetry, for instance). The final tumor regions for evaluation were:

- Whole tumor (WT), the union of all tumor structures (ED, NCR and ET)
- Tumor core (TC), the union of NCR and ET
- Enhancing tumor (ET), equivalent with the tumor label.

2.1.4.1 Metrics

The same two metrics as in the BraTS challenges were computed between the predicted segmentation and the reference segmentation, separately for each tumor region, specifically:

Dice similarity coefficient (DSC) is a measure of spatial overlap between the predicted masks ($\hat{Y}$) and the provided reference (Y), ranging from 0 (worst) to 1 (best). The definition from section 1.2.1.3 is repeated here for clarity:

$$DSC = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \tag{2.1}$$

Hausdorff distance (HD) is the largest distance between the boundary of the predicted segmentation and the boundary of the reference segmentation, so lower values are better. In the FeTS challenge, the 95$^{\text{th}}$ percentile (pc_{95}) of the HD is calculated, which is more robust to outlier pixels that can arise from noisy annotations:

$$HD_{95}(\hat{Y}, Y) = \max \left\{ \underset{\hat{y} \in \hat{Y}}{\text{pc}_{95}}\, d(\hat{y}, Y)\,,\ \underset{y \in Y}{\text{pc}_{95}}\, d(y, \hat{Y}) \right\}, \tag{2.2}$$

where $d(a, B) = \min_{a \in A} \|a - b\|$ is the distance of point a to the set of boundary points B.

Combining overlap-based and distance-based metrics is desirable in segmentation evaluation, as they are sensitive to different segmentation characteristics and hence complement each other (Maier-Hein et al. 2024).

2.1.4.2 Ranking

Beyond the purpose of benchmarking, a challenge is also an international competition and therefore needs official winners. Hence, a challenge ranking was computed based on the metric values for DSC and HD measured on all test cases for all evaluated algorithms (submissions by participating teams). The ranking method designed for the FeTS challenge tried to stay close to the ranking method of previous BraTS challenges while also taking into account the federated nature of the test data in the FeTS challenge, as well as its goal of benchmarking cross-institution generalization. Therefore, all algorithms were ranked per institution first using a rank-then-aggregate method (Maier-Hein et al. 2018; Wiesenfarth et al. 2021), similar to previous BraTS challenges, and then a consensus ranking between all institutions was calculated. Specifically, for each institution k among the K institutions, the algorithms were first ranked on all N_k test cases, considering the three tumor regions and two metrics. This resulted in $N_k \cdot 3 \cdot 2$ rankings for each algorithm at institution k. If an algorithm did not produce a valid segmentation for some test case (for example because

an error occurred during prediction), it was assigned the worst rank for the corresponding case-specific ranking. Next, the average rank across all test cases within institution k was computed, for each region-metric combination, which yielded $3 \cdot 2$ rankings per institution. The final rank of an algorithm was determined by averaging all $K \cdot 3 \cdot 2$ of its per-institution ranks. If there were ties in any ranking, they were resolved by assigning the minimum rank. Note that institutions with more test cases did not have a higher weight in this ranking method than smaller institutions. This was intentional, as institutions may have different distribution shifts, and submissions to the challenge should generalize to every institution equally.

2.1.5 Statistical Analyses

Every challenge ranking is based on a finite sample of test data, so it raises questions about whether the ranking differences between specific methods are statistically significant. Therefore, it is best practice to also analyze the stability of challenge rankings (Maier-Hein et al. 2018), which provides insights into how robust the ranking is to randomness in the data selection. For FeTS22, a bootstrap analysis was performed to produce blob plots (Wiesenfarth et al. 2021), which visualize the distribution of ranks each algorithm achieved across the bootstrap samples. In FeTS21, the significance map from (Wiesenfarth et al. 2021) was adapted by summing the number of significant comparisons for each algorithm pair across all sub-rankings, each of which is based on an individual institution-metric-tumor region combination. The one-sided Wilcoxon signed rank test at 5% significance level with adjustment for multiple testing according to Holm was employed for each comparison.

2.1.6 Technical Infrastructure

The main technical systems required by the FeTS challenge were:

- Challenge website. Required functionality: Convey information about the challenge to (potential) participants and provide a forum for answering questions.

- Submission platform. Required functionality: Allow participants to submit challenge contributions (the inference algorithms) and test their functionality.

- Federated evaluation system. Required functionality: Distribute the participants' submissions to the test data contributors' institutions, run the inference procedure, and collect results in the form of segmentation metrics. Figure 2.5 visualizes this federated evaluation workflow, which is a core component of the FeTS Challenge.

In the following, details of the system design and implementation are described. While the overall design was similar for FeTS21 and FeTS22, there were differences in the implementation details, which are briefly highlighted along the way.

Challenge website For FeTS21, a custom GitHub page (`https://fets-ai.github.io/Challenge/`) was used as an information portal, and GitHub discussions as a forum. In the following year (FeTS22), the Synapse platform (`https://www.synapse.org/fets`) served both purposes.

Submission platform Instructions on how to create a submission were provided through a GitHub repository (`https://github.com/FeTS-AI/Challenge/tree/main/Task_2`). Participants could then submit their inference code as containerized applications to a container registry. In FeTS21, Singularity images (Kurtzer et al. 2017) were pushed to a GitLab container registry (`https://gitlab.hzdr.de/`), while in FeTS22 Docker images (Merkel 2014) could be uploaded to the Synapse container registry. Each team could only submit once for the final test set evaluation, but the submission platform allowed them to debug the containers before that. In FeTS22, for instance, all submissions were checked in an isolated environment on cloud computing infrastructure at the DKFZ by the following procedure:

1. Convert the Docker image to a Singularity image file, as Docker was not supported by the IT departments of some data contributors.
2. Perform a compatibility test of the submission using the same evaluation software as in the testing phase. This consisted of evaluating the container on a small subset of the training data.
3. During the compatibility test, measure the GPU memory consumption and inference time. These were limited to 11 GB and 180 seconds per case, which was necessary to ensure that all submissions could be evaluated in the federation.
4. Report the results of the tests (metric values in case of success or error messages) back to the participants and, if successful, upload the Singularity image to cloud storage.

Federated Evaluation System In FeTS21, a custom Python script was integrated in the FeTS tool (Pati et al. 2022a), which had to be installed locally by the data contributors. In FeTS22, the MedPerf tool (Karargyris et al. 2023) was used for the evaluation. A benchmark in MedPerf comprises the steps of data preparation, inference, and evaluation, which were customized for the FeTS Challenge and implemented in an MLCube (`https:`

//github.com/mlcommons/mlcube). The inference step was modified by each participant to run their model. After installing the MedPerf client and dependencies locally, the benchmark was run at each institution and the results (metric values for each submission) sent back to the challenge organizers. Meta-data was transferred manually in FeTS22, but this can technically also be done when setting up the MedPerf client.

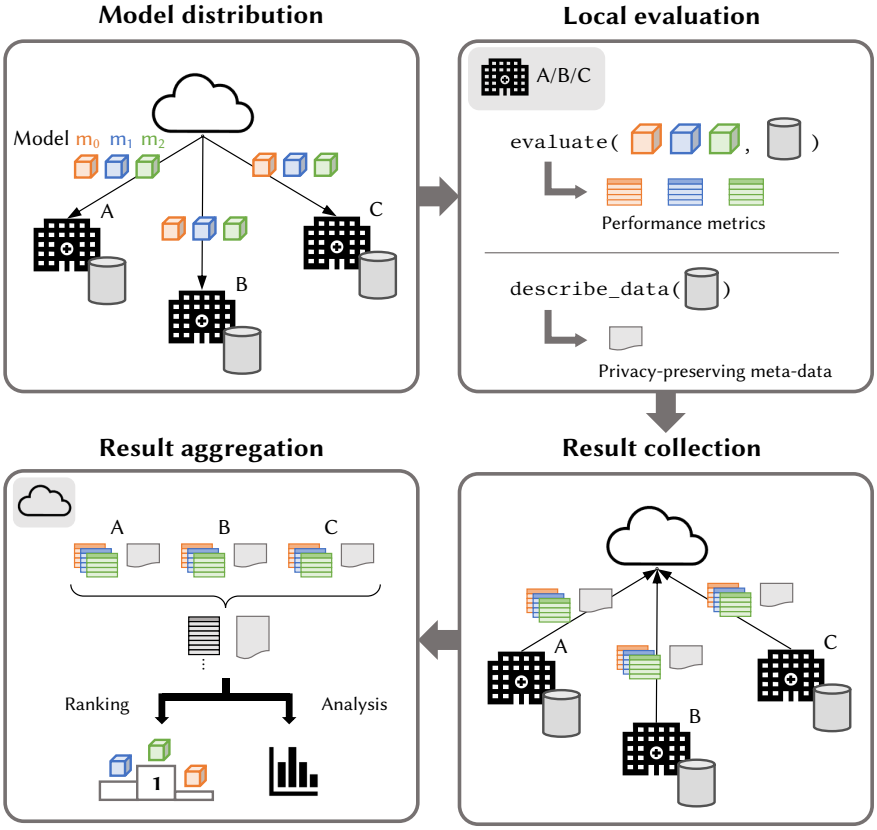

Figure 2.5: Diagram of the federated evaluation workflow used in the FeTS Challenge. The workflow consists of four steps: (1) segmentation models (colored cubes in the diagram) are sent from the benchmarking server (cloud) to the participating institutions (hospitals), which also hold the data (cylinders) for the test set. (2) At each institution, the local evaluation pipeline is run, which computes performance metrics for each model on each test case. Optionally, meta-data is collected in a privacy-preserving manner. (3) The metrics and meta-data results are sent back to the benchmarking server. (4) At the central server, the organizers combine the results from all institutions and perform a ranking of the evaluated models as well as additional analyses. Note that only three institutions and models are shown here as an example, but in the FeTS22 Challenge there were 32 hospitals and 41 models.

2.2 Failure Detection

Even robust automatic segmentation algorithms can fail on difficult cases or data with distribution shifts, so a mechanism to detect potential errors and notify the user is a crucial complementary component. This section focuses on automatic failure detection methods, as manual controls are expensive and do not scale to the continuously increasing amount of imaging data acquired. Since there is currently no standardized framework for evaluating failure detection methods, sections 2.2.1 and 2.2.2 revisit the task definition and propose an evaluation methodology to fill that gap. A benchmark is then designed by additionally selecting appropriate datasets (section 2.2.3), a self-configuring segmentation method (section 2.2.4), and diverse failure detection methods (section 2.2.5) from previous work. Details on the implementation and adaptations of individual methods are provided in the respective subsections.

2.2.1 Task Definition

Following the definition of failure detection for segmentation by Zenk et al. (2025b), consider a segmentation model, $f : \mathcal{X} \to \mathcal{Y}$, which generates a segmentation $\hat{y} = f(x)$ based on a three-dimensional image sample $x \in \mathbb{R}^{d_1 \times d_2 \times d_3}$. This study assumes failure detection is performed with a *confidence scoring function* (*CSF*) that provides a confidence score κ for each sample:

$$g : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}\,, \quad g(x, \hat{y}, f) = \kappa\,, \tag{2.3}$$

where $\mathcal{H}$ is the space of segmentation models and higher κ implies higher confidence in the prediction. Many CSFs only use a subset of these inputs, for example they can be independent of f and only use image and segmentation (x and $\hat{y}$). Other CSFs could even integrate f and g in the same model, but this is not the case for the methods studied in the remainder. Failure detection requires making a decision on whether to accept $\hat{y}$, which means in practice whether to use it for downstream analyses. The final model prediction is accepted if the confidence score exceeds a threshold τ:

$$g_\tau \circ f(x) = \begin{cases} f(x), & \text{if } g(x, f(x), f) \geq \tau \\ \text{abstain}, & \text{otherwise.} \end{cases} \tag{2.4}$$

This formulation interprets failure detection as a selective prediction task (Jaeger et al. 2022). An operating point for the rejection threshold τ can be found before testing using a calibration procedure as in (Geifman and El-Yaniv 2017).

2.2.2 Evaluation

To answer RQ 2.1 on page 29, this part describes the evaluation methodology for failure detection developed in this thesis. Failure detection aims to avoid risks associated with inaccurate segmentations. Here, a *risk function* $R(\hat{y})$ quantifies this risk. Higher risk values correspond to segmentations with more severe errors, but the concrete choice of risk function can vary depending on the application. One option could be, for example, a discrete risk function that assigns labels of "high", "medium", and "low" risk to predictions. In this study, however, segmentation metrics m were used as continuous risk functions, assuming that ground truth masks y are available, which is reasonable for benchmarking purposes. In the case that higher values of m indicate better segmentation quality and m is upper-bounded by 1 (for example the DSC score), the risk function can be defined as:

$$R(\hat{y}, y) = 1 - m(\hat{y}, y) \tag{2.5}$$

Note that ground truth data y was only required for evaluating the performance of failure detection methods in this thesis, but of course, the confidence scores produced by these methods did not depend on the ground truth.

2.2.2.1 Requirements and Pitfalls in Current Practice

An evaluation protocol for failure detection that measures progress in a realistic setting has to fulfill certain requirements, which were derived in this thesis based on theoretical considerations and related work. Furthermore, common evaluation pitfalls were identified, to highlight potential issues in current practice. The text in the following paragraphs, describing four requirements and pitfalls, was taken from (Zenk et al. 2025b). The original text was written by me and only slightly modified here to match the style of this thesis.

Requirement R1: Evaluate the failure detection task directly and allow comparison of all relevant solutions. Similar to how (Jaeger et al. 2022) argued for classification, a variety of proxy tasks for segmentation failure detection has been studied, each of them with their own metrics and restrictions, although failure detection is the commonly stated goal. To allow a comprehensive comparison and avoid excluding relevant methods, metrics are needed that summarize failure detection performance.
Pitfalls in current practice: A popular proxy task is out-of-distribution (OOD) detection (González et al. 2022; Graham et al. 2022). While OOD detection certainly is useful, it is not identical to failure detection. For example, when applying a segmentation model to a new hospital, all samples are technically OOD, but only some of them might turn out to be failures. Vice versa, in-distribution samples can also result in failures. Another

commonly studied task is segmentation quality estimation, which phrases failure detection as a regression task of segmentation metric values (Kohlberger et al. 2012; Valindria et al. 2017; Robinson et al. 2018; Liu et al. 2019; Li et al. 2022; Qiu et al. 2023). Although close to the task definition in section 2.2.1, it is slightly more restrictive, as confidence scores need to be on the same scale as the risk values. This kind of "calibration" can be desirable for some applications or to compute metrics like mean absolute error (MAE), but failure detection only requires a monotonous relationship between risk and confidence and the evaluation should not be restricted to methods that output segmentation metric values directly.

Requirement R2: Consider both segmentation performance and confidence ranking. Following Jaeger et al. (2022), in practice arguably the performance of the whole segmentation system matters, i.e., segmentation model and CSF. A desirable system has low remaining risk after rejection based on thresholding the confidence score, which can be achieved through (a) the CSF assigning lower confidence to samples with higher risk, i.e., better confidence ranking, or (b) avoiding high risks in the first place, i.e., better segmentation performance. These aspects cannot be easily disentangled, because the CSF might require architectural modifications that adversely impact segmentation performance, such as the introduction of dropout layers. The evaluation metric should hence consider both aspects. Beyond the choice of metric, this requirement also implies that a fair comparison between failure detection methods uses the same segmentation model for different CSFs, if possible.
Pitfalls in current practice: Most related works use metrics that ignore the segmentation performance aspect and focus on confidence ranking (Robinson et al. 2018; Liu et al. 2019; Jungo et al. 2020; Li et al. 2022), such as area under the receiver operating curve using binary failure labels (AUROC_f) and Spearman correlation coefficient (SC). As a side effect in the case of continuous risk definitions, exclusively considering confidence ranking while neglecting absolute risk differences can also lead to unexpected evaluation outcomes. Consider an example with four test samples resulting in risk values of $\{0.1, 0.5, 0.7, 0.72\}$ and perfect confidence ranking, i.e., the first sample has the highest confidence and so on. Switching confidence ranks between the first two samples has the same effect on the Spearman correlation as switching the ranks of the last two, but the first switch is more problematic from a failure detection perspective. This issue is, for example, relevant in scenarios where there is a group of test samples with similar, low risks and a smaller number of samples with higher, more variable risks, which is likely to happen in a failure detection scenario where failures are rare.

Requirement R3: Support flexible risk definitions. In contrast to image classification, there is no universal definition of what makes a segmentation faulty. The risk function depends ultimately on the specific application and can in particular be continuous. Therefore, a general evaluation protocol for failure detection, as required for a benchmark, should be flexible enough to support different choices.
Pitfalls in current practice: Several papers use a threshold on the Dice score to define failure (DeVries and Taylor 2018; Chen et al. 2020; Jungo et al. 2020; Lin et al. 2022; Ng et al. 2023), resulting in a binary risk function, which is reasonable if the specific application has a natural threshold. For many existing datasets, however, such a threshold cannot be determined easily, for instance when inter-annotator variability is unknown. In these cases, a continuous risk function like the (negative) DSC can avoid information loss and discontinuity effects. Hence, a general-purpose evaluation metric should be applicable to both discrete and continuous risk functions, which is not given for some popular metrics like area under the receiver operating curve (AUROC) for binary failure labels.

Requirement R4: Consider realistic failure sources. CSFs should be primarily judged on how successful they are in detecting realistic failures. These can happen for numerous reasons, but distribution shifts in data from different scanners and populations are especially important, as they are likely to be encountered in real-world applications. The data used for evaluating CSFs should hence reflect these failure sources, ideally covering different types of dataset shifts.
Pitfalls in current practice: While earlier works focused on in-distribution testing (DeVries and Taylor 2018; Chen et al. 2020; Jungo et al. 2020), there has been a development towards including test datasets from different centers or scanners in the evaluation (Mehrtash et al. 2020; González et al. 2022; Li et al. 2022; Ng et al. 2023). Some studies augment their test dataset with "artificial" predictions that are not produced by the actual segmentation model, for example by corrupting the segmentation masks or using auxiliary (weaker) segmentation models (Robinson et al. 2018; Li et al. 2022; Qiu et al. 2023). While this practice has the benefit of testing the CSF on a wide range of segmentation qualities, it may not be ideal for a benchmark on failure detection: Firstly, it contradicts R1, because only methods can be tested on the artificial test data that are independent of the segmentation model, excluding lines of work like ensemble uncertainty (Lakshminarayanan et al. 2017) or posthoc (González et al. 2022) methods, although they are usually applicable in failure detection scenarios. Secondly, the additional samples might put more emphasis on failure cases that would never occur in a realistic setting, decreasing the influence of practically relevant cases in the evaluation. It is important to note that realistic artificial images, unlike artificial predictions, can circumvent these drawbacks and meet requirement R4.

Table 2.3: Metric candidates for segmentation failure detection compared from the perspective of the requirements (R1–R3) from section 2.2.2.1. Segmentation performance is only taken into account by AURC, which also fulfills all other criteria. Abbreviations: area under the receiver operating curve using binary failure labels ($AUROC_f$), mean absolute error (MAE), Pearson correlation coefficient (PC), Spearman correlation coefficient (SC), area under the risk-coverage curve (AURC). Table adapted from (Zenk et al. 2025b).

Metric	Required confidence scale (R1)	Considers confidence ranking (R2)	Considers segmentation risk (R2)	Compatible with binary/continuous risk (R3)
$AUROC_f$	ordinal/real	yes	no	yes/no
MAE	same as risk	no	no	no/yes
PC	real	implicitly	no	yes/yes
SC	ordinal/real	yes	no	yes/yes
AURC	ordinal/real	yes	yes	yes/yes

2.2.2.2 Evaluation Protocol

Different protocols and metrics have been used in previous work to evaluate error detection or related tasks like segmentation quality estimation. To find out which metrics are most appropriate for the failure detection task definition from section 2.2.1, the most common metrics are compared in table 2.3 with respect to the requirements R1–R3 from section 2.2.2.1. The risk-coverage analysis (El-Yaniv and Wiener 2010), summarized by the scalar performance metric area under the risk-coverage curve (AURC), is the only one that fulfills all requirements and was hence used as the main evaluation protocol. Jaeger et al. (2022) recently recommended AURC for evaluating failure detection methods in the context of image classification. This thesis extended its application to semantic segmentation tasks.

In the benchmarking experiments from section 3.2, the DSC was used as the risk function, averaging it over all K foreground classes of a dataset:

$$R(\hat{y}, y) = 1 - \sum_{k=1}^{K} \mathrm{DSC}(\mathbb{I}(\hat{y} = k), \mathbb{I}(y = k)) \tag{2.6}$$

Here, $\mathbb{I}$ denotes the indicator function, transforming the multi-class label map into a binary map for each class. In section 3.2.3, the normalized surface dice (NSD) (Nikolov et al. 2021) is used instead of DSC to study the impact of the chosen risk function. NSD measures deviations from the predicted segmentation boundary to the ground truth, while DSC focuses on the volumetric overlap. Each experiment in the benchmark yielded a risk score

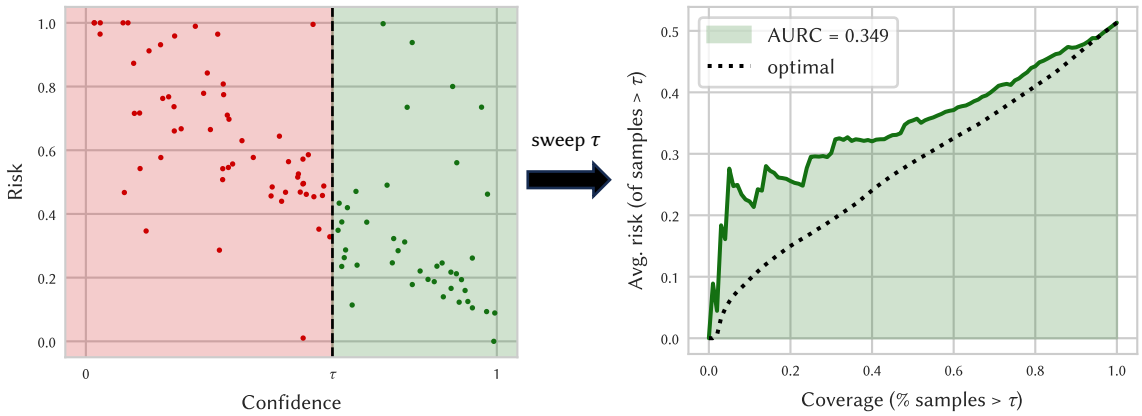

Figure 2.6: Example for the risk-coverage curve based on synthetic data. Left: The simulated experimental data consists of confidence scores and risks for each sample. For each possible confidence threshold, predictions for samples above the threshold would be accepted (green), while for low-confidence samples the model abstains from outputting a prediction (red). Right: Measuring the selective risk for each confidence threshold results in a risk-coverage curve, which can be summarized with the AURC. Figure adapted from Zenk et al. (2025b).

r_i and a confidence score κ_i for each test case x_i $(i = 1, \ldots, N)$. The risk-coverage curve was obtained by varying a confidence threshold τ, which would in practice determine which samples are considered erroneous. For each threshold, the selective risk R_s and the coverage C were measured:

$$R_s(\tau) = \frac{\sum_{i=1}^{N} r_i \cdot \mathbb{I}(\kappa_i \geq \tau)}{\sum_{i=1}^{N} \mathbb{I}(\kappa_i \geq \tau)}, \tag{2.7}$$

$$C(\tau) = \frac{1}{N} \sum_{i} \mathbb{I}(\kappa_i \geq \tau) \tag{2.8}$$

Intuitively, R_s quantifies the remaining risk after removing low-confidence samples, while C represents the fraction of retained samples. Failure detection is a trade-off between these two. To illustrate the risk-coverage curve, fig. 2.6 shows an example computed from artificial experiment results.

The area under the risk-coverage curve (AURC) can be used to summarize a whole curve into a scalar value, which facilitates comparing multiple methods concisely. Intuitively, it corresponds to the average selective risk (for example the average $1 - \text{DSC}$ risk) across varying confidence thresholds. In general, lower AURC values indicate better performance. There are two natural reference values for AURC: For a method that assigns

random confidence scores, the AURC is identical to the average overall risk, computed as $\text{AURC}_{\text{rand}} = \sum_i r_i / N$. In contrast, an optimal confidence method results in scores that perfectly order the risk values, meaning $\kappa_i \leq \kappa_j \Rightarrow r_i \geq r_j \; \forall \; i, j$. The corresponding optimal AURC can be determined computationally. AURC was implemented for the experiments (section 3.2) by adapting the publicly available implementation from (Jaeger et al. 2022) to the segmentation task, such that it is compatible with the continuous risk function. While AURC is the main metric in this benchmark, the effect of using alternative metrics like SC or PC will also be investigated in section 3.2.3.

2.2.3 Datasets

The criteria guiding the dataset selection for the benchmark were:

1. Public availability: This guarantees that the results can be reproduced by others and new work can build upon this benchmark.

2. 3D imaging modalities: class imbalance and higher hardware requirements make these images particularly challenging to segment.

3. Distribution shifts between training and test set: In practice, errors occur more frequently when distribution shifts are present, so a realistic evaluation scenario should incorporate them (R4 from section 2.2.2.1).

4. Previous work: existing results from previous works on datasets that meet the criteria above are helpful in discussing this study's results.

In total, six datasets with computed tomography (CT) and magnetic resonance imaging (MRI) scans were used to compare failure detection methods. Three additional datasets with different imaging modalities were studied to test the generalization capabilities of the methods beyond CT/MRI. Training and test sets were created from the original datasets by splitting them on a patient level, so that all images of one patient (if there are multiple) are guaranteed to be in the same split. Details on the number of training and test cases are provided in table 2.4, and a small set of CT/MRI example images is depicted in fig. 2.7. Additional data samples are printed in appendix A.2. For model development, each training set was further subdivided in a cross-validation manner into five folds, which use 20% of the cases for validation and 80% for training. All datasets are described below, with a short reason for inclusion in the benchmark. The following paragraphs (six dataset descriptions) are taken from Zenk et al. (2025b). They have been originally written by me and were slightly modified here to match the style of this thesis.

Table 2.4: Summary of datasets used in this study. The #Testing column contains case numbers for each subset of the test set separated by a comma, starting with the in-distribution test split and followed by the shifted "domains". The number of classes includes one count for background. Datasets below the separating line are not part of the main benchmark; instead, they were included to explore how the insights on CT/MRI transfer to other modalities. Table adapted from (Zenk et al. 2025b).

Dataset	#Classes	#Training	#Testing	Modality	Shift in test set
Brain tumor (2D)	2	939	313, 313 × 4	MRI	4 Artificial corruptions
Brain tumor	4	235	50, 50	MRI	Higher prevalence of LGG
Heart	4	190	60, 190, 100, 100	MRI	Unseen scanner vendors
Prostate	2	26	6, 30, 19, 13, 12, 12	MRI	Unseen institutions
Covid	2	160	39, 50, 20	CT	Unseen institutions
Kidney tumor	4	367	122	CT	–
Echocardiography	3	80	55	3D-US	–
Retinal fluids	4	34	6, 6, 24	OCT	Unseen scanner
Optic cup/disc (2D)	3	640	160, 101, 159	RGB	Unseen institutions

Brain Tumor (2D)

Despite being 2D, this simplified version of the FeTS 2022 dataset (Menze et al. 2015; Bakas et al. 2017; Zenk et al. 2025a) was included in the benchmark, as it allows for quick experimentation. Data preparation for each case consisted of cropping the original images around the brain, selecting only the axial slice from the 3D images with the largest tumor extent, and resizing that slice to 64 × 64 pixels. Each case has four MRI sequences: T1, T1-Gd, T2, FLAIR. All publicly available cases were split randomly into a training and a test set. To introduce shifts in the test set, four artificial corruptions were applied to each test case using the `torchIO` library (Pérez-García et al. 2021), producing four additional corrupted versions per test case: affine transforms, bias field, spike and ghosting artifacts. Due to the low image resolution, only the whole tumor region was used as a label for this dataset.

Brain Tumor

The BraTS 2019 dataset (Menze et al. 2015; Bakas et al. 2017; Bakas et al. 2019) contains information about the tumor grade (high-grade glioma (HGG), or low-grade glioma (LGG)) for each training case. To simulate a population shift with more LGG cases during testing, all publicly available cases were split into a training and a test set, such that there are 167 HGG and 26 LGG cases in the training set and 50 cases for each grade in the testing set. Note that LGG cases are often harder to segment (Bakas et al. 2019). Each case consists of four MR sequences (T1, T1-Gd, T2, FLAIR). The labels for this dataset are nested tumor

regions: WT, TC, and ET. The hierarchical ordering is ET $\subseteq$ TC $\subseteq$ WT. A similar dataset is used in (Hoebel et al. 2022), but here a few LGG cases were included during training to make the setup more realistic.

Heart

The M&Ms dataset (Campello et al. 2021) provides short-axis MRI data from four scanner vendors. For the training set, only samples from vendor B were included, while the testing set comprised 30 patients from vendor B and data from the remaining three vendors. Each patient contributes two images, corresponding to the end-diastolic and end-systolic phases, respectively. The labels in this dataset include the left ventricle, the right ventricle, and the left ventricular myocardium. Although this dataset was also employed in (Kushibar et al. 2022), a different data split was applied here, motivated by the findings of (Full et al. 2020), which indicate that generalization from vendor B to vendor A poses the greatest challenge.

Prostate

This dataset is a collection of two data sources: The prostate dataset from the Medical Segmentation Decathlon (Simpson et al. 2019; Antonelli et al. 2022) was employed for training and in-distribution testing. Additional testing data was provided by (Liu et al. 2020), which includes data prepared from (Bloch et al. 2015; Lemaître et al. 2015; Litjens et al. 2023). Segmentation performance was evaluated using the T2 MRI sequence, focusing solely on the whole prostate label. This setup aligns with the approach in (González et al. 2022), with one difference: The 'RUNMC' institution in (Liu et al. 2020) also contributed data to the training set (Simpson et al. 2019; Antonelli et al. 2022), so it was excluded from this benchmark to avoid duplicates.

Covid

This dataset is a collection of three data sources: The COVID-19 CT Segmentation Challenge dataset (Clark et al. 2013; An et al. 2020; Roth et al. 2022) was split into 39 cases for testing and 160 for training. Additional test cases were drawn from (Morozov et al. 2020) and (Jun et al. 2020). The dataset includes a single foreground label representing lesions associated with COVID-19. This setup followed the configuration from (González et al. 2022).

Kidney Tumor

The publicly available CT scans and annotations from the KiTS23 dataset (Heller et al. 2021; Heller et al. 2023) were randomly divided into training and test sets. While the test

set does not include an explicit distribution shift, it contains a sufficient number of difficult cases suitable for evaluating failure detection. The same three nested regions as defined in the challenge were used as labels: kidney + cyst + tumor, cyst + tumor, and tumor.

Non-CT/MRI Datasets

This benchmark's main focus was on CT and MRI imaging data, as these are the most common modalities used in medical image segmentation (Maier-Hein et al. 2018). To investigate how the insights on CT/MRI transfer to other modalities, the following additional datasets were explored:

- Echocardiography (Carnahan et al. 2021): Published through the MVSeg Challenge 2023, this dataset contains 3D-ultrasound (US) images of the heart and annotations for the two mitral valve leaflets. All scans were acquired at the same hospital and with the same scanner model, so there is no distribution shift at test time and cases were split randomly between training and test set.
- Retinal fluids (Bogunović et al. 2019): Optical coherence tomography (OCT) is a popular modality in ophthalmology to image the retina. The RETOUCH challenge provided this dataset, which was acquired from three different scanners. Two scanners were included in the training set, while the third was only part of the test set. Three types of retinal fluids were annotated in the images, which have high inter-annotator variability (mean DSC = 0.73 according to Bogunović et al. (2019)).
- Optic disc/cup (2D) (Fumero et al. 2011; Sivaswamy et al. 2015; Orlando et al. 2020): As an additional 2D dataset, this collection of three fundus photography RGB-image datasets was used in previous work on domain generalization (Wang et al. 2020b). The segmentation targets are two nested regions: optic cup $\subset$ optic disc. Models were trained only on one of the subsets (Orlando et al. 2020) and the other two were unseen before testing.

Example images are shown in fig. 2.8. Results on these additional datasets are reported in section 3.2.2, while results for the CT and MRI datasets can be found in section 3.2.1.

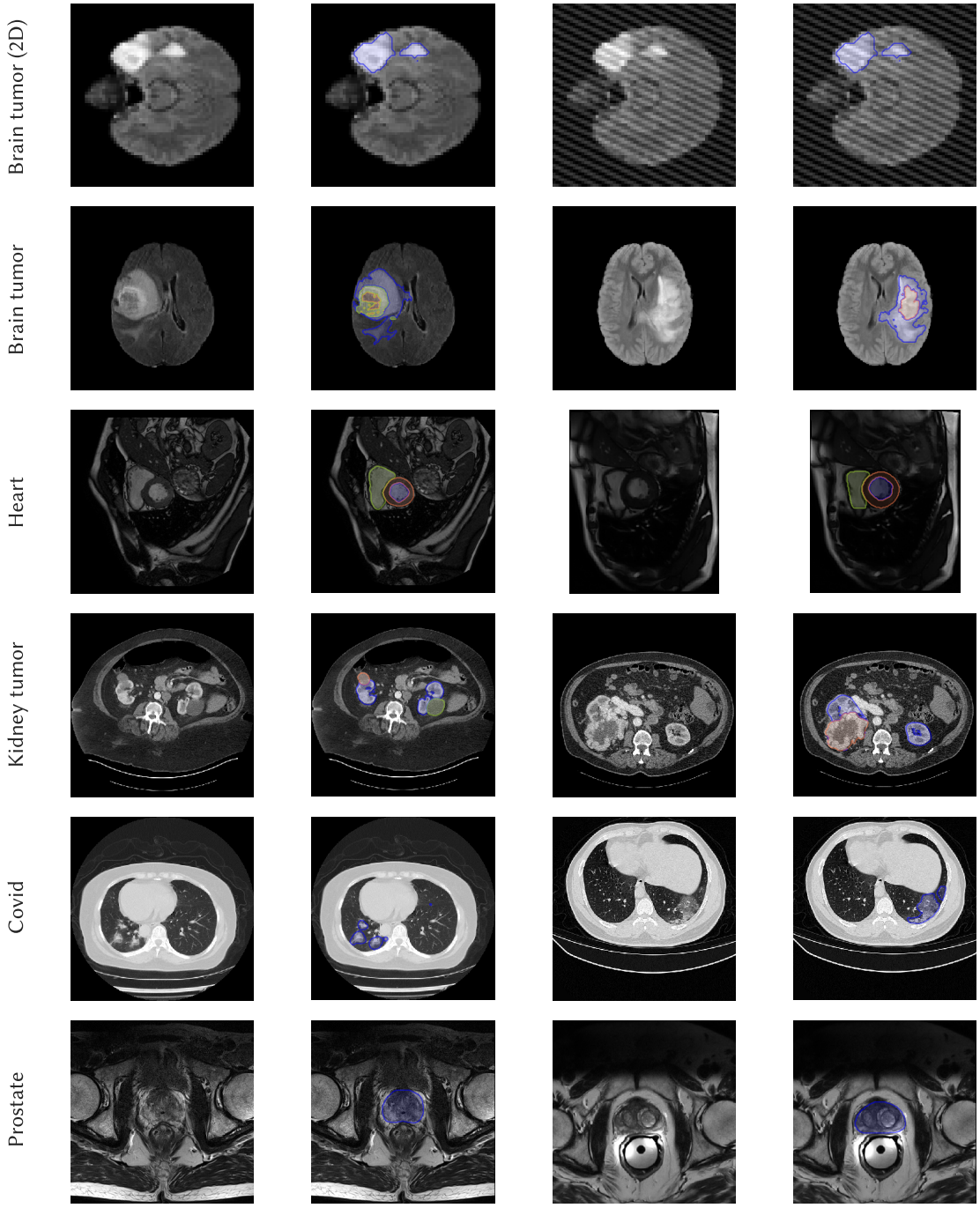

Figure 2.7: Example images from the CT/MRI datasets. Two pairs of (image, ground truth) from the test set are shown per dataset. The left pair is an in-distribution sample while the right pair is a sample with distribution shift. For the kidney tumor dataset, there is no distribution shift. Consistent windows were used for the CT datasets.

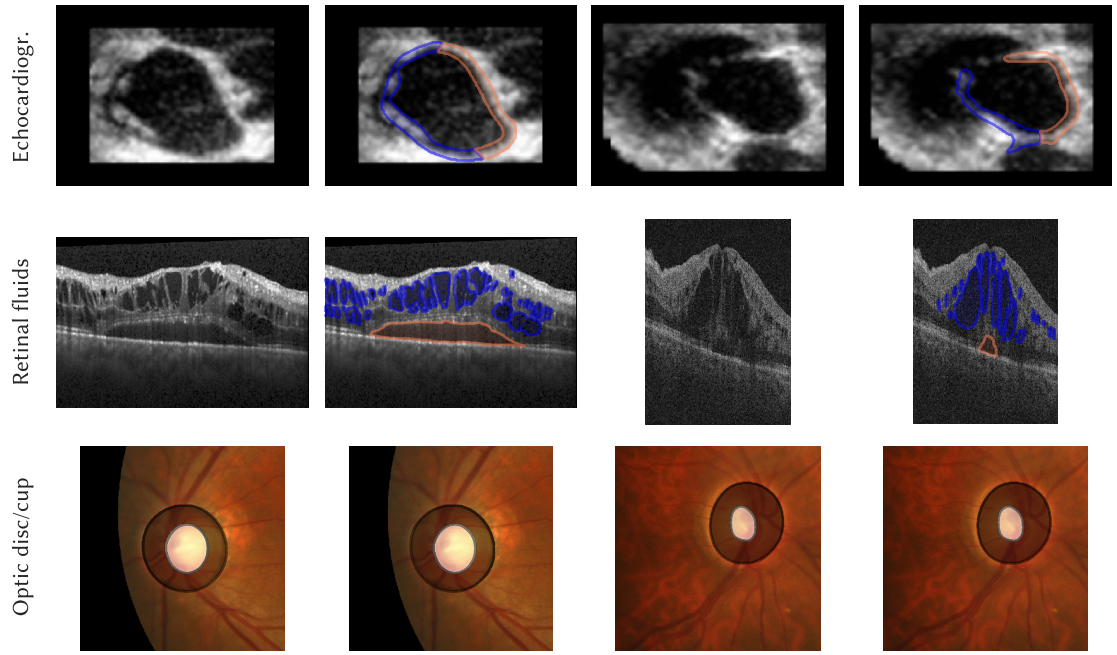

Figure 2.8: Example images from the non-CT/MRI datasets used for the benchmark. Two pairs of (image, ground truth) from the test set are shown per dataset. The left pair is an in-distribution sample while the right pair is a sample with distribution shift. For the echocardiography dataset, there is no distribution shift.

2.2.4 Segmentation Algorithm

A system capable of segmentation failure detection consists of a segmentation algorithm and a failure detection method. For some failure detection methods, modifications on the segmentation algorithm are required; these are detailed in section 2.2.5. Here, the default segmentation model is described.

2.2.4.1 Preprocessing

Image and segmentation preprocessing was performed using the nnU-Net library (Isensee et al. 2021a). Their pipeline includes the following steps:

1. Cropping to the nonzero image part, to remove irrelevant background.

2. Intensity normalization. By default, each sample and imaging modality is normalized independently. The mean and standard deviation across all voxels are computed to normalize all intensities x by $(x - \mu)/\sigma$ (z-score normalization). For CT images,

nnU-Net clips intensities to the 0.5 and 99.5 percentiles of the foreground intensity distribution before applying z-score normalization. The percentiles, mean and standard deviation are computed across the entire dataset in this case.

3. Resampling to a common image spacing. For each spatial image dimension, the median spacing is used as the target spacing. Images are by default interpolated with third-order splines, and segmentations with linear interpolation (plus argmax). Some special cases with strong anisotropies are handled differently; details can be found in (Isensee et al. 2021a).

For the brain tumor 2D dataset, only z-score normalization was used, as it was created from 3D data using a data preparation pipeline that already included cropping and resampling.

2.2.4.2 Network Architecture and Training

The segmentation networks employed in this study are based on the U-Net architecture (Ronneberger et al. 2015). Below, the architecture and training procedure are described in detail, and a summary of hyperparameters is provided in table 2.5.

The default U-Net implementation contains multiple stages, each of which consists of a convolution block containing two convolutional layers, followed by instance normalization (Ulyanov et al. 2017) and a leaky rectified linear unit (ReLU) activation function. In an ablation study (section 3.2.2.2), residual connections (He et al. 2016) were added to each convolution block, and the number of blocks per stage increased to three. Both network configurations are visualized in fig. 2.9. Within each stage, the stride and kernel size for the convolutions were adapted to the dataset automatically, based on the spacing and size of image patches fed into the network. The number of output channels was set to the dataset-specific number of classes. Dropout was applied selectively at the end of each stage for the five innermost stages, with a dropout rate of 0.5, following Kendall et al. (2016). This setup facilitated the use of test-time Dropout, without excessively regularizing the network. For the brain tumor 2D dataset, a U-Net implementation from the MONAI framework (Cardoso et al. 2022) was used, with 5 stages, two residual units per stage, and a dropout rate of 0.3 after each convolutional layer.

As large volumetric images are problematic to fit into limited graphics processing unit (GPU) memory, patches (meaning smaller cutouts) were extracted from the preprocessed data using nnU-Net's data loading functionality and put into the network. The patch size was configured automatically. During training, the sum of the Dice loss and cross-entropy (CE)* was employed as the loss function, and the networks were optimized until

*For some datasets (brain tumor, kidney tumor, and optic disc/cup), the target structures are hierarchical,

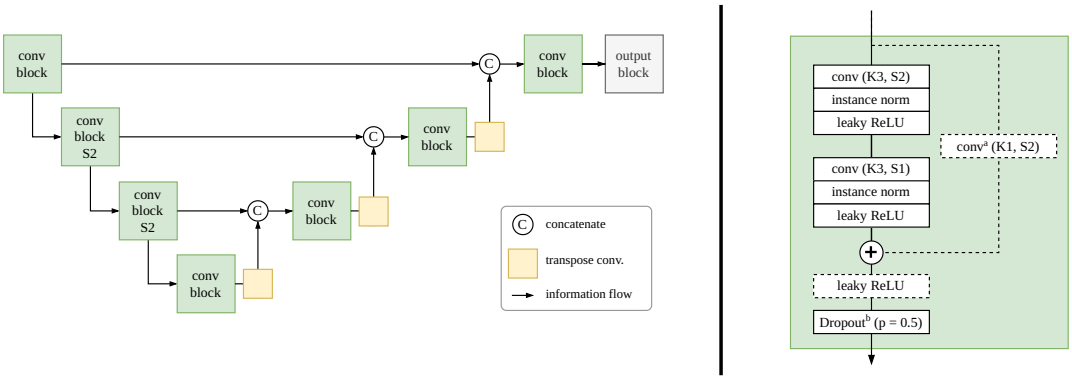

Figure 2.9: Schematic of the segmentation network backbone (U-Net). Left: Overview of the network, with the usual stride-2 convolutions (S2) for downsampling, transposed convolutions for upsampling, and skip connections. The output block consists of a single convolutional layer with kernel size 1 (K1). The depth of the network depends on the dataset; here, a network with four stages is shown. On the right, a detailed view of the convolution blocks is depicted. The dashed lines indicate components that are only present for residual blocks, not in the default convolution block. Two special cases are described in this figure with superscripts: [a] This convolutional layer is only present in residual blocks that change the feature dimension. [b] Dropout is activated only in the lowest five U-Net stages.

the validation loss plateaus using the stochastic gradient descent (SGD) optimizer with Nesterov momentum. The batch size was maximized for each dataset given the available GPU memory. Data augmentations included rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction, and mirroring, except for the brain tumor (2D) dataset. For this special case, only mirroring augmentations were used to make the network susceptible to artificial corruptions.

2.2.4.3 Inference Procedure

Test-time prediction (also called inference) was conducted using the dataset-specific preprocessing parameters from the respective training set. For datasets with 3D images, sliding window inference was performed on individual test cases. Patches were sampled from the preprocessed images in a sliding window manner with an overlap of 50%, and predictions for all patches were averaged using a Hann-window weighting scheme (Pérez-García et al. 2021). This approach assigns greater weight to central voxels in the patch

overlapping regions. In this case, sigmoid functions were applied after the network's output layer instead of softmax and the *binary* CE was used in the loss.

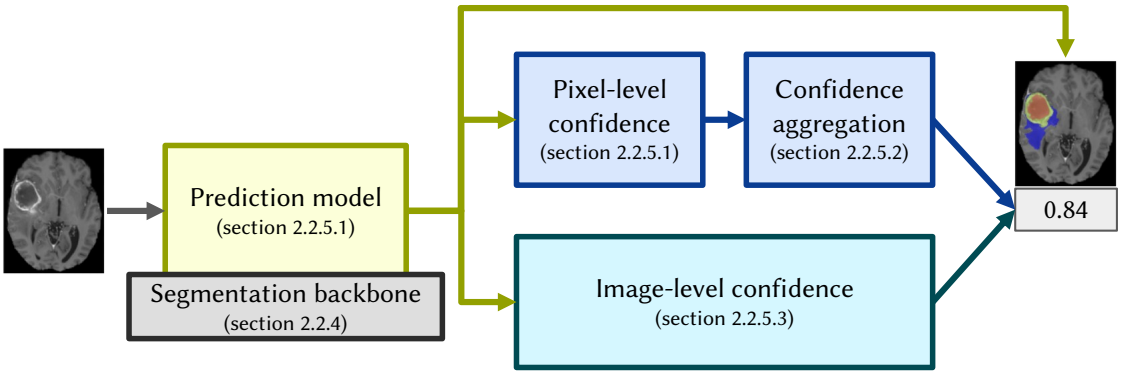

Figure 2.10: Overview of the components involved in failure detection. The prediction model is based on an exchangeable segmentation backbone and generates segmentation probabilities and masks. These are then used by the failure detection methods to produce confidence scores, which can be derived either directly at the image level or by aggregating pixel-level confidence maps. This pipeline structure follows Kahl et al. (2024) and extends it by image-level methods. Table 2.6 summarizes all confidence aggregation and image-level confidence methods implemented for the benchmark. The depicted image, mask and confidence score are not from real experiments but created just for illustration purposes.

while reducing the influence of boundary voxels, thereby mitigating edge artifacts. In the case of 2D datasets, inference was completed with a single forward pass through the network.

2.2.5 Failure Detection Methods

In the formulation from section 2.2.1, failure detection involves outputting a confidence score for each sample, which can be used to filter the model predictions. This can be implemented on a high level either by producing these (image-level) scores directly based on the complete case data or by first generating a pixel confidence map, which is aggregated in a second step to the confidence score. Figure 2.10 gives an overview of the failure detection method components described here and shows how this translates to the structure of this section. All methods in this benchmark are combinations of the components described here. The main goal behind which methods to include in the benchmark was to evaluate a wide range of approaches applicable to the failure detection task, while also considering the popularity in related literature and the feasibility of re-implementation, which may be limited by method complexity or the lack of a public official source code. A comprehensive discussion of the method selection is included in section 4.2. All methods were implemented in PyTorch (Paszke et al. 2019).

Table 2.5: Hyperparameters for the segmentation and failure detection methods used in the failure detection benchmark. Abbreviations: CC: connected components, (B)CE: (binary) cross-entropy. Table adapted from (Zenk et al. 2025b).

Method	**Parameter**	**Default value**	**Changes for brain tumor 2d**
		Segmentation	
U-Net	loss function	Dice + (B)CE	Dice
	optimizer	SGD + momentum (0.99)	AdamW
	learning rate	0.01	0.001
	learning rate decay	polynomial (exponent 0.9)	-
	weight decay	0.00003	0.00001
	batch size	2 (heart: 4)	32
	normalization layer	instance	batch
		Pixel confidence	
MC-Dropout	Number of samples	10 (kidney: 5)	-
Deep ensemble	Number of samples	5	-
		Pixel confidence aggregation	
Non-boundary-weighted	boundary width	4	
Patch-based	patch size	10^3	10^2
RF (simple features)	boundary width	4	
	connectivity for CC	3	2
		Image-level failure detection	
Quality regression	loss function	L2	
	optimizer	AdamW	
	learning rate	0.0002	
	learning rate decay	cosine	
	weight decay	0.0001	
	batch size	2 (heart: 4)	32
Mahalanobis	Max. feature dim.	10000	
VAE	loss function	BCE + $\beta \cdot$ KL-div.	
	β	0.001	
	optimizer	Adam	
	learning rate	0.0001	
	learning rate decay	-	
	batch size	6	32

2.2.5.1 Prediction Model and Pixel-level Confidence

Kahl et al. (2024) distinguish between the *prediction model*, which is responsible for the final predicted pixel-wise class scores, and the *uncertainty measure*, which computes a pixel-wise confidence map based on the prediction model's output. Importantly, the prediction model can also be used by image-level failure detection methods (section 2.2.5.3). Here, three different prediction models from this benchmark are presented, along with the uncertainty measures applied to each of them.

Single network This simple baseline uses the output of a single U-Net as the prediction. Confidence maps are computed from softmax probabilities through one of two uncertainty measures: the predictive entropy (PE) and the maximum softmax per pixel, defined in background section section 1.2.3.

MC-Dropout (Gal and Ghahramani 2016) This prediction method generates multiple samples by activating dropout layers at test time. Inference is repeated M times (with different active subnetworks randomly sampled through dropout), which results in M softmax maps. For this study, the default is $M = 10$; only for the kidney tumor dataset it is reduced to $M = 5$, due to the long inference times required for these large image volumes. All M probability maps are averaged to obtain the final prediction, and a pixel confidence map is computed with one of two uncertainty measures: either the PE of the probabilities averaged across samples or the mutual information (MI) of the whole sample of probabilities.

Deep ensemble (Lakshminarayanan et al. 2017) This method also generates multiple prediction samples but uses an ensemble of networks instead of dropout. In this benchmark, the ensemble consists of five networks trained with different random seeds on the same data. In contrast to Lakshminarayanan et al. (2017), adversarial training is not used for simplicity, following previous application to medical images (Mehrtash et al. 2020). As for MC-Dropout, the pixel-wise mean over the output probabilities of all ensemble members serves as the final prediction, and confidence maps are computed with the PE and MI uncertainty measures.

Some datasets (brain tumor, kidney tumor) have overlapping regions as outputs, so the segmentation network probabilities were produced by a sigmoid function for each region. To derive a single-channel confidence map from this prediction, a confidence map for each region was computed first, using the uncertainty measures described above, and these were aggregated across regions by taking the minimum confidence per pixel.

2.2.5.2 Aggregation of Pixel-level Confidence

The previous section described methods that output a real-valued pixel confidence map in addition to the actual segmentation. To decide whether the predicted segmentation as a whole should be accepted or flagged as potentially erroneous by the failure detection system, aggregation methods are required, which take in the pixel confidence map and output a scalar confidence score. Each implemented method is described below and also summarized in table 2.6. Detailed hyperparameters are listed in table 2.5.

Mean This naive baseline simply averages all pixel confidence values.

Foreground mean Identical to mean above, but averages only pixel confidence values inside the predicted foreground region.

Non-boundary-weighted (Jungo et al. 2020) Often, the model confidence is low at the boundaries between two predicted classes, because of inherent uncertainty in the exact boundary location. Jungo et al. (2020) and Kahl et al. (2024) observed that this can result in an undesired correlation between mean confidence and object size, as the boundary is longer for larger objects. To fix this, the boundary pixels are excluded from confidence aggregation. Specifically, a region around the boundary (4 pixels wide) is removed from the confidence map before computing the mean.

Patch-based (Kahl et al. 2024) This method shares the same motivation as the non-boundary-weighted aggregation, but uses a different approach to avoid the object size bias. First, confidence scores are computed in a sliding-window manner for patches of a fixed size, by averaging confidence across pixels in the patch. As in (Kahl et al. 2024), a patch size of 10^D was used for D-dimensional images The resulting patch-level confidence scores are aggregated by using the minimum confidence among all patches as the final confidence score, based on the idea that any low confidence region may indicate an error.

The methods explained so far are simple and computationally inexpensive, but this may limit their ability to aggregate complex confidence map patterns. An alternative are learning-based aggregation methods that require training a model on suitable data. In this benchmark, two such methods are included:

Regression forest (RF) trained on radiomics features (Jungo et al. 2020) The idea behind this method is to learn the relation between the confidence map and a segmentation risk (in the sense of section 2.2.1) through supervised learning. Jungo et al. (2020) proposed to extract hand-crafted radiomics features (Gillies et al. 2015) from the confidence map and

use the DSC scores for the associated prediction to fit a regression forest (RF). Radiomics features are usually computed on a region of interest (ROI) in the image (for example, on a lesion), but for this application they are determined by thresholding the confidence map. The threshold is tuned on a validation set, as described by Jungo et al. (2020). In this study, the feature extractor implemented in the pyradiomics library (Griethuysen et al. 2017) is used with default parameters, and z-normalization is applied. If individual radiomic features are not defined in special cases for confidence maps, missing values are imputed using the 'mean' strategy. Regression forests are then fitted using the scikit-learn library (Pedregosa et al. 2011), with the DSC for each class and the generalized DSC (Crum et al. 2006) as targets. The final confidence score is the mean over the estimated per-class DSCs.

Regression forest (RF) trained on simple features (Zenk et al. 2025b) Developed within the scope of this thesis, this is a simplification of the previous method. It does not require the definition of a ROI and replaces radiomics features with five simpler, hand-crafted features:

1. Mean confidence in the predicted foreground
2. Mean confidence in the predicted background
3. Mean confidence in the boundary region
4. Foreground size relative to the whole image size
5. Number of connected components in the predicted foreground

Each of these features intuitively contains information about important confidence map or prediction characteristics, which may be helpful to estimate the prediction quality. The configuration of the RF is identical to the method trained on radiomics features.

To fit regression forests on simple or radiomics features, a training set with examples of (prediction, confidence map, target DSC score) is required. This requirement is in practice not easily fulfilled for medical imaging datasets with notoriously small size. In this benchmark, the cross-validation outputs of running the inference procedure for pixel confidence methods on each validation split were used to create the required training data for the regression forests while leveraging all available samples.

2.2.5.3 Image-level Confidence

Confidence scoring methods that do not require pixel confidence maps are also valid approaches for the failure detection task. Four methods were implemented for this study (table 2.6 gives an overview and table 2.5 lists detailed hyperparameters):

Pairwise DSC (Roy et al. 2019) This approach requires a segmentation model to output a set of $M > 1$ plausible predictions for each example. The DSC is computed between each pair of predictions, and the final confidence score is defined as the mean of all pairwise DSC values. MC-Dropout or deep ensembles are used in this study to produce M discrete segmentation masks. As DSC is computed for each class separately, another average over classes is necessary for datasets with multiple foreground classes to arrive at image-level confidence scores.

Quality regression (Robinson et al. 2018) Failure detection is closely related to estimating the value the risk function takes on for unseen examples. As deep neural networks can learn to approximate arbitrary functions, they can also be trained to predict the segmentation risk (or quality) directly for a given image and prediction. Assuming that segmentation quality is measured with continuous segmentation metrics, this approach can be thought of as a regression task. The quality regression network gets tuples of (image, prediction) as input, and segmentation quality is the target variable. Note that this approach differs from the RF methods from section 2.2.5.1 in the type of inputs and the regression model. In this study, segmentation quality is measured by computing the DSC with the ground truth for each foreground class, so the network outputs multiple quality estimates, which are averaged to derive a scalar confidence score. The architecture of the regression network follows the segmentation network's dataset-specific encoder but adds residual connections to each stage. Global average pooling and a linear layer are appended after the bottleneck for producing the quality predictions. As in (Robinson et al. 2018), the L2 loss is the training objective, which is optimized with AdamW (Loshchilov and Hutter 2019). Similar to the random forests used for confidence aggregation, predictions of a segmentation model have to be generated before training a quality regression network. Using segmentation model predictions on its training set would severely overestimate the quality on unseen data, so the cross-validation predictions of a single segmentation network are used instead. Preprocessing for the quality regression network comprised the following steps: (1) normalize the image with the z-score and transform the prediction mask to one-hot encoding. (2) crop the raw data to a dataset-specific bounding box around the foreground (during testing, the predicted foreground is used); (3) resize the image and prediction by a factor of 0.5 if necessary to train on a GPU with 11 GB memory. The data augmentation pipeline for the resulting patches included randomized zoom and mirroring for image and prediction, as well as Gaussian noise and intensity scaling for the images only. As the target DSC qualities of validation predictions were on average still high, additional pseudo-predictions with lower segmentation quality were introduced, by randomly (probability of 1/3) corrupting segmentation masks with affine transformations.

Mahalanobis-distance (González et al. 2022) This method follows the approach of dataset shift (or OOD) detection by modeling the training data distribution as a multivariate Gaussian distribution over high-dimensional feature space. Specifically, in the first step, feature representations of all patches in the training data are extracted by a pretrained segmentation network, using the outputs of the last convolutional layer in the bottleneck. The dimensionality of these feature maps is reduced through iterative average pooling until it is below a certain threshold (in this study 10 000). Finally, a Gaussian distribution is fitted to the flattened features. During testing, features are extracted in the same way for each sliding-window patch and a confidence score is computed as the Mahalanobis distance between the patch and the Gaussian training distribution. Patch confidences are aggregated in a sliding window manner following González et al. (2022).

Variational autoencoder (VAE) (Liu et al. 2019) Also treating failure detection as an OOD detection problem, this method is based on the idea that segmentations in medical images often have a characteristic shape. Deviations from this shape may indicate errors in the prediction. A VAE is trained on the ground truth segmentations of the training set to learn the distribution of characteristic shapes. The confidence score for unseen test samples is computed as the scalar loss value of the VAE, which is a lower bound of the likelihood the VAE assigns to this prediction.

The VAE consists of an encoder, which is a sequence of [convolution (kernel size 3, stride 2), instance norm, leaky ReLU]-blocks, and a symmetric decoder with transpose convolutions. The number of feature maps per block is [32, 64, 128, 256, 512] (or [16, 32, 64, 128, 256, 512] for the Covid and Kidney tumor datasets). In the bottleneck, the feature space is projected to 1D by a fully connected layer with output dimension 512, which is twice the latent representation dimension. On a high level, the β-VAE's loss function (Higgins et al. 2016) for an example x can be written as the weighted sum

$$L_{VAE}(x) = L_{\text{reconstruction}}(x) + \beta \cdot L_{\text{KL}}(x)\,, \tag{2.9}$$

where the second term is the Kullback-Leibler divergence between approximate posterior and prior distribution, which has a regularizing effect on the total loss. As the training data are binary segmentation masks for this failure detection method, binary CE is used as the reconstruction loss. Data preprocessing for the VAE consisted of the following steps: (1) convert the segmentation mask (ground truth during training, prediction during testing) to one-hot encoding. (2) crop the input mask to a dataset-specific bounding box around the foreground (during testing, the predicted foreground is used); (3) resize the mask by a factor of 0.5 if necessary to train on a GPU with 11 GB memory. The data augmentation pipeline for the resulting patches included randomized affine and mirroring transforms.

Table 2.6: Overview of failure detection methods included in the benchmark. Confidence aggregation methods from section 2.2.5.2 are colored dark blue and image-level methods from section 2.2.5.3 are light blue. Each method computes confidence scores based on different inputs (img: imaging data, seg: predicted segmentation, seg $\times M$: set of Monte Carlo samples of predicted segmentations, conf. map: confidence map). These inputs are processed by a CSF, which sometimes includes a model with learnable parameters. The output column provides a brief summary of what is produced by the methods and how. Each method in this overview is combined in the benchmark with a prediction model (and pixel confidence method if applicable) from section 2.2.5.1.

Method	**Input**				**Model**	**Output (brief summary)**
	img	seg	seg $\times M$	conf. map		
Mean				✗	—	Mean confidence
Foreground mean		✗		✗	—	Mean confidence (exclude predicted background)
Non-boundary		✗		✗	—	Mean confidence (exclude predicted boundary)
Patch-based				✗	—	Min. patch confidence (mean per patch)
RF (radiomics features)				✗	Regression forest	Regressed DSC based on radiomics features
RF (simple features)		✗		✗	Regression forest	Regressed DSC based on 5 heuristic features
Pairwise DSC			✗		—	Mean DSC between all pairs of segmentations
Quality regression	✗	✗			DNN	Regressed DSC
Mahalanobis	✗				Gaussian distr.	Mahalanobis distance to training distribution
VAE (seg)		✗			VAE	log likelihood of segmentation (VAE loss)

3 Results

Following the structure of the materials and methods part, section 3.1 in this chapter reports results for the benchmarking studies on the robustness of brain tumor segmentation models, the Federated Tumor Segmentation (FeTS) Challenges. The work on benchmarking segmentation failure detection methods across multiple datasets and realistic distribution shifts is presented in section 3.2, providing insights into which methods perform reliably across datasets.

Disclosure

Section 3.1 is based on the manuscript summarizing the FeTS Challenges, which has been accepted for publication (Zenk et al. 2025a), so portions of the text resemble the original manuscript text.
Section 3.2 is derived from a previously published article (Zenk et al. 2025b), so portions of the text resemble the original manuscript text, in accordance with the publisher's license.
If parts of the text replicate sections from the corresponding manuscripts, this is explicitly stated beforehand.

3.1 Generalization

The goal of the studies in this section was to assess the generalization capabilities of brain tumor segmentation algorithms when evaluated on data from institutions that did not contribute to the training data. Dataset shifts, such as differences in acquisition or population, are expected to occur in this setting and could lead to model failures. Towards that goal, two international competitions were organized in consecutive years (the FeTS challenges 2021 and 2022) and their results were analyzed to answer the research questions (RQs) introduced in section 1.4:

RQ 1.1: Do current brain tumor segmentation algorithms generalize "in the wild"?

RQ 1.2: Which algorithm and dataset characteristics affect generalization?

RQ 1.3: Which practical hurdles are associated with federated evaluation?

The two iterations of the challenge, which took place in 2021 and 2022, are presented in section 3.1.1 (FeTS21) and in section 3.1.2 (FeTS22), respectively. Both challenges share the same concept, but FeTS22 was a larger study in terms of dataset size and number of evaluated algorithms, so the results of FeTS21 are kept short here, and the focus lies on the scientific insights and practical experiences gained from FeTS22.

3.1.1 Results of the FeTS challenge 2021 (Pilot Study)

This challenge was the first to evaluate submissions in a real-world federation of institutions, that is, with geographically distributed testing data. It can, therefore, be seen as a pilot study, which was extended in the following year's FeTS challenge 2022 (section 3.1.2). Although the results of the two years are overall consistent, a short description of the main results from 2021 is given below for completeness.

3.1.1.1 Participating Teams

In total, 14 teams registered for the challenge and made submissions on the validation set. From those, 4 teams prepared a submission to the generalization benchmark, out of which one was not functional, so eventually 3 algorithms were evaluated on the test dataset. The description of the three individual contributions below is a summary of the teams' publications and adapted from Zenk et al. (2025a).

Team Alpaca (Nalawade et al. 2021) trained a model with federated learning (McMahan et al. 2017) and developed a new weight-aggregation logic based on the average validation Dice similarity coefficient (DSC) scores of each training institution. Hyperparameters used for training the network were selected based on the performance of the previous round of federated training. For the first 5 rounds, the learning rate was set to 10^{-3} and the number of epochs per round to 10. In each round after that, if the calculated average DSC was $0.5 < \text{DSC} \leq 0.8$, the learning rate was changed to 10^{-4} and the epochs per round to 5. If the DSC was larger than 0.8, the learning rate was further reduced to 10^{-5}.

Team CUHK (Yin et al. 2021) extended the nnU-Net segmentation method that won the BraTS 2020 challenge (Isensee et al. 2021b) with a test-time adaptation mechanism, to dynamically adjust the model parameters at test time, thus compensating for distribution shifts due to the varying image acquisition conditions. They adapted an approach from image classification (Wang et al. 2020a), which notes that higher entropy of model predictions usually reflects a notable domain shift. Therefore, the unsupervised objective for model parameter adaptation was the minimization of the prediction entropy of test samples.

Team MBI (Pawar et al. 2021a) proposed orthogonal encoder-decoder convolutional neural networks (CNNs) for brain tumor segmentation in two stages. An orthogonal network is an ensemble of three networks trained on axial, sagittal, and coronal 2D slices. Training and prediction were performed in two stages: in stage-I, a coarse segmentation for the whole 3D volume was predicted slice-wise using an orthogonal U-Net. In stage-II, the labels from stage-I were used to crop the whole tumor region, and seven orthogonal networks were used to predict a fine segmentation label for the region of interest. The final segmentation label was estimated using the averaged probability of all eight predictions. Heavy data augmentation, consisting of geometric transformation and random contrast, was used to avoid overfitting and improve the generalization.

3.1.1.2 Challenge Results

To answer RQ 1.1 (page 68), the segmentation metric values for all models and 21 institutions are shown in fig. 3.1. The institution with ID 22 is not included, as a technical issue made its results incorrect. The two best teams achieved good performance, both in terms of DSC and Hausdorff distance (HD), for most of the institutions. However, a few outlier institutions (07, 11, 19) had lower median performance, which indicates a lack of generalization. Even in the institutions that reported high median performance, there were individual cases with worse metric values, most prominently institution 10.

Comparing the ranking of the different algorithm submissions in fig. 3.2 (left) shows that the team CUHK won most of the subrankings. Each subranking was computed for one metric measured for a single tumor region at one institution, for example ranking all models by DSC for the whole tumor (WT) region in institution 01. The ranking stability analysis (fig. 3.2, right) confirms the clear ranking: The submission by the team Alpaca was significantly inferior to the others in most subrankings, and CUHK had a significant advantage over MBI in 24 of the 126 subrankings.

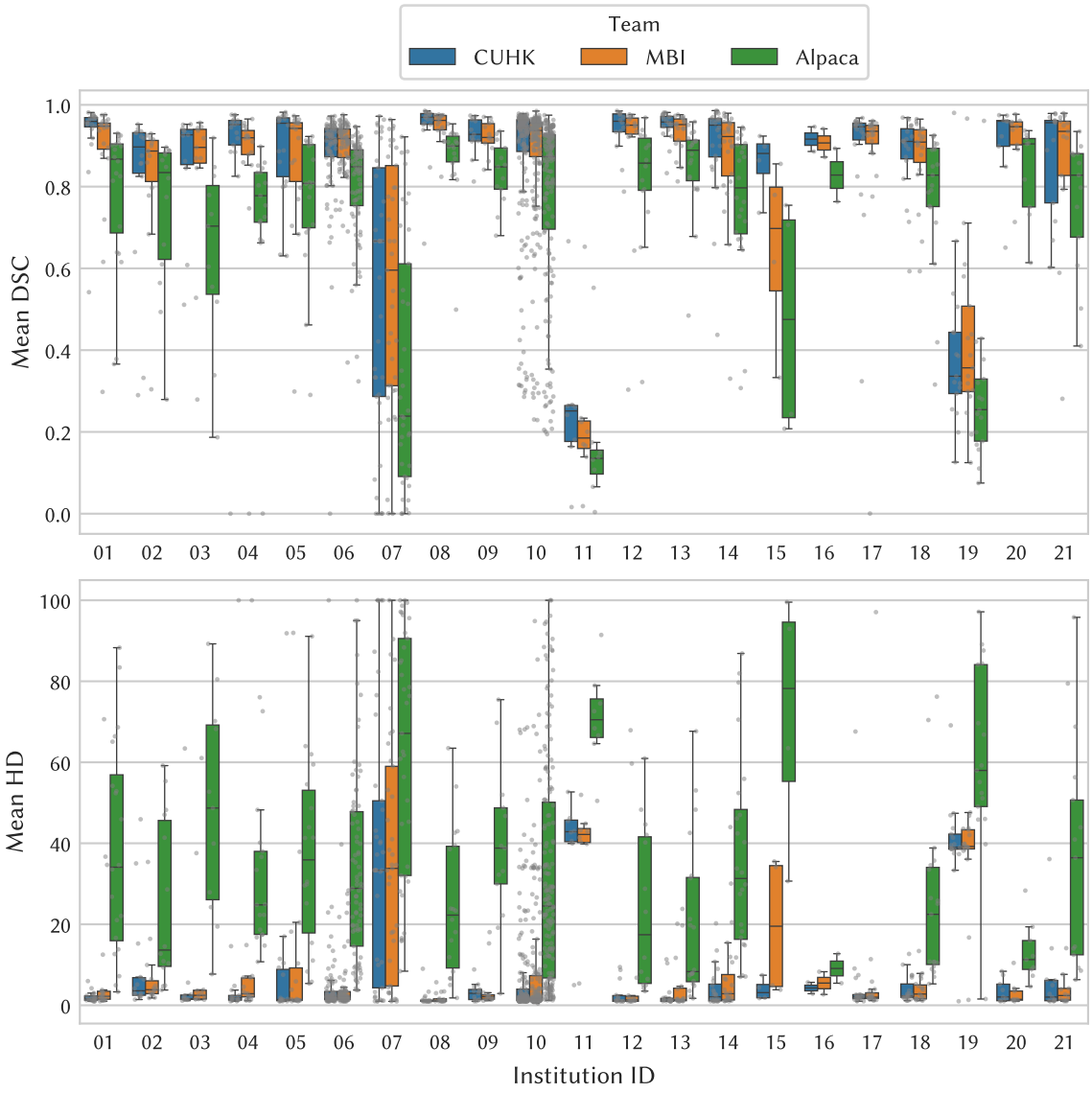

Figure 3.1: Distribution of metric values for each institution and participating team of the FeTS21 challenge. The mean over tumor regions is reported for DSC and HD, and for HD, values were clipped to 100 for clarity. These results clearly show the performance differences between the teams. While the best algorithm (team CUHK) performs well on the majority of institutions, failures occur for individual institutions (especially IDs 07, 11 and 19) and test cases.

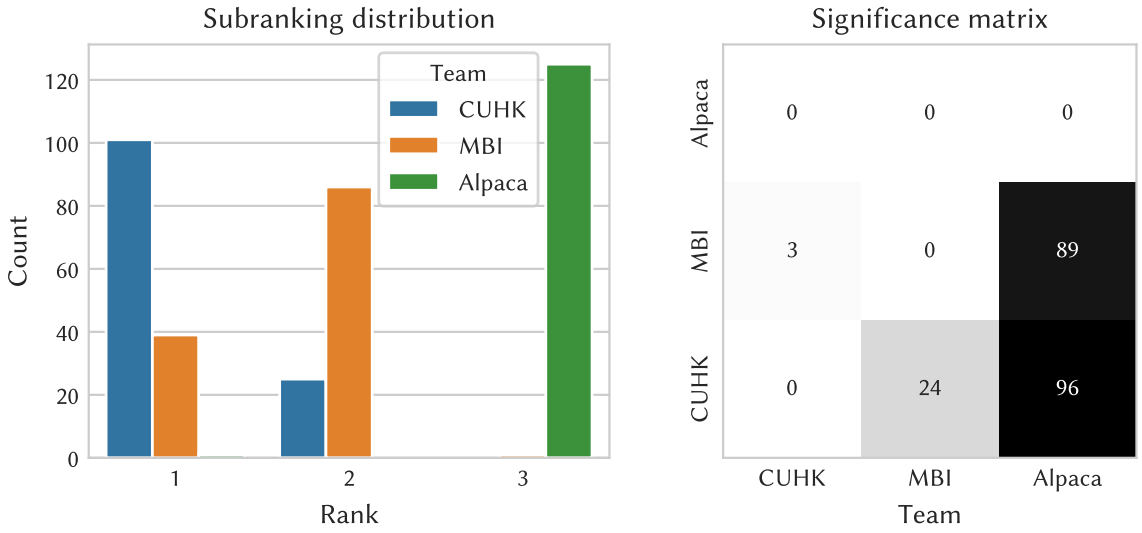

Figure 3.2: Ranking statistics for the FeTS21 challenge. Left: The distribution of the subrankings (126 in total, one for each combination of the 21 institutions, 3 tumor regions, and 2 metrics) shows that CUHK secures most first positions and Alpaca is consistently third. Right: For each subranking, a significance test is performed and the number of significant pairwise comparisons is added. This confirms that CUHK's performance is significantly superior to MBI. Figure adapted from (Zenk et al. 2025a).

3.1.1.3 Practical Experiences and Conclusions

As a pilot study for federated evaluation in a challenge setting, one goal of FeTS21 was to gather experiences and identify areas for improvement. Here the main issues encountered during the challenge organization are summarized, including comments about improvements in the following challenge iteration (FeTS22).

Regarding the technical implementation, one observation was that the submission system consisted of too many parts and did not use a standard challenge platform, making it hard to use for challenge participants. Moreover, upload speeds for singularity images were a bottleneck for some participants. This motivated using a new, established challenge platform for FeTS22 (synapse.org). During the federated evaluation, several problems were encountered, mostly related to GPU compatibility but also caused by incomplete installations or data issues. Although the evaluation system was tested before installing it in the institutions, the heterogeneity of IT systems and personnel led to unexpected issues in the federation. Therefore, the software testing was extended significantly in FeTS22 by also involving a subset of data contributors in the federation prior to the actual evaluation to find solutions to common issues.

Finally, the challenge analysis of FeTS21 and, consequentially, insights from it were

limited because no additional information apart from the segmentation metric results was available due to the federated evaluation. Meta-data about the institutions or individual cases would have been helpful, as would access to a subset of images for which the algorithms perform badly. In FeTS22, both institution-level meta-data and individual test set images could be used for the analysis.

3.1.2 Results of the FeTS challenge 2022

3.1.2.1 Participating Teams

The challenge attracted the interest of 35 registered teams, of which 6 teams submitted to the generalization benchmark. As one failed the compatibility test described in section 2.1.6, eventually 5 submissions were admitted to the federated evaluation (IDs: F01–05). To increase the algorithmic diversity, I additionally converted 36 algorithms submitted to the BraTS 2021 challenge to include them in the challenge analysis (IDs: B01–B36). This was possible because the training data was identical for FeTS22 and BraTS 2021 (as mentioned in section 2.1). Submissions B01–B36 are described in their respective publication; the appendix contains a list of references in table A.2. A summary of the five official contributions, F01–F05, is presented in the following five paragraphs, which are taken from (Zenk et al. 2025a). The original text was written by me based on the participants' papers and slightly adapted here to match the style of this thesis. Algorithm characteristics for all submissions are also analyzed in section 3.1.2.5 and summarized in table 3.1.

Team Sanctuary (F01) (Jiang et al. 2022) based their submission on the nnU-Net contribution for BraTS 2020 (Isensee et al. 2021b), extending it with test-time adaptation through batch normalization (BN) statistics. Unlike the conventional approach of collecting and freezing BN statistics during training, their method leverages test data information to dynamically correct internal activation distributions, thus addressing domain shift issues. In their approach, BN statistics are recalculated based on the batch at prediction time. Notably, the algorithm utilized a batch size of 1 during testing. Furthermore, the team employed an ensemble strategy involving six models trained on distinct training data folds. Each of these models underwent adaptation using test-time BN.

Team Graylight Imaging (F02) (Kotowski et al. 2022) built upon the 3D nnU-Net framework, incorporating a customized post-processing step specifically designed for the tumor core (TC) region and previously used in (Kotowski et al. 2021). The post-processing method named *FillTC* relabels voxels surrounded by TC to necrotic core/necrocyst (NCR). This iterative post-processing is sequentially applied to each 2D slice, first in the axial direction and subsequently in the coronal and sagittal directions. The rationale behind this approach is grounded in clinical expertise, suggesting that significant tumors typically lack voids of healthy tissue. Furthermore, if a given region is surrounded by NCR or enhancing

tumor (ET), it is deemed to be part of the TC.

Team NG research (F03) (Ren et al. 2021) adapted their submission from the BraTS 2021 challenge, which makes heavy use of model ensembling. The ensemble comprises five models of diverse architectures, which are combined with mean softmax: three are based on U-Net (Ronneberger et al. 2015), one on HRNet (Wang et al. 2021), which is a CNN that maintains multi-resolution branches, and one on CoTR (Xie et al. 2021), which combines convolutional and self-attention blocks in the encoder. Their models were refined by several strategies: Randomized data augmentations, including affine transforms, mirroring, and contrast adjustment, were employed during training to enhance model robustness. Furthermore, a post-processing step was integrated that selectively discarded ET predictions falling below a specified volume threshold.

Team vizviva (F04) (Peiris et al. 2022) employed an encoder-decoder architecture based on volumetric vision transformers. In this setup, the encoder partitions a 3D scan into patches, subsequently processing them through layers that amalgamate the outputs of 3D Swin transformer blocks (Liu et al. 2021) and 3D CSwin transformer blocks (Dong et al. 2022). For the decoder, 3D Swin transformer blocks and patch expansion layers are utilized to reconstruct the processed information. The training strategy involved a combination of cross-entropy and Dice loss. Additionally, to bolster the model's resilience against adversarial examples, virtual adversarial training introduces an extra loss term.

Team HPCASUSC (F05) (Shi et al. 2022) used a 3D U-Net model and added improvements inspired by (Isensee et al. 2021b). They used region-based training, which uses the WT, TC and ET regions as labels during training instead of NCR, ED and ET. Further, they increased the batch size to 24 and used batch normalization layers instead of instance normalization. Data augmentation consisted of random mirroring, rotation, intensity shift, and cropping.

3.1.2.2 Findings from the Annotation Quality Control

While dataset diversity was the central goal of the FeTS challenge, it also bore risks with respect to the annotation process. Each institution contributing data to the test set also generated the reference segmentation independently, which complicates annotation quality control. Fortunately, a subset of institutions shared their datasets with the FeTS challenge organizers after the federated evaluation, allowing to perform quality control on the annotations for 1201 cases from 16 institutions. The main goal of the quality control was to quantify the frequency of annotation errors and estimate whether they affect the challenge results. In total, 125 cases (10.4 %) were identified by visual examination to have insufficient quality, with a median of 5 erroneous samples per institution. These were subsequently not used for the challenge ranking and analysis. The observed errors

were diverse and differed between the institutions: Individual reference segmentations appeared to be inaccurately hand-drawn, some resembled an extremely noisy automatic segmentation, and others were completely empty. Apart from segmentation mistakes, another reason for exclusion was the quality of imaging data: In some cases, errors during registration or skull-stripping prevented an accurate segmentation. Furthermore, for one institution, 11 duplicated scans were found.

The most consistent annotation errors between institutions, however, were related to bright blood products. When bleeding occurs near the tumorous brain region, this often appears as hyperintensities in both native T1-weighted (T1) and contrast-enhanced T1-weighted (T1-Gd) sequences. According to the official annotation protocol, however, the ET region should only contain the enhancements that are not hyperintense in T1. In 43 cases, this convention was not obeyed and T1-Gd hyperintense regions were labeled as ET irrespective of their appearance in T1. Figure 3.5 (c) shows an example case with this error.

A second common issue is the subjective delineation of the TC region. This occurs because the official TC definition also includes non-enhancing tumor components, which can be difficult to differentiate from edematous or infiltrated areas. Since the inter-annotator discrepancies caused by this are in line with the annotation protocol, cases with large TC parts, possibly interpreted as non-enhancing tumor regions, were not considered errors. However, it is worth mentioning that 46 cases could potentially fall into this category. One example is shown in fig. 3.5 (d).

For the two most salient issues described above, similar cases could also be found in the training set, highlighting a prevalent issue in the field of medical image segmentation: the reference segmentations used for algorithm evaluation may not always represent the ground truth, which is usually inaccessible for radiological images. Furthermore, annotators are known to disagree with each other (or even themselves at different time points), resulting in inter- and intra-rater variability in creating these reference segmentations. For the BraTS challenge, the median inter-rater agreement has been estimated previously to DSC scores of 0.87, 0.86 and 0.77 for the WT, TC and ET regions (Menze et al. 2015).

In conclusion, the annotation quality control revealed two issues that occurred consistently across institutions. However, about 90 % of the inspected test cases had acceptable segmentations, and figs. A.3 and A.4 in the appendix show there are only minor differences in the rankings and DSC distributions, respectively, before and after quality control. Although it is possible that annotation quality differs in the part of the test set that could not be shared, these results provide an estimate for the frequency of potential annotation errors and suggest that they do not considerably affect the challenge results.

3.1.2.3 Challenge Analysis

To find out whether current brain tumor segmentation algorithms generalize "in the wild" (RQ 1.1 on page 68), 41 models were evaluated on test cases from 32 institutions using federated evaluation. Figure 3.3 shows the mean DSC values per institution and model, revealing that the top models achieved good results with mean DSC scores around 0.9 for most institutions. However, there were also some sites on which the DSCs were considerably worse, for example institution IDs 12, 15, 16 and 24. Note that the top models' scores were largely consistent within each institution (column in the heatmap of fig. 3.3), which indicates that they shared a lack of generalization to the institutions with lower performance. These trends were less pronounced but similar for the HD metric; the corresponding diagram can be found in the appendix (fig. A.1). Another observation from these results is that there were six institutions for which not all algorithms could be evaluated due to technical reasons (details in section 3.1.2.6). This illustrates a different aspect of robustness: Even though the algorithm might generalize to the dataset at these sites, technical compatibility can still prevent successful deployment in federated systems.

While fig. 3.3 gives a high-level picture of the challenge results for all models, it does not contain information about the performance for individual test cases. Figure 3.4 provides additional insights: The upper part reveals a lack of robustness of the top 20 models by highlighting considerable gaps between average-case generalization, measured by median DSC across cases, and worst-case generalization, measured by the 10th percentile DSC. Moreover, and surprisingly, the models did not necessarily generalize better on institutions represented in the training set than on those not contributing to training set cases. The lower part of fig. 3.4 focuses on the best model, showing that there were outlier cases with low DSC for most institutions, which can be considered failures. This is also observed when looking at DSC of individual tumor regions instead of the mean (fig. A.2 in the appendix). Together, these findings suggest that case-specific failure sources exist, which might be distinct from the institution-level distribution shift.

3.1.2.4 Qualitative Failure Case Analysis

One conclusion from the previous section was that segmentation algorithms might fail on individual samples, although they perform decently on the majority of cases from a given institution. This raises the question of what common failure sources are in the dataset. To gain a clearer picture and find a qualitative answer to the aspects of dataset characteristics in RQ 1.2 (page 68), a visual review was performed for a subset of test samples for which the ensemble of the 15 best methods obtained poor segmentation metric values. This analysis was only possible on samples the organizers had access to, which correspond to those used for annotation quality control (see section 3.1.2.2). The following points

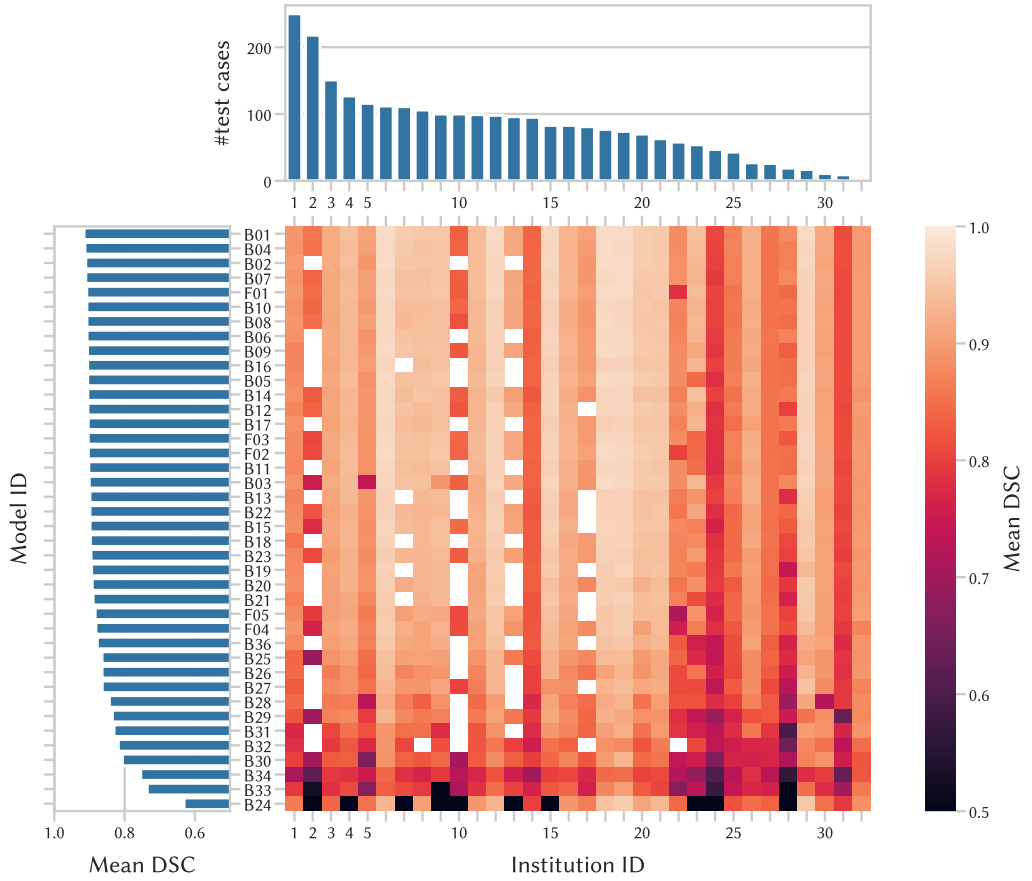

Figure 3.3: Aggregated challenge results of the FeTS22 challenge for each evaluated model and institution. Each tile in the heatmaps represents the DSC value of a single model, averaged over all test cases and tumor regions of one institution. The values were clipped at 0.5 and white tiles indicate evaluation runs that failed due to technical issues. Models are sorted by mean DSC (bar plot on the left) and institutions by their test set size (bar plot at the top). The best models achieved similar performances within each institution, as apparent from the vertical structures in the heatmap. However, for some institutions the performance of all models dropped, indicating a lack of robustness. Figure adapted from (Zenk et al. 2025a).

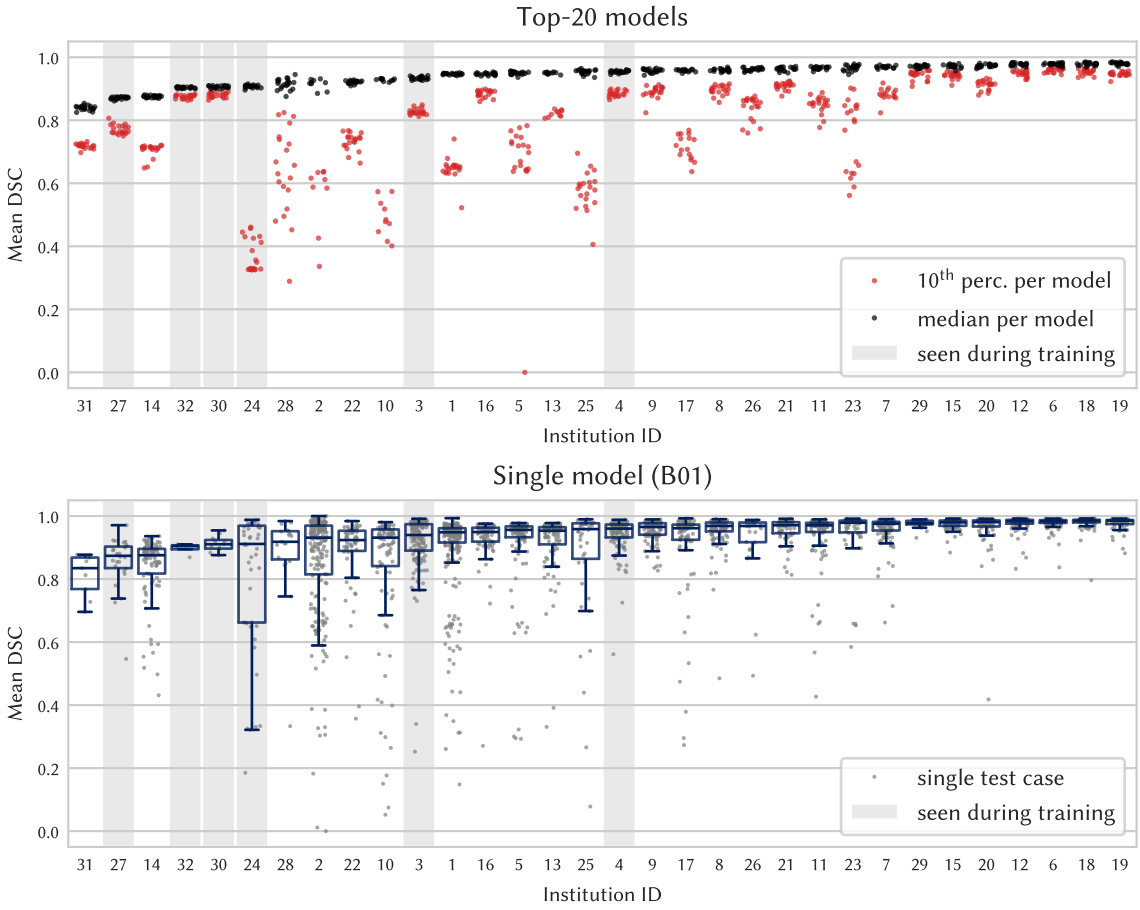

Figure 3.4: Average-case and worst-case performance of the top 20 algorithms in FeTS22 (top), and case-level challenge results of the top-ranked algorithm (bottom) for each institution of the test set. While most institutions were not seen during training, some also contributed (different) cases to the training set ("seen during training"). Top: For each institution and the 20 best-ranked models, the median and 10th percentile of DSC (mean across the three tumor regions) are shown as a measure for average-case and worst-case performance, respectively. Low values for the median or a large gap to the 10th percentile indicate robustness issues, which occurred in several institutions. Bottom: Detailed results for one model, where each gray dot represents the DSC score for a single test case. Despite the high median performance for most institutions, there were often individual cases with reduced performance, even for institutions seen during training. Figure adapted from (Zenk et al. 2025a).

describe the most frequent failure cases for each tumor region:

- WT: T2-hyperintensities were sometimes labeled as edema (ED), although they are caused by other pathologies. Figure 3.5 (a) shows an example where the prediction segments one half of the symmetrical hyperintensity near the ventricles, although it is far from the enhancing tumor.
- TC: The evaluated models occasionally labeled "random" parts near the ET region as NCR, despite little evidence in the imaging data for necrotic or cystic tissue (fig. 3.5 (b, d)). This is likely due to the inherent ambiguity in non-enhancing tumor regions, which were a separate label in BraTS challenges before 2017 but were later merged into the NCR label (Bakas et al. 2019). Similar issues were observed during the annotation quality control (section 3.1.2.2).
- ET: Small contrast enhancements were sometimes not detected, especially if they were distant from the main tumor (fig. 3.5 (b)). A second common mistake related to ET happened for cases in which part of the tumor is hyperintense both in the T1 and T1-Gd sequences (fig. 3.5 (c)). The ET label should then only contain the areas which are enhanced compared to T1, but often the prediction covered the whole T1-Gd hyperintense region. Some reference segmentations of the training set also contain the latter flaw, which is discussed in more detail in section 3.1.2.2.

Another known failure source in brain tumor segmentation is small tumor region size, which mostly affects the TC and ET regions. Especially DSC is overly sensitive to minor differences in segmentation masks when the target objects are small (Reinke et al. 2021). In cases with low-grade gliomas, TC or ET can even be absent completely. The BraTS convention for these cases is that a prediction obtains perfect scores if it is also empty (DSC = 1, HD = 0) and worst scores if it is non-empty (DSC = 0, HD = 373.15). In the FeTS22 test set, the number of cases with empty reference regions was 1 (WT), 49 (TC) and 117 (ET), which corresponds to about 4.5 % of all cases. Figure 3.6 illustrates that also cases with small, non-empty tumor regions were difficult to segment for state-of-the-art models. The two-dimensional histogram compares region volume in the reference segmentation with the achieved segmentation metrics for the top 10 models evaluated in the FeTS challenge. While noisy DSC values are expected for small tumor regions, the HD values also show an increase in outliers for small regions, which indicates that localizing these lesions is still challenging.

3.1.2.5 Challenge Ranking

The official challenge ranking, presented at the medical image computing and computer assisted intervention (MICCAI) conference 2022, was computed based on the results of

the five original FeTS submissions on all institutions and is reflected in their model IDs, F01–F05. In a second, extended ranking, the 36 models converted from the BraTS 2021 challenge (B01–B36) were added. Institutions with IDs 2, 10, 13, and 17 were excluded from this ranking, as many BraTS models could not be evaluated on them due to technical issues. In the extended ranking, the best original FeTS submissions were pushed back to the ranks 7 to 9, and the BraTS models took the top positions instead; they hence represent the state of the art in terms of generalizability. Ranking stability with respect to randomness in test data selection was analyzed by recomputing the ranking for 1000 bootstrap samples in fig. 3.7. It shows that the ranking was more stable for the DSC metric than for HD. Furthermore, the ranking order depended on the tumor region and metric, which suggests that the teams optimized different aspects of their algorithm.

As a qualitative summary of the teams' algorithmic design choices, table 3.1 contains descriptions of salient algorithm characteristics for all submissions, and the extended ranking. The differences between the submissions mainly affected the network architecture, post-processing, and the number of models in the ensemble. With respect to the question of which algorithm characteristics affect generalization (RQ 1.2 on page 68), no single characteristic stood out among the top teams, but a few trends could be observed: U-Net variants were still the most popular network architectures and achieved very good results without requiring extensions like attention blocks or residual connections. This point is supported by the fact that implementing the algorithm in the nnU-Net framework (Isensee et al. 2021a) increased the chances for a top position: 9 of the 20 top submissions were (partially) based on nnU-Net, while within the lower-ranked half of submissions none used its self-configuring capabilities. Other popular design choices were using model ensembling and removing implausibly small predicted regions in a post-processing step. The only algorithm that adapted to distribution shifts was model F01, which recomputed the batch normalization statistics at test time. However, it was superseded in the ranking by the BraTS 2021 models, which did not have access to information about the training set partitioning. Therefore, the FeTS challenge could not clarify if there is a better way to use multicentric training data for enhancing algorithmic robustness than simply training on the pooled data.

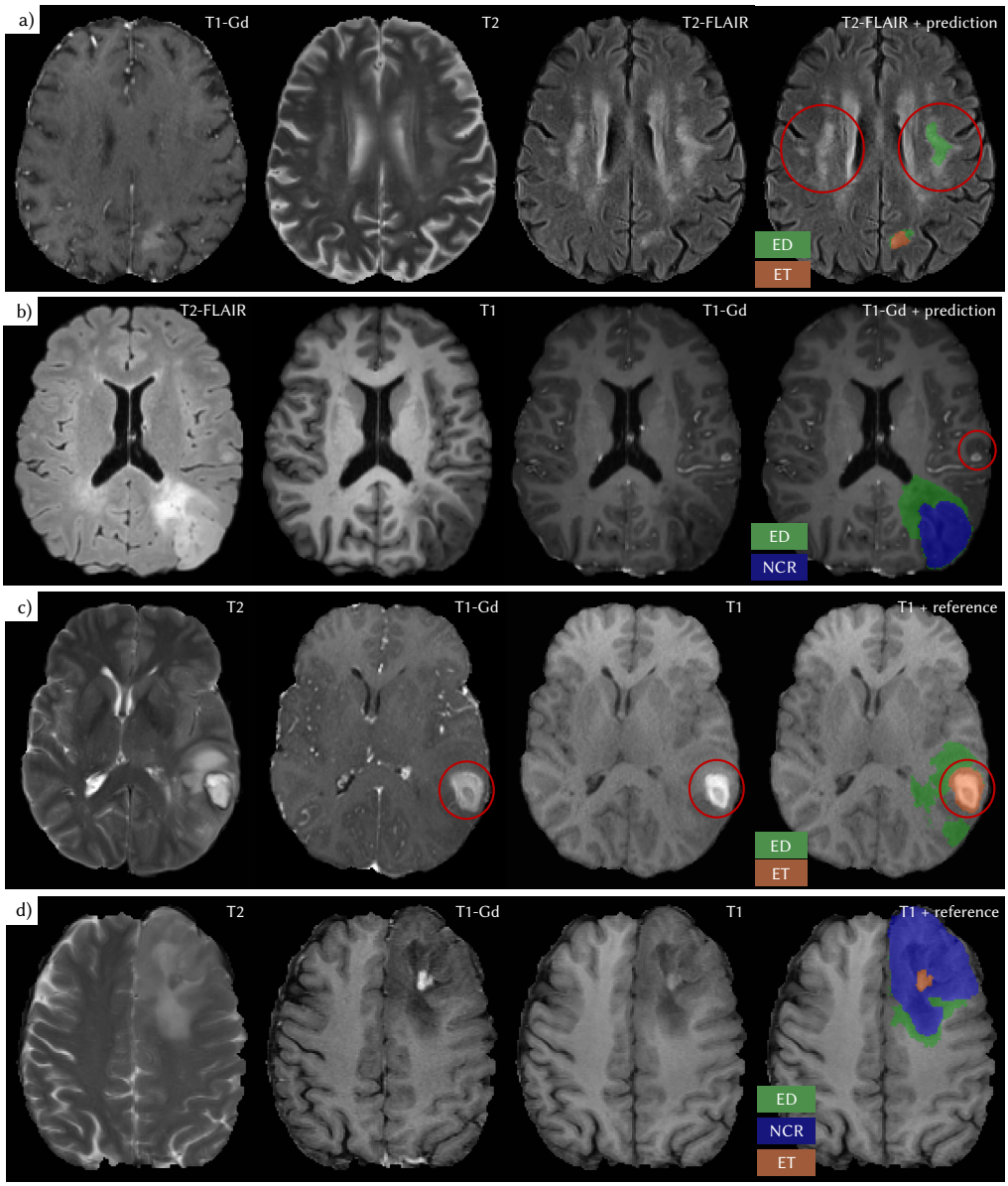

Figure 3.5: Examples of common segmentation errors. For each case (row), the three most relevant MRI sequences are shown. The depicted segmentation masks are the predictions of the best model (ID: B01) in (a, b) and the reference segmentation in (c, d), respectively. (a) False positive ED prediction. The symmetric hyperintensity is caused by a pathology that is not related to the tumor. (b) A small contrast enhancement is missed, separate from the larger tumor in the lower right. (c) Blood products are bright in T1 and T1-Gd, so they can be confused with ET. This issue occurs in some reference segmentations and test set predictions. (d) The segmentation of non-enhancing TC parts (labeled as NCR here) is difficult and often differs between annotators. Figure adapted from (Zenk et al. 2025a).

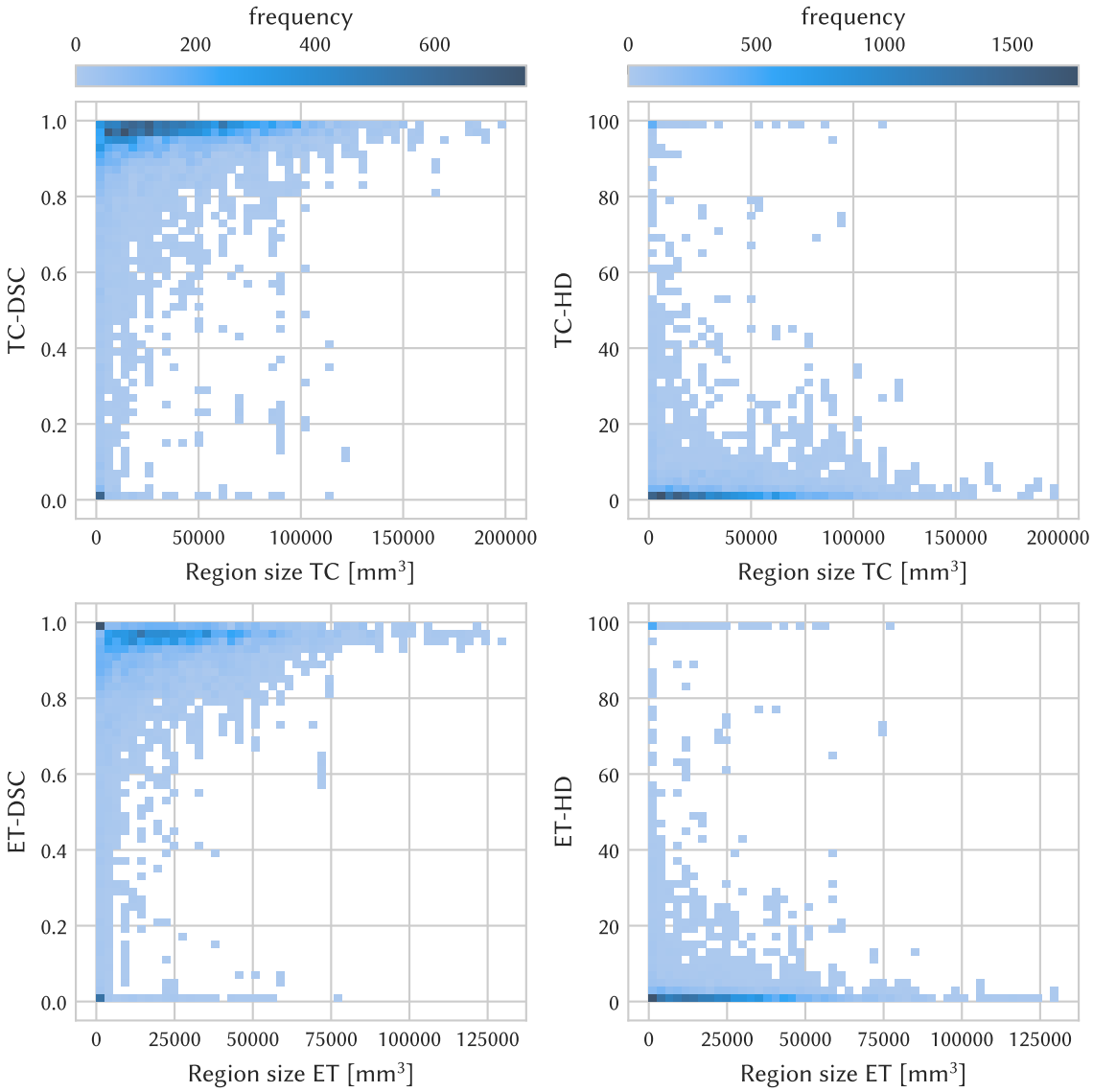

Figure 3.6: 2D-Histograms of true region size versus segmentation metrics for all test cases, accumulated over the predictions of the ten best models. Each row shows a different tumor region (TC and ET), whereas the columns correspond to two metrics (DSC and HD). HD was clipped to 100, and the color bars apply per column. The WT region is not shown as it is usually quite large, and this figure focuses on the effect of small tumor volumes. In all four panels, these small regions result in a wider distribution of DSC and HD values. In the extreme case of regions with zero size, the metrics become binary, which explains the accumulations at these points in the diagram. While DSC is known to be noisy for small regions, HD shows that it is harder for the models to localize them correctly. Figure adapted from (Zenk et al. 2025a).

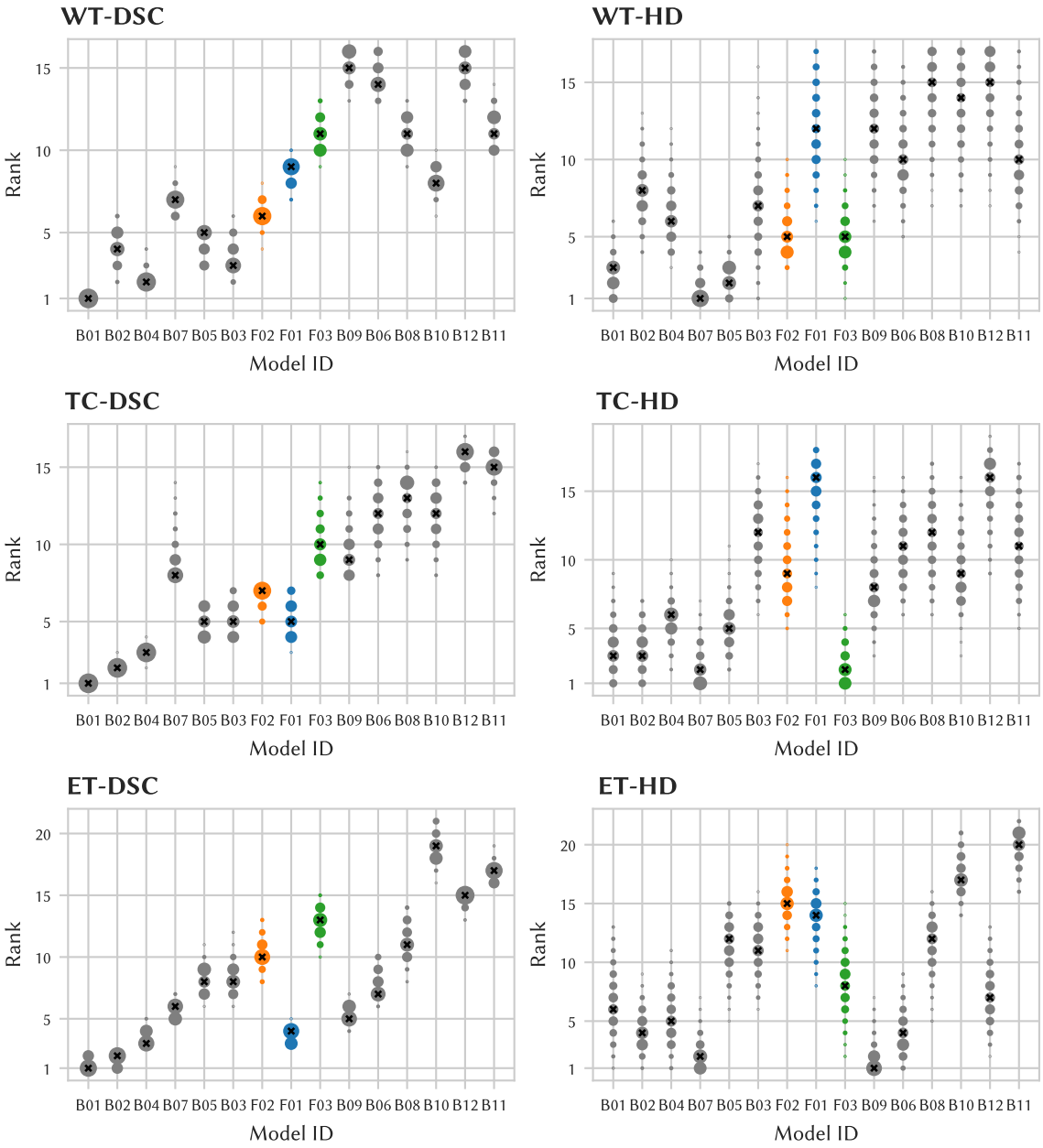

Figure 3.7: Ranking stability for each region and metric evaluated in the FeTS22 challenge, based on 1000 bootstrap samples. The size of the circular markers is proportional to the fraction of bootstrap samples for which an algorithm occupied a specific rank. Black crosses denote the median rank across bootstrap samples. For clarity, only the top 15 models are shown in the final ranking order, and the official FeTS submissions are colored. The ranking based on DSC is usually more stable than for HD, and the differences between the tumor regions indicate that the algorithms have distinct strengths.

Table 3.1: Extended ranking and algorithm characteristics of all models evaluated in the FeTS Challenge 2022. Algorithm characteristics were extracted from the participants' method description papers. The keywords in the table are described in more detail in appendix A.3. '-' denotes that nothing was reported for this field. Abbreviations: CNN = convolutional neural network, CE = cross-entropy, VAT = virtual adversarial training. Table adapted from (Zenk et al. 2025a).

Model ID	Rank	Architecture	Loss	Post-processing	Ensemble size	nnU-Net
B01	1	U-Net, larger encoder	CE, batch Dice, region-based	ET (small to NCR)	10	yes
B02	2	U-Net, larger encoder, multi-scale skip block	focal loss, jaccard, region-based	-	30	no
B04	3	U-Net	CE, Dice, TopK, region-based	-	5	yes
B07	4	U-Net, residual blocks, transformer in bottleneck	CE, Dice	ET (small to NCR)	3	yes
B05	5	U-Net	CE, Dice	ET (drop disconnected), TC (fill surrounded), WT (drop small components)	5	yes
B03	6	U-Net, larger encoder	CE, batch Dice, region-based	ET (small to NCR)	5	no
F02	7	U-Net	CE, Dice	TC (fill surrounded)	5	yes
F01	8	U-Net	CE, Dice, region-based	ET (small to NCR)	5	yes
F03	9	CoTr, HR-Net (CNN), U-Net, U-Net++	multi, region-based	ET (small to NCR)	5	yes + other
B09	10	U-Net, larger encoder, residual blocks	Dice, focal loss	ET (small to NCR)	5	no
B06	11	HNF-Net (CNN), attention	CE, genDice, region-based	ET (small to NCR)	5	no
B08	12	U-Net, multiple encoders	CE, Dice, region-based	ET (small to NCR)	4	no

—table continued on next page—

Model ID	Rank	Architecture	Loss	Post-processing	Ensemble size	nnU-Net
B10	13	U-Net	CE, Dice, generalized Wasserstein Dice	-	8	no
B12	14	U-Net, larger encoder, residual blocks	Dice, region-based	ET (small to NCR)	4	no
B11	15	U-Net, modality co-attention, multi-scale skip block, transformer in bottleneck	CE, region-based	ET (drop small components)	-	no
B17	16	U-Net	CE, Dice, region-based	ET (convert to NCR based on auxiliary network), drop small components	10	yes + other
B14	17	U-Net	CE, Dice, batch Dice, region-based	ET (small to NCR)	15	yes + other
B16	18	U-Net	batch Dice, region-based	ET (small to neighboring label), drop small components	5	no
B15	19	-	-	-	-	-
B13	20	HarDNet (CNN)	CE, Dice, focal loss, region-based	-	3	no
B19	21	U-Net, attention	Dice, region-based	-	1	no
B18	22	U-Net, attention	CE, Dice, region-based	-	1	no
B21	23	-	-	-	-	-
B22	24	U-Net, multiple decoders	CE, Dice, region-based	TC (remove outside of WT), drop small components, morph. closing	1	no
B20	25	2-stage, 2D, CNN, U-Net, U-Net++, residual blocks	Dice	-	29	no
B23	26	CNN, neural architecture search	CE, Dice, region-based	-	5	no

—table continued on next page—

Model ID	Rank	Architecture	Loss	Post-processing	Ensemble size	nnU-Net
F05	27	Swin Transformer	CE, Dice, VAT, region-based	-	1	no
F04	28	U-Net	Dice, region-based	-	1	no
B36	29	U-Net	CE, Dice	-	1	no
B26	30	2D, U-Net, attention, residual blocks	CE, Dice	-	-	no
B27	31	U-Net, attention, residual blocks	Dice, region-based	-	5	no
B25	32	2D, U-Net, residual encoder	Dice	-	1	no
B24	33	2-stage, U-Net, residual blocks	CE, Dice, region-based	ET (small to NCR)	5	no
B28	34	2D, U-Net, residual encoder, transformer	CE, Dice, region-based	-	1	no
B30	35	2-stage, U-Net	-	-	1	no
B29	36	U-Net, multi-stage	BCE	fill holes	1	no
B31	37	2D, U-Net++	Dice, boundary distance	-	3	no
B32	38	2-stage, CNN, Graph NN	CE	-	1	no
B35	39	CNN, larger encoder, residual blocks	Dice, boundary, region-based	ET (small to NCR)	1	no
B33	40	2D, U-Net	Dice	-	1	no
B34	41	-	-	-	-	-

3.1.2.6 Practical Experiences with Federated Evaluation

The FeTS challenge was the first to employ federated evaluation in a challenge benchmarking setting. Consequently, various practical problems concerning its organization or implementation had to be solved. Here, the most important experiences are described and put in a general context ("lessons learned"), to answer RQ 1.3 (page 68) and aid other researchers in the future with similar projects.

The federated evaluation demanded significant amounts of time and coordination, especially for setting up the software and resolving technical issues at federated site. While the challenge participants developed their algorithms in the training phase, the evaluation system was already tested on a small subset of institutions. This allowed for finding solutions to common problems encountered during these preliminary tests, which could later speed up the troubleshooting during the large-scale setup. In the official test set evaluation phase, a compatibility test was conducted once at each institution, which consisted of sanity checks on the local data and running the complete evaluation pipeline using a reference model. The reference model was based on a nnU-Net submission to a previous BraTS challenge (Isensee et al. 2021b). To identify potential issues with the software and the data, the compatibility test was run for a standardized toy dataset and for the local test data, respectively. Eventually, setting up the evaluation system at each site took from a few days to several weeks. Running the actual evaluation was fast compared to the time spent on evaluation system setup: On a single-GPU reference hardware, executing the evaluation pipeline for all 41 models on 100 cases required 86 hours. In the real-world federation, however, the inherent heterogeneity of IT systems led to diverse technical issues during setup and evaluation, which required remote support from the organizers to resolve. Coordinating the troubleshooting with each institution's representative through shared log files, emails or video calls resulted in slow feedback loops, making communication the primary bottleneck in the FeTS challenge. This significantly extended the duration of the complete evaluation process (including setup), which ranged from a few weeks to half a year. In conclusion, the practical hurdles and organizational effort of conducting a large-scale federated evaluation should not be underestimated, and extensive technical monitoring and support are likely necessary.

One particularly frequent issue within the FeTS challenge was compatibility with the heterogeneous GPU hardware in the federation. As a countermeasure, a specific base docker image compatible with all common GPU models available at the time of the challenge was recommended for official challenge submissions. However, this recommendation was not followed by the 36 models adapted from docker images submitted to the BraTS 2021 challenge, as their base images were chosen before the FeTS challenge took place. Consequently, the official challenge submissions did not encounter errors related

to GPU compatibility at any institution, whereas some BraTS submissions could not be run everywhere, resulting in the missing model evaluations from fig. 3.3. This experience highlights that the (GPU) hardware present in a federation needs to be assessed early on, to determine effective compatibility solutions that can be used by challenge participants and guarantee a successful federated evaluation.

Finally, for a federated challenge such as FeTS, case-specific meta-data for the challenge test set can facilitate advanced analyses beyond pure ranking and segmentation metric values. In the FeTS challenge, meta-data was only available at the institution level, which could be used for a descriptive data analysis. To identify failure sources, however, case-specific information about the tumor region size and (limited) data sharing were more helpful. As data sharing is usually not easily done in federated setups, the experiences from the FeTS challenge suggest that additional meta-data on the patient level should be collected for the test set to facilitate deeper insights.

3.2 Failure Detection

The goal of this study was to compare methods that can notify the user when a segmentation model's prediction is suspected to contain errors. To this end, the following RQs were posed in the introduction:

RQ 2.1: What are best practices and pitfalls related to the evaluation of segmentation failure detection?

RQ 2.2: Which failure detection algorithms are reliable across multiple datasets?

RQ 2.3: How to aggregate pixel-level confidence into image-level scores for failure detection?

A unified evaluation protocol was presented in section 2.2.2, which addresses RQ 2.1 and allows benchmarking a variety of methods from section 2.2.5. The results of this large-scale benchmark are reported in Section 3.2.1, providing answers to research questions (RQs) 2.2 and 2.3. Section 3.2.2 shows extensions to the original benchmark that investigate how failure detection methods are impacted by the network architecture or dataset size and whether they generalize to other imaging modalities. Finally, section 3.2.3 presents an analysis of the evaluation protocol and clarifies how the results change if alternative protocols are used, thereby providing quantitative results for RQ 2.1.

3.2.1 Failure Detection Benchmark Results

This part initially reports results for the segmentation performance obtained on the six different datasets that were used for benchmarking failure detection methods (section 3.2.1.1) to illustrate that the segmentation models produce enough failures that can be detected. Based on this foundation, results for the two classes of failure detection methods are presented, first for methods based on the aggregation of pixel confidence maps (section 3.2.1.2) and then for methods that perform image-level confidence scoring directly (section 3.2.1.3). An overall comparison of the two classes is integrated in the latter section.

3.2.1.1 Segmentation results

Following the methodology from section 2.2.4, five U-Nets with different random seeds were trained for each of the five cross-validation folds, resulting in 25 models per dataset

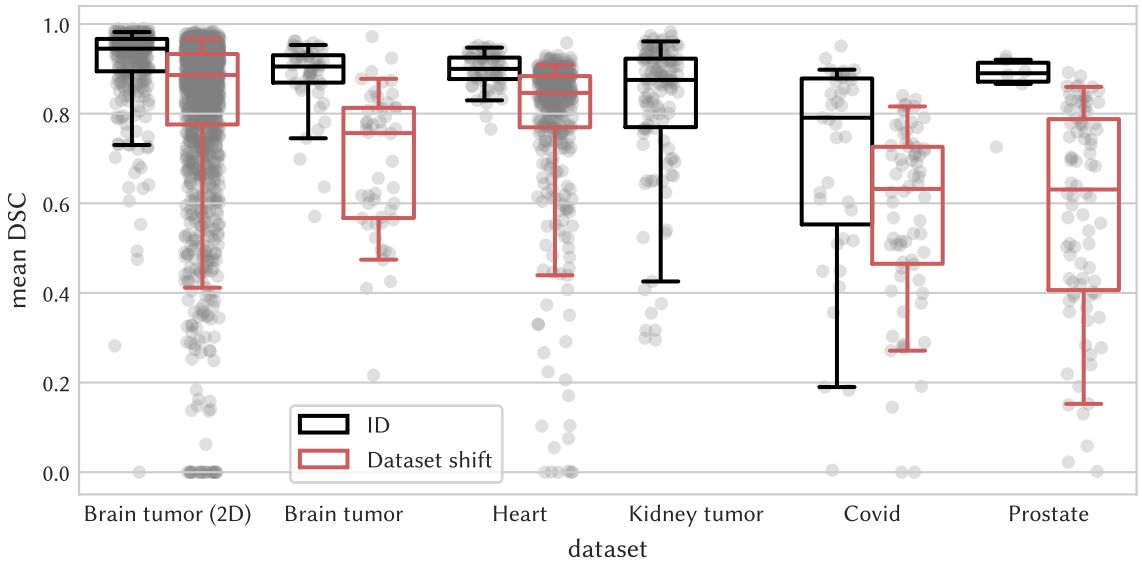

Figure 3.8: Test set segmentation performance measured by DSC, of a single U-Net trained on the first fold of the training data. Boxes are laid over scores for individual test cases (gray points), showing the median and interquartile range (IQR) of the DSC distribution, while whiskers extend to the 5th and 95th percentiles, respectively. Each dataset includes samples originating from the same distribution as the training set (in-distribution, ID) and samples exhibiting a dataset shift. Typically, performance was higher on in-distribution samples compared to those affected by a distribution shift; however, notable exceptions were observed, such as in the Kidney tumor dataset (which lacked distribution shifts) and the Covid dataset, where several in-distribution failure cases were also present. Figure adapted from (Zenk et al. 2025b).

from section 2.2.3. Most datasets contain samples with distribution shifts in their test set to simulate a realistic deployment scenario with unseen data characteristics at test time. As fig. 3.8 shows, failures occurred more frequently for samples with dataset shifts, but the kidney tumor and Covid datasets show that low DSC score segmentations were also found for in-distribution samples. These results reflect the performance of a single U-Net, which is usually inferior to ensembling multiple predictions. Figure 3.9 visualizes the performance gains from using an ensemble of networks or MC-Dropout, which were trained on the same data as the single network baseline. While the improvements from MC-Dropout were negligible, the ensemble could increase the mean DSC by +0.008 for the brain tumor 2D dataset (single net: 0.833), +0.011 for the brain tumor dataset (single net: 0.785), -0.002 for the heart dataset (single net: 0.791), +0.016 for the kidney tumor dataset (single net: 0.834), +0.009 for the Covid dataset (single net: 0.628) and +0.045 for the prostate dataset (single net: 0.566). Individual samples exhibited much larger

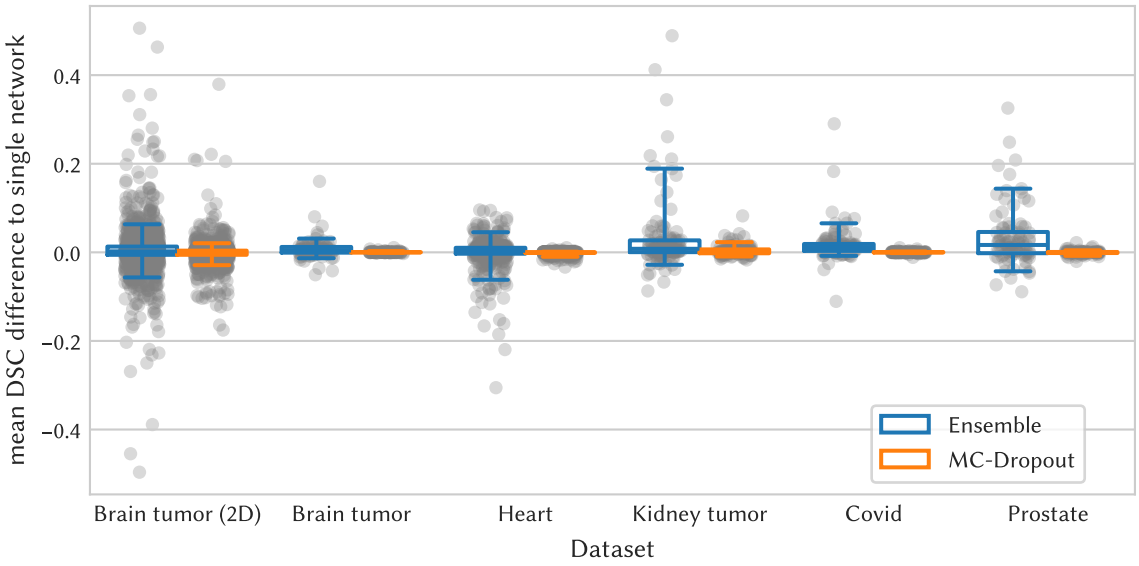

Figure 3.9: Performance difference (DSC scores) on the test set between a single network, a model that averages the predictions of multiple forward passes through the same network with enabled test-time dropout (MC-Dropout), and an ensemble containing four additional models trained independently on the same data. The median DSC differences are larger for Ensemble than for MC-Dropout, but overall still small compared to the deltas on individual test cases. Most distributions for the Ensemble are skewed toward positive values, indicating that it can already avoid some failures by improving the segmentation performance.

differences (in both directions), turning a failure case into a successful prediction and vice versa. Hence, such differences should not be neglected when evaluating failure detection methods, which is captured by evaluation protocol requirement R2 from section 2.2.2.1. For this reason, whenever possible, comparisons between failure detection methods in the next sections are made based on the same underlying prediction model, for which the options are single network, MC-Dropout, or Ensemble. Although the suggested risk-coverage analysis and the area under the risk-coverage curve (AURC) metric allow a fair comparison even between different prediction models, any such comparison is transparently reported in the following because a difference in AURC can be due to enhanced failure detection performance *or* better segmentation performance by the prediction model.

3.2.1.2 Pixel confidence aggregation Methods

The segmentation performances from the previous section determine the risks associated with each prediction. Failure detection methods try to output a confidence score that

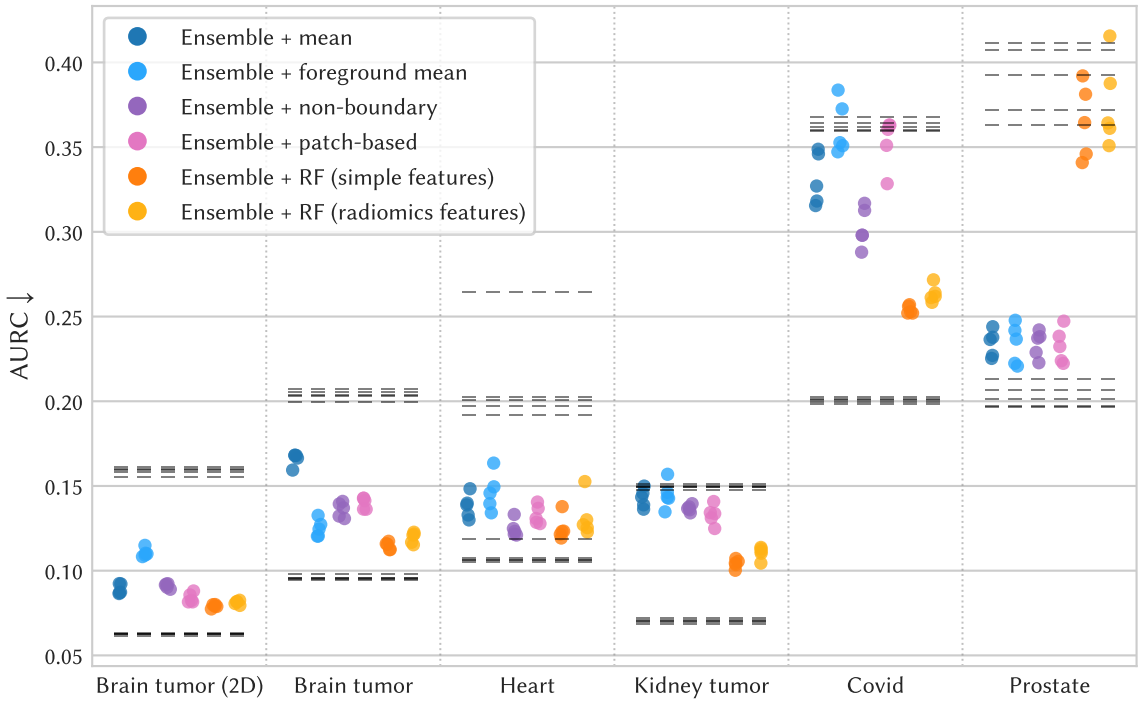

Figure 3.10: Comparison of aggregation methods from section 2.2.5.1 in terms of AURC scores for all datasets (lower is better). Each colored dot corresponds to a single experiment, meaning a prediction model + aggregation method, trained on one fold of the training data. PE was used as a pixel-wise uncertainty. Gray "–" markers visualize values for the random/optimal confidence baselines, which differ between the models trained on different folds due to their varying segmentation performance (section 2.2.2.2). Aggregation methods based on regression forests (RF) show performance advantages on most datasets but fail catastrophically on the prostate dataset, possibly due to the small training set size. Figure adapted from (Zenk et al. 2025b).

estimates that risk. This section focuses on the family of methods that aggregate pixel confidence maps (section 2.2.5.1). Evaluating the AURC metric for all aggregation methods based on the confidence maps produced by a deep ensemble results in the experimental data shown in fig. 3.10. The best-performing aggregation method among those that do not require training (mean, foreground mean, non-boundary, patch-based) differed between the datasets. On four of six datasets, all of these methods improved upon random confidence scores, but for the kidney tumor and Covid datasets, some of them outperformed this baseline only marginally. The two methods based on DSC regression through a regression forest (RF), which is trained on features of the prediction and confidence map, achieved lower (better) AURCs on five of six datasets. However, the Prostate dataset

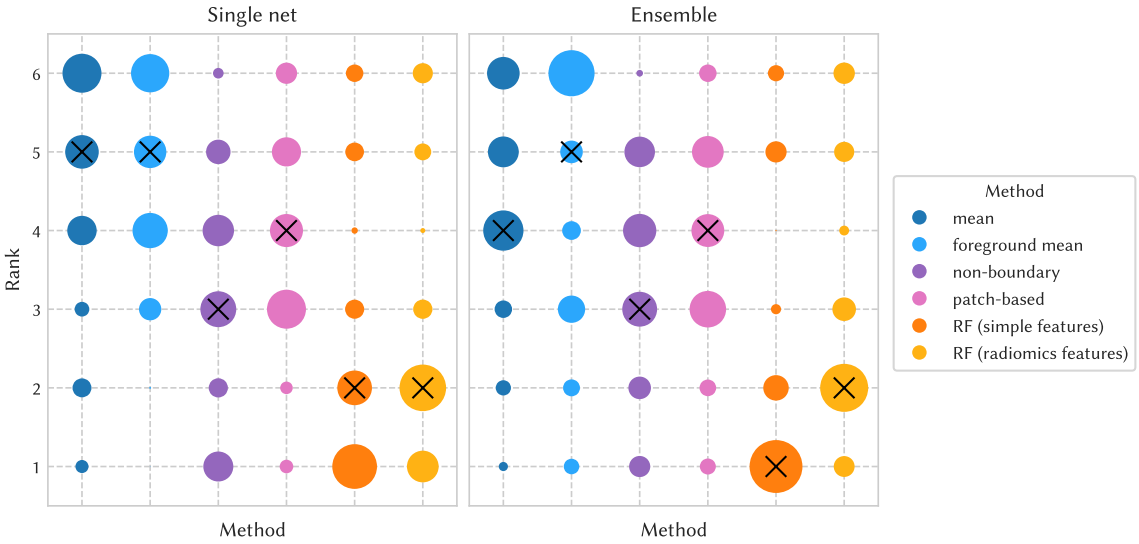

Figure 3.11: Ranking distribution obtained through bootstrapping ($N = 1000$) for aggregation methods, compared between two different prediction models (single net and ensemble). The bootstrapped ranks for all six datasets are combined in this figure to provide an overview across datasets. Here, the area of each circular marker is proportional to the ranking count across bootstrap samples, and the black crosses indicate the median rank. There are only small differences between the ranking distributions when using a single network or ensemble. Mean and foreground-mean are usually worst, while failure detection methods based on a regression forest (RF) work best.

exemplifies an important exception, as their performance dropped to the random level on it. A possible explanation for this behavior is the small training set size for the prostate dataset ($N = 26$), which may have been insufficient to train a generalizable regression forest on the simple/radiomics features.

Figure 3.10 focuses on the performance of different confidence aggregation methods based on the predictions of a deep ensemble and confidence maps derived from them via predictive entropy. Both of these components—in the language of section 2.2.5.1, the prediction model and pixel-wise uncertainty measure—can be modified to check how robust the results are with respect to a particular choice. Changing the prediction model from an ensemble to a single network did not alter the overall aggregation method ranking by much (fig. 3.11), suggesting that aggregation methods can be transferred to different pixel confidence methods. On average, mean confidence or mean foreground confidence were at the two last ranking positions, suggesting that they are a weak baseline, even though they are commonly used. The two pixel-wise uncertainty measures for ensembles explored in this study were predictive entropy (PE) and mutual information (MI). Comparing the

Table 3.2: AURC scores (×100) on the test sets for different pixel-level uncertainty measures (PE and MI) and aggregation methods, averaged across five prediction models trained on different folds. The best AURC (lower is better) is highlighted for each dataset row. Each prediction model is an ensemble of five networks. Performance gains by switching from PE to MI are largest for the kidney, Covid, and prostate datasets.

Aggregation →	mean		foregr. mean		non-boundary		patch-based	
Dataset ↓	MI	PE	MI	PE	MI	PE	MI	PE
Brain tumor (2D)	8.8	8.9	11.0	11.1	9.1	9.1	**8.4**	**8.4**
Brain tumor	14.0	16.6	**12.4**	12.5	13.3	13.6	13.8	14.0
Heart	13.6	13.8	14.5	14.6	12.6	**12.5**	13.0	13.3
Kidney tumor	12.8	14.3	13.3	14.5	13.1	13.7	**12.5**	13.3
Covid	30.5	33.1	34.5	36.1	**27.7**	30.3	32.4	35.3
Prostate	23.1	23.4	**22.7**	23.4	23.1	23.4	22.9	23.3

AURCs of four aggregation methods based on these two measures showed that MI had a slight advantage on most datasets.

3.2.1.3 Image-level methods

Moving on to failure detection methods that do not require a pixel confidence map (section 2.2.5.3), fig. 3.12 shows that the pairwise DSC achieved consistently the best AURC scores. It is applicable to any prediction method that produces multiple samples, such as deep ensembles or MC-Dropout. While ensembles worked overall best in the experiments reported here, pairwise DSC also performed well in combination with MC-Dropout, highlighting its flexibility and robustness. The second-best image-level failure detection method was the quality regression network. However, it revealed weaknesses in the prostate and Covid datasets. For the prostate dataset, this behavior was similar to the regression forests from the previous section, It could be due to the small training set size ($N = 26$), which may not have been enough to train a generalizable regression network. The Mahalanobis and VAE-based methods achieved significantly worse AURCs on most datasets than the aforementioned methods. As the Mahalanobis method was proposed for a single network, it suffered the disadvantage of using a weaker prediction model. AURC still allowed to fairly compare it to other methods, as it considers both segmentation performance and failure detection, which cannot be disentangled in practice either (see section 2.2.2.1). The VAE rarely performed better than the random baseline, indicating that it is either unsuitable for failure detection or difficult to adapt to new datasets.

The evaluation protocol developed within this thesis also enabled a comparison between the methods based on pixel confidence map aggregation in the previous section and the

Table 3.3: AURC scores (×100) on the test sets for different failure detection methods. Mean and standard deviation (std) are computed across five prediction models trained on different folds. A color map is applied on each 'mean' column, ranging from light yellow (worse) to dark green (best). PE was used for pixel uncertainty. Mean foreground aggregation and RF (radiomics features) were not included in the comparison, as they are similar to mean aggregation and RF (+ simple features), respectively. In this comprehensive comparison, pairwise DSC can also outperform methods based on pixel confidence aggregation. The quality regression network (the second-best image-level method) and the RF achieve comparable failure detection performance, switching ranks between datasets frequently. Abbreviations: Ens.: Ensemble, Single: Single network.

	Brain-2d		Brain		Heart		Kidney		Covid		Prostate	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Ens. + mean	8.9	0.3	16.6	0.4	13.8	0.7	14.3	0.5	33.1	1.5	23.4	0.8
Ens. + non-boundary	9.1	0.1	13.6	0.4	12.5	0.5	13.7	0.2	30.3	1.2	23.4	0.8
Ens. + patch-based	8.4	0.3	14.0	0.3	13.3	0.6	13.3	0.6	35.3	1.5	23.3	1.0
Ens. + RF (simple features)	7.9	0.1	11.5	0.2	12.5	0.7	10.4	0.3	25.4	0.2	36.5	2.2
Ens. + Quality regression	9.6	0.6	11.0	0.2	13.1	0.6	9.1	0.4	29.8	1.1	30.2	1.3
Ens. + pairwise DSC	7.6	0.1	10.5	0.1	11.8	0.5	8.4	0.2	24.0	0.3	23.0	0.6
Single + Mahalanobis	13.3	0.4	15.3	1.7	15.0	0.6	15.1	0.9	28.5	0.7	33.7	2.4
Ens. + VAE (seg)	12.0	0.5	22.5	1.2	19.3	2.9	14.7	0.6	37.4	1.1	25.8	0.9

image-level methods above. A selection of these methods is shown in table 3.3, and ranking stability is analyzed in section 3.2.3 (fig. 3.17). Pairwise DSC worked consistently best across datasets and was, hence, the overall winning algorithm for this benchmark. The best aggregation method from section 3.2.1.2 was a RF trained on simple features. Compared to the image-level methods presented here, it was closest to the performance of a quality regression network. This makes intuitive sense because both are regression models that learn to estimate the true DSC score. It is interesting to see, however, that the RF was competitive or even better on individual datasets, given that it operates on only five features.

To understand the strengths and weaknesses of the best-performing method, pairwise DSC, a deeper analysis of this method's results is provided in the following. This analysis focuses on its calibration, which means in this context that the pairwise DSC between ensemble predictions ideally is a good approximation of the true DSC, computed between ground truth and prediction. Although failure detection relies on the ranking of confidence scores, not their calibration, these two goals are closely related if the DSC is used as a risk function. Figure 3.13 shows that pairwise DSC was correlated with true DSC but also that it tended to overestimate the latter, especially on the Covid dataset. For all datasets, there were samples with high pairwise DSC and low true DSC, which are the practically most problematic because they are failures that would not be detected as such.

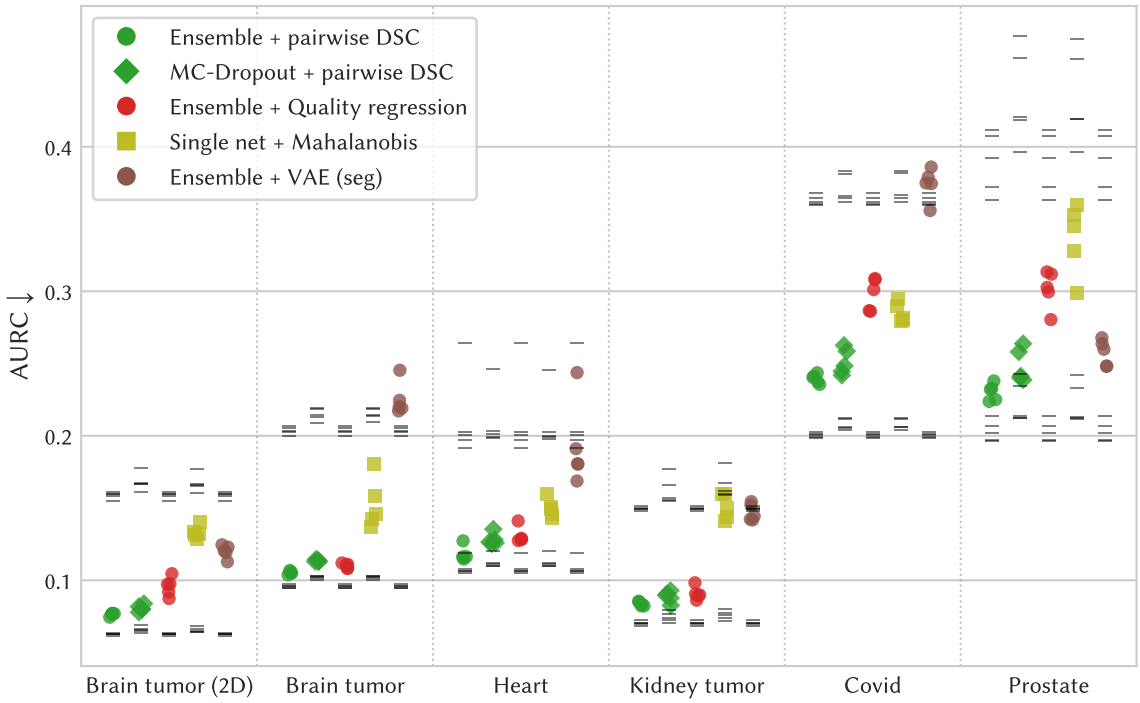

Figure 3.12: Comparison of image-level failure detection methods from section 2.2.5.3 in terms of AURC scores for all datasets (lower is better). Each colored marker corresponds to a single experiment, meaning a prediction model + aggregation method, trained on one fold of the training data. The marker shape represents differences in prediction models. Gray "–" markers visualize values for the random/optimal confidence baselines, which differ between the models trained on different folds due to their varying segmentation performance (section 2.2.2.2). Pairwise DSC performs clearly best across all datasets. The lowest AURCs are achieved in combination with the ensemble, but MC-Dropout also yields good results. Quality regression networks are usually the next-best option, but they show a performance drop on the Covid and Prostate datasets. Figure adapted from (Zenk et al. 2025b).

To gain intuition about the behavior of ensemble + pairwise DSC, fig. 3.14 shows qualitative results for a few selected failure cases from each dataset. In the Brain tumor (2D) dataset example, ensemble predictions consistently included a false-positive region in the posterior part of the brain, likely due to bias field artifacts. The ensemble members also disagreed on the extent of the actual tumor, resulting in low pairwise DSC. For the 3D brain tumor sample, cases of low-grade glioma (LGG) frequently exhibited inherent ambiguity in the tumor core region (orange), which resulted in inconsistent ensemble predictions and a reduced pairwise DSC.

In the heart example, segmentation errors occurred despite the clear visibility of the

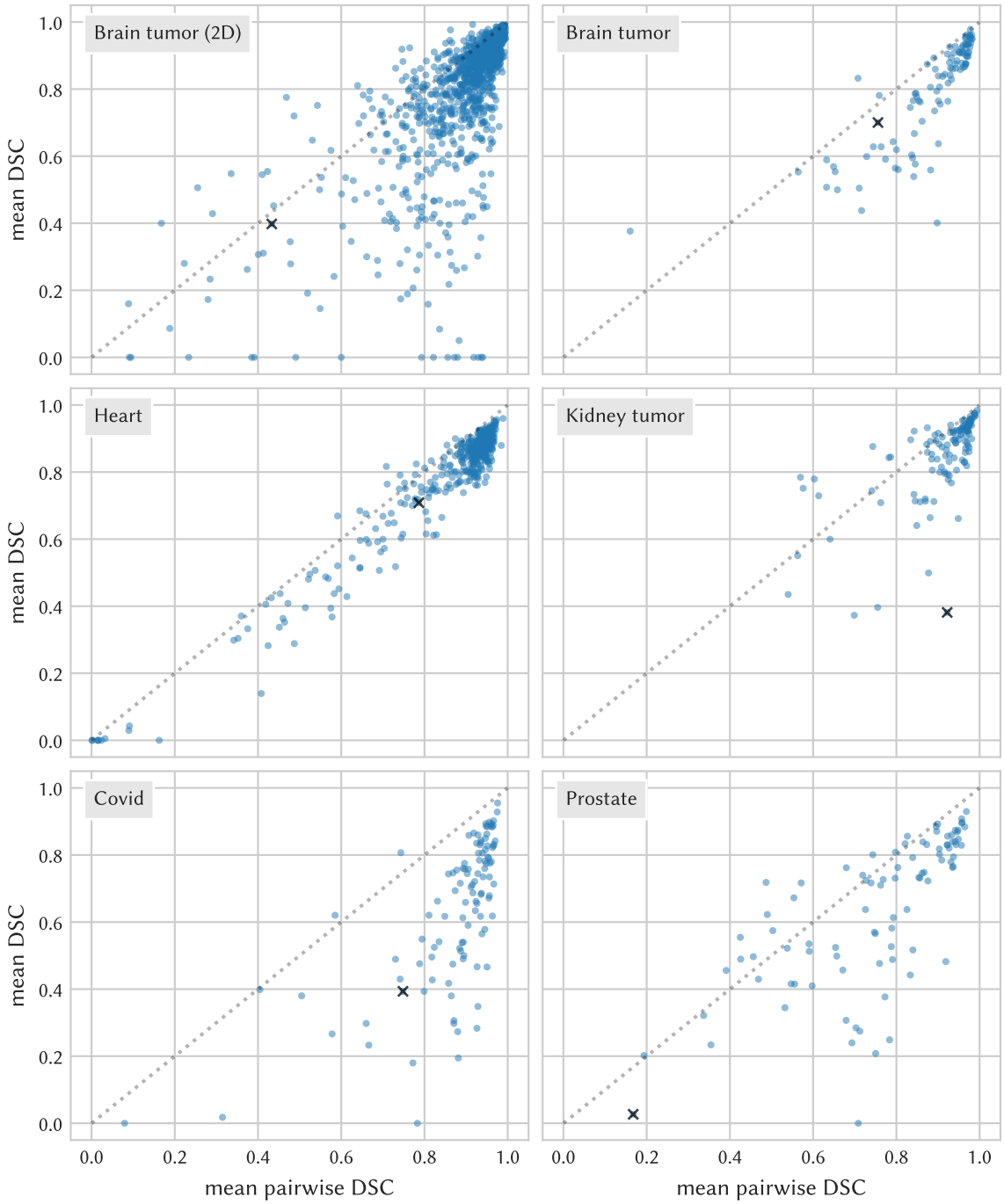

Figure 3.13: Scatter plot of confidence scores produced by ensemble + mean pairwise DSC versus true DSC scores for all datasets. Data samples marked with 'x' are visualized in detail in fig. 3.14. The correlation between the two quantities is clearly visible, but the pairwise DSC also tends to overestimate the true DSC, especially for the Covid dataset.

target regions. Variability between ensemble members was primarily observed in the right ventricle (green) region, while under-segmentation of the myocardium (orange) was not captured. Consistent false positives in the orange region further indicate that ensemble members often made similar mistakes, leading to potential issues with pairwise DSC reliability.

For the Kidney tumor CT image, the ensemble consistently made errors, such as segmenting masses beyond the kidneys or identifying additional kidneys. These obvious errors covered large regions and resulted in high inter-model agreement, making these failures harder to detect with pairwise DSC.

In the lung CT scan, the ensemble identified the rough location of Covid lesions, but their extent often deviated from the ground truth. The largest lesion was relatively consistent across ensemble members, leading to over-optimistic DSC scores. This observation was characteristic of the Covid dataset, as shown by fig. 3.13, and might point towards annotation inconsistencies between the training and test sets due to differences in annotators across the three subsets of the Covid data.

In the Prostate dataset, significant differences in acquisition techniques between training and test sets resulted in diverse image appearances and severe segmentation errors. In this example, the presence of an endorectal MR coil, which was absent during training, made ensemble predictions unstable. Due to the variability in predictions, the pairwise DSC was low, and these failures remain detectable.

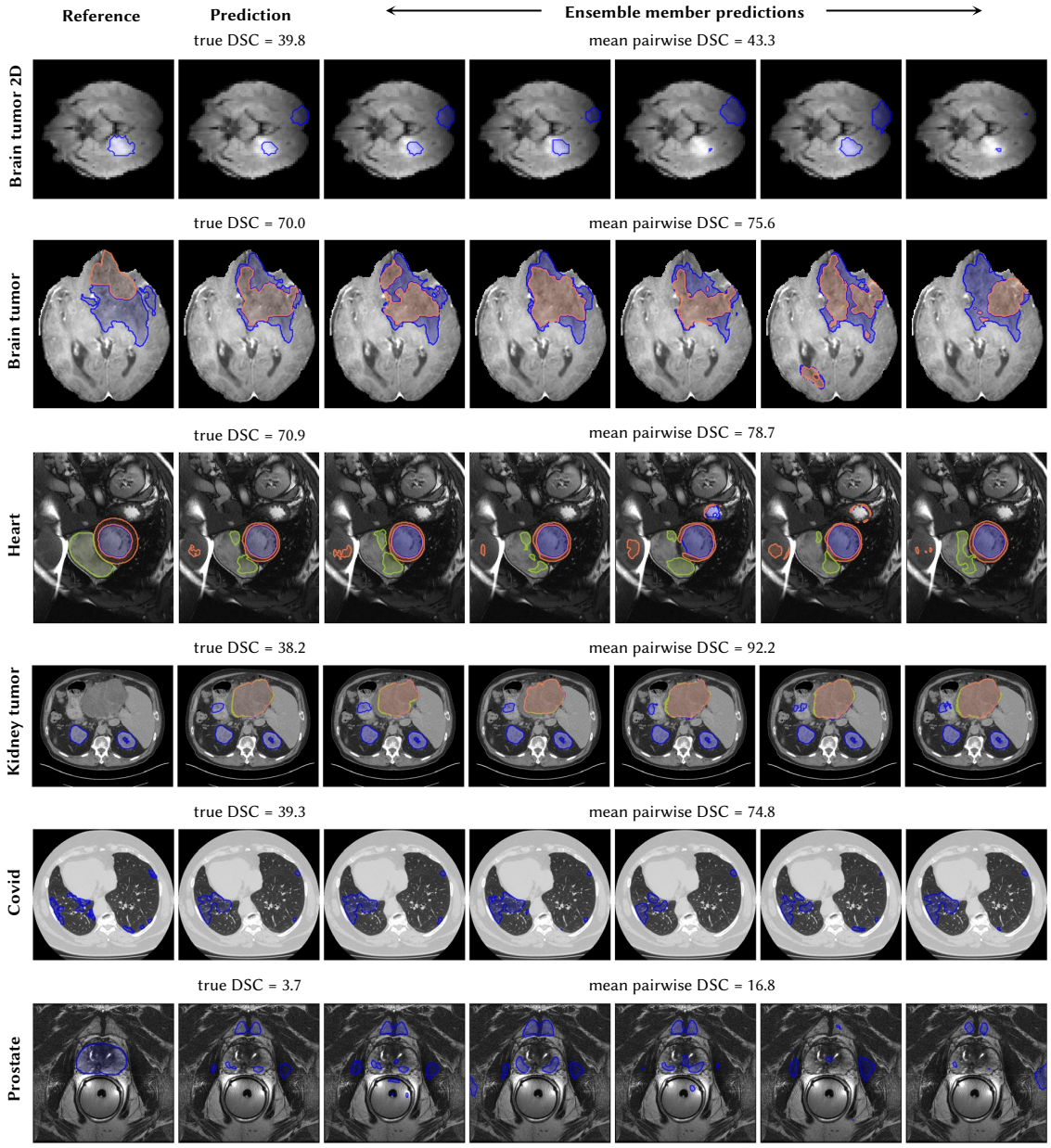

Figure 3.14: Qualitative analysis of ensemble predictions on all datasets (rows). Each row shows a failure case, together with the reference (ground truth), the ensemble prediction, and individual masks of its members, trained with different random seeds. The ensemble members often disagree in test cases with segmentation errors, which leads to low pairwise Dice as desired. However, all examples also contain regions where all ensemble members consistently predict a faulty segment. If such regions are large, as for the kidney tumor, they can result in an unnoticed failure. Figure adapted from (Zenk et al. 2025b).

3.2.2 Extensions of the Benchmark

While the results shown so far focused on the high-level question "Which failure detection method works best?", the benchmark proposed in this study can be easily extended to answer more questions on particular components of the failure detection system. Three examples are reported in this section, investigating how the results generalize to datasets from other medical imaging modalities than computed tomography (CT) and magnetic resonance imaging (MRI), how sensitive the methods are to changes in the segmentation backbone network, and how performance is affected when reducing the training set size.

3.2.2.1 Results on non-CT/MRI Datasets

As failure detection is also relevant for other modalities than CT and MRI, the comparison between different methods is extended here to include the additional datasets from section 2.2.3. Segmentation performance (fig. 3.15) revealed distinct patterns across datasets. The heart ultrasound (US) dataset achieved relatively high and consistent DSC scores, with only a few low-DSC failure cases, likely due to the absence of distribution shifts in the test set. In contrast, the optical coherence tomography (OCT) dataset exhibited low mean DSC scores, with a median DSCs for in-distribution cases only marginally above out-of-distribution (OOD) cases, which may be due to high inter-annotator variability. Finally, the large 2D optic disc/cup training set apparently did not prevent drops in DSC caused by distribution shifts in the test set.

The failure detection results (table 3.4) on the additional datasets were similar to the CT/MRI datasets: the pairwise DSC method consistently outperformed the other approaches. Among the remaining methods, performance differences were minor when comparing within the same prediction model, that is, a single network or ensemble. A notable deviation from previous results was that the Mahalanobis method ranked second for the heart ultrasound dataset, potentially indicating that samples with low DSC for this dataset exhibited characteristics of OOD samples. For the OCT dataset, the variance across different training folds was high, which can be explained by the small size of the training and test sets.

These observations suggest that the heart ultrasound dataset did not contain enough failure cases and that the OCT dataset was too small for reliable failure detection evaluation, so neither of them was included in the main benchmark. The 2D optic disc/cup dataset has the advantage that it can be used for fast experimentation, similar to the brain tumor 2D dataset, so it can be recommended for future benchmarking efforts.

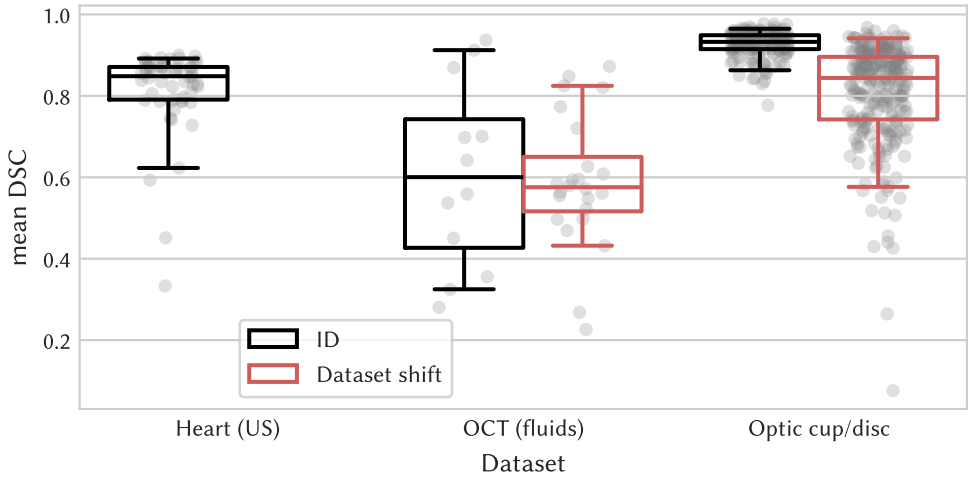

Figure 3.15: Segmentation performance for all test cases (gray points), measured by DSC, of a single U-Net trained on the first fold of the training data. Boxes show the median and IQR of the distribution, while whiskers extend to the 5th and 95th percentiles, respectively. Two datasets also include samples with distribution shifts besides the samples that originate from the same distribution as the training set (in-distribution, ID). While the Optic cup/disc dataset displays a clear performance gap between the ID and distribution shift subsets, the heart US dataset has only a few failure cases. DSC scores have high variance for the OCT dataset. Figure adapted from (Zenk et al. 2025b).

Table 3.4: AURC scores (×100) for the same methods as in table 3.3, but with three datasets from different modalities (US, OCT and fundus RGB photograph). Mean and standard deviation (std) are computed across five prediction models trained on different folds. The color map applied to each 'mean' column ranges from light yellow (worse) to dark green (best). PE was used for pixel uncertainty. Pairwise DSC outperforms the other methods by a clear margin on these additional datasets, while the differences between the remaining methods' AURCs scores are smaller. Table adapted from (Zenk et al. 2025b).

	Heart-US		OCT-fluids		Optic c/d	
	mean	std	mean	std	mean	std
Ensemble + mean	14.9	0.2	26.5	2.4	9.5	0.1
Ensemble + non-boundary	14.6	0.2	25.3	2.6	9.4	0.1
Ensemble + patch-based	15.0	0.3	25.4	2.9	9.3	0.2
Ensemble + RF (simple features)	15.4	0.3	27.4	2.3	10.3	0.5
Ensemble + Quality regression	15.2	0.4	27.3	2.6	9.6	0.5
Ensemble + pairwise DSC	13.6	0.2	22.9	3.5	8.6	0.1
Single net + Mahalanobis	13.9	0.2	38.7	5.2	10.3	0.3
Ensemble + VAE (seg)	15.4	0.1	27.3	2.4	9.6	0.2

Table 3.5: AURC scores (×100) of failure detection methods with different segmentation backbones on the heart and kidney tumor datasets. The *default*, *residual* encoder, and *wide* U-Net backbones are included for the heart dataset, but the wide U-Net ran out of memory for the kidney tumor dataset. Mean AURC are computed across three prediction models trained on different folds. A color map is applied within each dataset block (6 × 3 cells for heart, 6 × 2 cells for kidney), ranging from light yellow (worse) to dark green (best). Only minor changes in the ranking order can be observed when switching between backbones, with the pairwise DSC consistently dominating. Absolute differences in AURC between backbones are small on the heart dataset, but the residual encoder improves the scores significantly on the kidney dataset. Table adapted from (Zenk et al. 2025b).

Dataset:	Heart			Kidney tumor	
Backbone:	default	residual	wide	default	residual
Failure detection method					
Single net + mean	14.8	16.8	17.1	16.0	14.6
Single net + Mahalanobis	15.1	14.4	16.0	14.8	16.5
Single net + RF (simple features)	13.2	13.4	13.4	11.3	8.6
Ensemble + Quality regression	13.2	13.5	13.1	8.9	7.7
MC-Dropout + pairwise DSC	12.9	13.1	12.9	8.8	8.4
Ensemble + pairwise DSC	12.0	11.8	11.9	8.3	6.8

3.2.2.2 Influence of the Segmentation Backbone

The segmentation backbone can have two different effects on failure detection performance: On the one hand, the segmentation accuracy can be improved with suitable architectures, which also reduces the overall risk of failures and, simultaneously, AURC. On the other hand, switching the network architecture also has an effect on the feature representations, which might, in turn, affect the confidence scores of some failure detection methods.

An extensive evaluation of different backbone architectures is beyond the scope of this thesis, but to demonstrate how the proposed benchmark can be used for studying the impact of network architectures, two variants of the default backbone were compared on the heart and kidney tumor datasets. The default backbone was a U-Net, while the variants were: (i) A "wide" U-Net, which doubles the number of filters per convolutional layer of the default. (ii) A residual encoder U-Net, which has a larger encoder with residual connections, as described in section 2.2.4. Due to graphics processing unit (GPU) memory limitations, the wide U-Net could not be trained on the kidney tumor dataset.

The results reported here were obtained by averaging the test set scores of three independent runs, which were trained on different folds of the training data. The effect of the architecture on the segmentation performance was mixed: For the heart dataset, the default

U-Net achieved the best performance (mean DSC = 0.784), while the variants suffered performance drops (mean DSC = 0.780 for the wide U-Net and 0.761 for the residual encoder U-Net). Overfitting or lack of OOD robustness are possible reasons for these drops. The situation was reversed for the kidney tumor dataset, for which the residual encoder U-Net boosted the mean DSC to 0.868, compared to 0.833 for the default U-Net. Some of this gain can be attributed to the increased patch size for the residual encoder U-Net on the kidney tumor dataset.

Evaluating the failure detection results revealed that the clear differences in DSC scores on the heart dataset did not directly translate to AURC values. The RF and quality regression methods apparently could compensate for the additional failures, and pairwise DSC even slightly outperformed the default U-Net. On the kidney tumor dataset, AURC decreased for all methods except Mahalanobis when using residual encoder U-Nets, as expected based on the better segmentation performance. Although some methods swapped ranks for different architectures, for example quality regression and MC-Dropout + pairwise DSC on the kidney tumor dataset, the overall ranking obtained in the main benchmark was largely stable across backbones. This suggests that the segmentation backbone affects most failure detection methods similarly.

3.2.2.3 Influence of the Training Set Size

The success of deep learning models critically depends on the amount of training data, yet small sample sizes are common for medical image segmentation datasets. Failures occur more frequently in applications with scarce training data, making failure detection methods especially important for them. This section reports the results of experiments on the impact of the training set size on the failure detection methods. The heart dataset was chosen for this purpose, as it has large training and test sets and also small image shapes, allowing for fast experimentation. The number of training cases was halved iteratively from the original 152 down to a minimum of 8 scans, which correspond to 4 patients, as the heart dataset has two scans per patient.

Figure 3.16 shows how the segmentation and failure detection performance evolved with varying training set size. A single fold was trained for each method. DSC scores remained high even with 38 training cases but dropped below this number, especially for samples with distribution shift. Although the AURC also worsened with few training samples for the pairwise DSC method, the difference to the optimal AURC stayed small. Especially the Mahalanobis and RF methods, but also the quality regression network suffered from small training sets, falling behind the single network + mean baseline in the ranking for 8 training cases. This is additional evidence of the weakness of quality regression and RF with small training sets, which was also observed in the main benchmark.

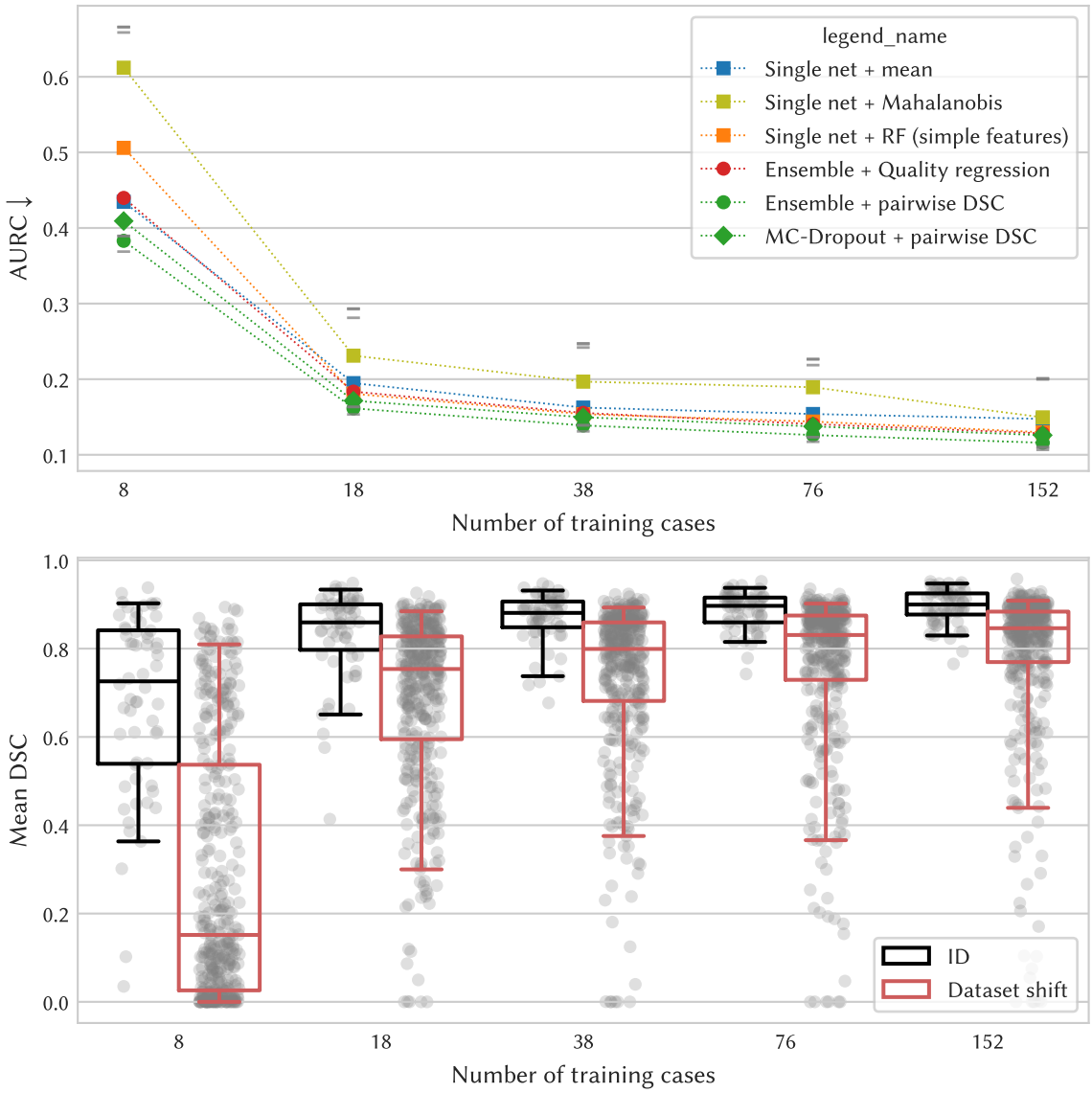

Figure 3.16: Experimental results obtained when varying the number of training samples in the heart dataset, while keeping the test set fixed. Top: AURCs for different failure detection methods. Different colors represent the methods and different markers the underlying prediction model. The random and optimal performance baselines are marked with gray horizontal markers. Bottom: DSC values for a single U-Net trained on datasets with varying sizes. The test set consists of in-distribution (ID) and distribution-shifted cases. Performance worsens for both AURC and DSC with smaller training sets, but pairwise DSC stays close to the optimal baseline. Figure adapted from (Zenk et al. 2025b).

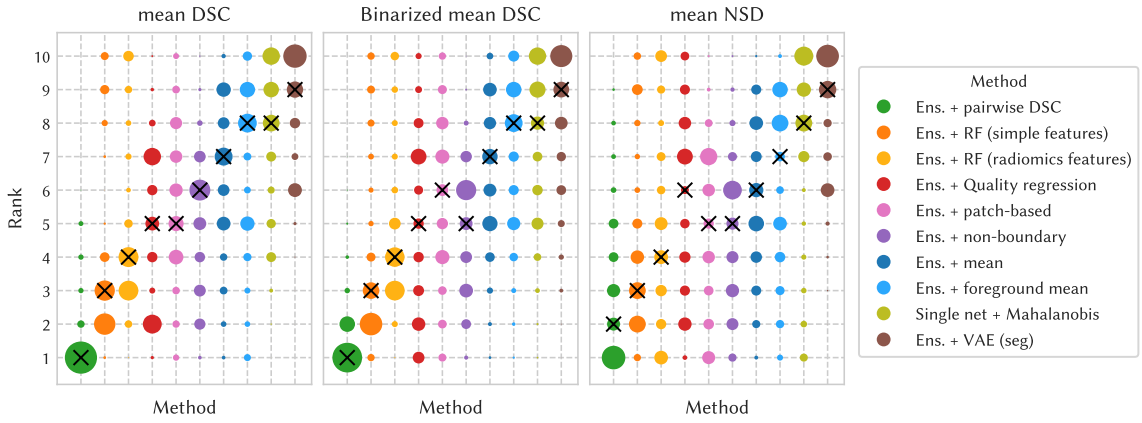

Figure 3.17: Ranking distribution plots based on 1000 bootstrap samples, compared between different risk functions (mean DSC, binarized mean DSC, and mean NSD). The results of all training set folds and datasets were accumulated. The circle area is proportional to the frequency with which a method occupies the corresponding rank on the y-axis. Black 'x'-markers indicate median ranks. The method order is consistent between all sub-plots and identical to the ranking order based on the mean DSC risk. Overall, the ranking distributions are similar for the three risk functions, but the variance in the ranking distributions is higher for binarized DSC and NSD. This shows that the choice of failure detection method is only moderately affected by the risk function changes introduced here. Figure adapted from (Zenk et al. 2025b).

3.2.3 Analysis of the Evaluation Protocol

While the previous sections focused on which failure detection method works best, here a closer look is taken at the evaluation protocol described in section 2.2.2.2, which is a central contribution of this benchmark. The evaluation protocol mainly relies on two components: the risk function, which measures the degree of failure/risk associated with a segmentation mask, and the failure detection metric, which summarizes the experimental data of confidence scores and risks into a scalar measure of performance. For both of these components, alternatives to the methodology chosen for the benchmark are investigated here.

First, the benchmark ranking is recomputed after switching the risk function. The original risk function was $1 - \text{DSC}$ (see eq. (2.6)), which was substituted with two alternatives before re-running the evaluation pipeline: (a) Binarized DSC converted the continuous risks based on DSC into binary failure/success labels, by applying a dataset-specific (manually chosen) threshold. (b) normalized surface dice (NSD) replaced the DSC with the NSD metric, which is a boundary-based metric for segmentation tasks that is complementary to DSC. Evaluating the same experiments with these risk functions

yielded bootstrapped ranking distributions shown in fig. 3.17. For most methods, the median ranks stayed constant across risk functions; in a few cases, changes of at most 1 rank occurred. The higher variance in the ranking distributions for binarized DSC and NSD risks indicate lower ranking stability, however. The main conclusion from this analysis is that the proposed evaluation protocol allows to choose the risk function flexibly, and that the benchmark's findings are robust to moderate changes in the risk function.

The second part of this analysis focuses on failure detection metrics. While this benchmark used AURC as the main metric, alternatives from the literature are the Pearson correlation coefficient (PC) and Spearman correlation coefficient (SC). Notably, SC only considers the correlation between confidence and risk *ranking*. Table 3.6 compares the benchmark results when evaluating these three failure detection metrics. The same subset of methods as in table 3.3 is shown for clarity. Within the groups of three columns for each dataset, the relative performance differences between methods (visible in the color map) were similar between AURC, PC and SC. While individual deviations existed, the method identified as best in the benchmark, pairwise DSC, also achieved the highest scores for the alternative metrics. Hence, the choice of failure detection metric seems to have little impact on the overall benchmark ranking. However, weaknesses of metrics that are purely based on the confidence ranking (such as SC) while ignoring segmentation performance can have consequences in special cases. Figure 3.18 demonstrates, for example, that SC favored a single network as the prediction model over an ensemble when quality regression was used as a failure detection method, although the ensemble achieved lower risk for virtually all coverage levels. In contrast, the AURC was lower (better) for the ensemble, as expected from the risk-coverage curve. The same trend was found across datasets for the quality regression method (table 3.6). This observation confirms the importance of considering segmentation performance in a fair failure detection evaluation, as initially argued in requirements R2 from section 2.2.2.1.

The analysis in the previous two sections replaced the risk function or failure detection metric in the evaluation pipeline. This final analysis revisits requirement R1 in section 2.2.2.1 and determines the effect of evaluating a different but seemingly similar task. OOD detection is such a proxy task, which is often used for evaluating new methods, although failure detection is explicitly stated as its goal. To evaluate OOD detection, test set samples were labeled as OOD and in-distribution (ID), so that classification metrics like area under the receiver operating curve (AUROC) could be applied. The confidence scores of the evaluated methods were then used for classification. For four datasets from this benchmark, OOD labels could be created by treating samples with distribution shifts as OOD and the rest as ID. Applying this evaluation protocol on the experimental data resulted in the scores depicted in fig. 3.19, which produced a completely different ranking than the failure detection results. One of the worst methods in terms of AURC, the

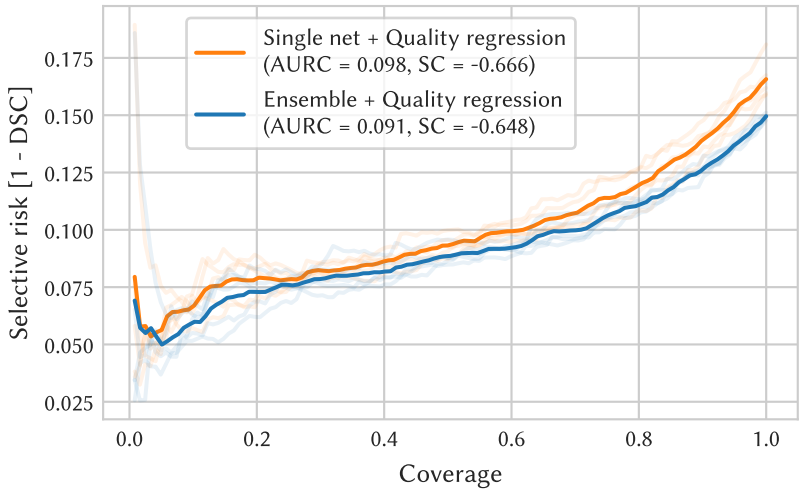

Figure 3.18: Comparison of the risk-coverage (RC) curves of two failure detection methods that differ only in the prediction model (single net versus ensemble). Solid lines are the mean over five repetitions of the experiment, which trained prediction models on different folds. Faint lines depict RC curves of individual runs. SC values seen in the legend assign better performance to the single net model. At the same time, AURC favors the ensemble, raising the question of which metric better captures the practically relevant performance aspect. The RC curves show that the ensemble is superior for virtually all coverage levels, which suggests that AURC is better suited to compare failure detection methods that use different prediction models.

Mahalanobis detector, now achieved the highest AUROC values across datasets. This illustrates that OOD and failure detection are indeed distinct tasks which require specialized evaluation protocols.

Table 3.6: Comparison of the failure detection metrics AURC, Spearman correlation coefficient (SC), and Pearson correlation coefficient (PC) for different failure detection methods. Due to space limitations, the six datasets are distributed over two sub-tables. Values are averaged across 5 training folds, and background color coding is applied per column, ranging from light yellow (worst) to dark green (best). Within each dataset group, the colors are similar between the failure detection metrics, which indicates that the benchmark results do not change significantly when using these alternative, correlation-based metrics. In particular, pairwise DSC remains the best-performing method. Abbreviations: Single = Single network; Ens. = Ensemble. Table adapted from (Zenk et al. 2025b).

Dataset	Brain tumor (2D)			Brain tumor			Heart		
Metric	AURC	PC	SC	AURC	PC	SC	AURC	PC	SC
Method									
Ens. + mean	0.089	-0.596	-0.693	0.166	-0.258	-0.300	0.138	-0.533	-0.618
Ens. + non-boundary	0.091	-0.489	-0.619	0.136	-0.336	-0.569	0.125	-0.795	-0.725
Ens. + patch-based	0.084	-0.611	-0.737	0.140	-0.441	-0.489	0.133	-0.638	-0.679
Ens. + RF (simple features)	0.079	-0.686	-0.787	0.115	-0.741	-0.797	0.125	-0.686	-0.772
Single + Quality regression	0.097	-0.501	-0.584	0.115	-0.786	-0.862	0.134	-0.662	-0.637
Ens. + Quality regression	0.096	-0.432	-0.556	0.110	-0.759	-0.840	0.131	-0.659	-0.620
Ens. + pairwise DSC	0.076	-0.697	-0.821	0.105	-0.812	-0.884	0.118	-0.964	-0.819
Single + Mahalanobis	0.133	-0.126	-0.314	0.153	-0.470	-0.491	0.150	-0.356	-0.448
Ens. + VAE (seg)	0.120	-0.028	-0.392	0.225	0.174	0.222	0.193	-0.358	-0.137

Dataset	Kidney tumor			Covid			Prostate		
Method	AURC	PC	SC	AURC	PC	SC	AURC	PC	SC
Ens. + mean	0.143	-0.264	-0.001	0.331	-0.124	-0.247	0.234	-0.713	-0.792
Ens. + non-boundary	0.137	-0.305	-0.056	0.303	-0.163	-0.358	0.234	-0.699	-0.785
Ens. + patch-based	0.133	-0.203	-0.135	0.353	-0.059	-0.030	0.233	-0.692	-0.802
Ens. + RF (simple features)	0.104	-0.538	-0.484	0.254	-0.603	-0.600	0.365	-0.066	-0.065
Single + Quality regression	0.098	-0.715	-0.666	0.305	-0.440	-0.374	0.319	-0.472	-0.491
Ens. + Quality regression	0.091	-0.703	-0.648	0.298	-0.421	-0.355	0.302	-0.388	-0.443
Ens. + pairwise DSC	0.084	-0.672	-0.753	0.240	-0.664	-0.708	0.230	-0.735	-0.811
Single + Mahalanobis	0.151	-0.055	-0.028	0.285	-0.370	-0.426	0.337	-0.346	-0.390
Ens. + VAE (seg)	0.147	-0.252	0.067	0.374	0.040	-0.058	0.258	-0.333	-0.646

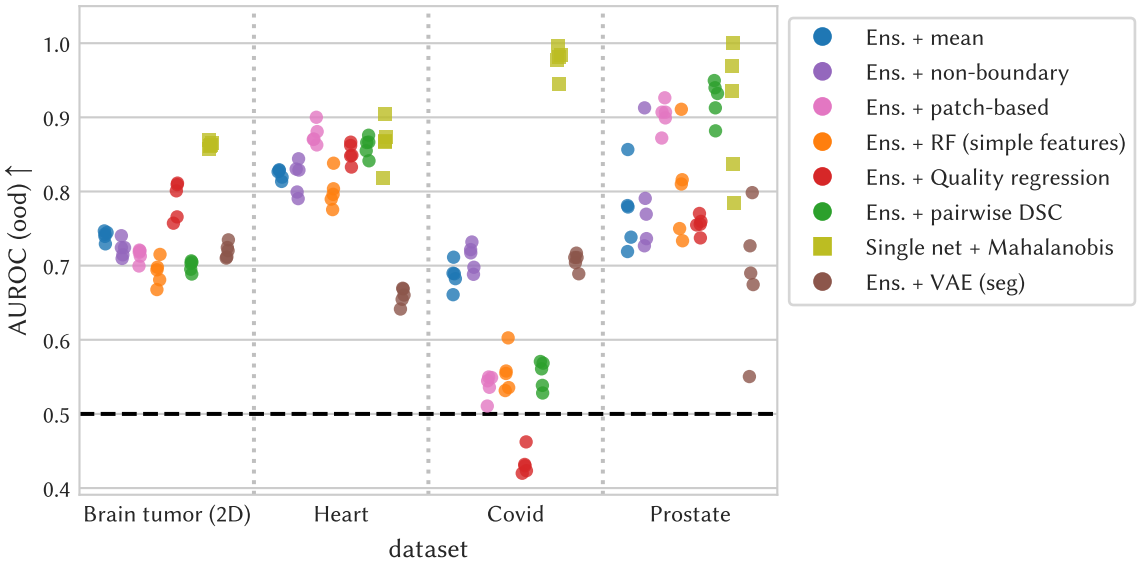

Figure 3.19: Comparison of OOD-AUROC scores for all datasets and different failure detection methods (higher is better). Each experiment was repeated using 5 folds, corresponding to the different markers per method. Compared to the main benchmark, the kidney tumor and brain tumor datasets were excluded because they do not contain samples with distribution shifts in the test set. The dashed black line indicates random performance. Single net + Mahalanobis consistently achieves the highest scores, while the best method for failure detection, pairwise DSC, clearly falls behind in the ranking. This suggests that OOD and failure detection are distinct tasks. Figure adapted from (Zenk et al. 2025b).

4 Discussion

Two aspects of the robustness of medical image segmentation methods were studied in this thesis: their ability to generalize to unseen hospitals and to estimate how reliable the predicted segmentation is, in particular also flagging potential failures. The discussion below follows this structure, focusing first on the generalization aspect (section 4.1) and then on failure detection (section 4.2).

Disclosure

Section 4.1 is based on the manuscript summarizing the FeTS Challenges, which has been accepted for publication (Zenk et al. 2025a), so portions of the text resemble the original manuscript text.
Section 4.2 is derived from a previously published article (Zenk et al. 2025b), so portions of the text resemble the original manuscript, in accordance with the publisher's license.
If parts of the text replicate sections from the corresponding manuscripts, this is explicitly stated beforehand.

4.1 Generalization

The FeTS Challenges 2021 and 2022 were the first competitions to evaluate submitted algorithms in a real-world federation, keeping the test data decentralized. In 2022, 41 models were evaluated in the medical image analysis task of brain tumor segmentation at 32 institutions to test their robustness on diverse datasets. The 2021 challenge was a preceding pilot study on a smaller scale (3 models on 22 institutions), so the discussion here focuses on the findings from 2022.

4.1.1 Interpretation of the Challenge Results

From the perspective of the research questions (RQs) formulated initially (page 28), the challenge results can be interpreted as follows.

RQ 1.1 concerns generalization "in the wild", which first raises the question of how closely the FeTS Challenge approximates an "in the wild" scenario or, equivalently, how wide was the range of possible clinical application cases covered. Although the geographical diversity was high, the majority of patients still originated from North America and Europe. The metadata collected from the federation showed a diverse patient population in terms of age, sex, and genetic tumor variants, as well as a variety of image acquisition settings covering different scanner models and protocols. Given that 9 and 14 independent institutions from Europe and North America (table 2.2 and fig. 2.2) contributed to the test sets, respectively, it is plausible that the FeTS Challenge simulates an "in the wild" scenario at least in these regions. The main interest of RQ 1.1 was the generalization of brain tumor segmentation algorithms, which was quantified in section 3.1.2.3 in terms of segmentation metrics, finding that the models generalized well to most institutions but also showing a consistent performance drop at a small subset of institutions (fig. 3.3). Moreover, the average-case performance ranged from 0.84 to 0.98 across the evaluated institutions, while the worst-case performance ranged from 0.37 to 0.96 (fig. 3.4). Determining the targeted performance level based on human inter-rater variability would require multiple annotators for each patient, which was not feasible for the FeTS Challenge. As a rough estimate, in a previous challenge on brain tumor segmentation (Menze et al. 2015), the authors measured a mean inter-rater variation of 0.78 Dice similarity coefficient (DSC). Hence, based on the experimental results, worst-case robustness was problematic at 13 of 32 institutions, demonstrating that brain tumor segmentation is still not solved "in the wild" (RQ 1.1).

RQ 1.2 asked how algorithm and dataset characteristics influence generalization. Due to the competition format of the study, *algorithm characteristics* could only be extracted from the participants' reports, and ablation studies were difficult to implement. Therefore, specific components of the algorithms that improved generalization could not be isolated. Common characteristics among the top-ranked teams included the use of 3D U-Nets, especially implementations based on nnU-Net. Ensembling methods and brain tumor-specific post-processing were also associated with performance gains. The initial hypothesis of the FeTS challenge was that the institution partitioning information provided for the training data would be beneficial for developing generalizable models, as it contains information about distribution shifts. However, only one team developed a method that explicitly addressed such shifts, while the rest trained their models on the pooled training data without considering its multi-institutional nature. Further studies are needed to de-

termine whether domain generalization or test-time adaptation methods can improve the robustness of brain tumor segmentation algorithms. Regarding *dataset characteristics*, the challenge results showed that segmentation performance on institutions that contributed to the training data was, surprisingly, similar to or even worse than the results on unseen institutions. There could be two reasons for this: (a) The data from unseen institutions was so similar to data from training institutions that there was no generalization gap, or (b) differences in annotation style and quality led to noise in the segmentation metrics, obscuring differences in generalization. The annotation quality control experiments, which covered about half of the test cases, provided some evidence that annotation mistakes were not a major issue in the test dataset. It is plausible, however, that the segmentation metric values are slightly optimistic because of the semi-automatic annotation procedure. In this procedure, nnU-Net was part of a model ensemble that produced an initial mask, which was later refined by local experts. The refined reference segmentations are likely biased toward this model, which is very popular among challenge participants. In fact, this issue is not specific to the FeTS challenge and highlights the importance of annotation quality control. Assuming that annotation differences between institutions only had a minor effect, the challenge results and qualitative analysis of individual predictions (section 3.1.2.3) did not identify institution-specific dataset characteristics that led to failures. Instead, they found associations between performance drops and properties of individual test cases. In the FeTS challenge, only qualitative properties of failure cases were determined. To systematically study case-specific failure modes (as done in Roß et al. (2021), for example), additional meta-data collection or annotation is necessary.

RQ 1.3 aimed at the previously unknown practical hurdles of federated evaluation when used for image analysis competitions. While section 3.1.2.6 described specific hurdles, here federated competitions are compared to traditional ones, for which the test data is centralized at the organizers' institution. The comparison is structured according to the efforts required before, during, and after the official competition period. In the preparation phase, after defining the challenge mission and evaluation, data must be collected, annotated, and quality controlled. While data collection and annotation can be parallelized and scaled more easily for federated challenges, quality control is significantly harder due to the limited access to raw data and the large number of independent annotators. Additional federated infrastructure increases the workload compared to a centralized setup. Executing a federated challenge involves higher organizational and technical hurdles than centralized challenges. Algorithm submissions must be compatible with diverse hardware systems, and coordinating a federated evaluation via personal communication can become a bottleneck. For the FeTS Challenge, the duration varied widely between different collaborators (from a few weeks to 6 months), mostly determined by their responsiveness. After the ranking is computed based on the raw challenge results,

the analysis of federated and centralized competitions is similar, as long as no access to the imaging data is required. If the latter is federated, qualitative analyses of individual failure modes are much more difficult than with centralized data. Maintaining a federated benchmark over multiple years is attractive because the initial setup costs are amortized. However, changes in hardware, staff, or regulations may require additional efforts by organizers and are more likely when multiple collaborators are involved. In conclusion, federated competitions require additional effort and time in all challenge phases. If these hurdles can be overcome, a federated design allows scaling up the test dataset size and diversity considerably. Therefore, federated evaluation may be best suited for a future type of "phase 2" challenge, which is organized after the performance in a medical image analysis task has saturated on small research datasets, to focus on algorithmic robustness on a large scale. Finally, federated challenges are also valuable in that they establish collaborative networks, which can provide long-term research output beyond the original competition setting.

4.1.2 Comparison to Related Work

The FeTS challenge is a direct extension of the long-standing BraTS challenge series (Menze et al. 2015; Bakas et al. 2019; Baid et al. 2021) towards benchmarking generalization in unseen hospitals. Bakas et al. (2019) found good median in-distribution generalization, but also outliers that highlight a potential lack of robustness. They speculated that these outliers might vanish if more training data is added, but the FeTS challenge showed that such failure cases still occur when training on a large cohort of 1251 patients. The latest instance of the BraTS challenge (Bakas et al. 2024) continued the line of work on generalization to distribution shifts started with the FeTS Challenge, but no results have been published so far. Common segmentation errors of automated segmentation methods observed in previous BraTS challenges (Bakas et al. 2019; Baid et al. 2021) persisted in the FeTS challenge. The two most prominent issues found in this study (bright blood products and unclear non-enhancing tumor core) have not been described in detail so far, and their presence in the training data was found during the annotation quality control of the FeTS challenge, which is an important step towards future data cleaning efforts.

As detailed by the related work section 1.3.1, there are a few other international competitions that focus on generalization across different clinics. The M&Ms challenge (Campello et al. 2021) evaluated methods on data acquired at new centers with different scanners. However, they investigated the task of cardiac segmentation and chose the conventional approach, in which multi-centric data is collected centrally. Consequently, their dataset was significantly smaller: the test set comprised 160 test cases from six centers, of which two centers (64 cases) did not contribute to the training data. The advantage of their

central collection is that the number of test samples per scanner vendor was more balanced. The M&Ms challenge found that state-of-the-art segmentation models with strong data augmentation are currently still superior to methods for domain/test-time adaptation, which agrees with the FeTS challenge results. The MIDOG challenge by Aubreville et al. (2023) focused on the task of detecting cells undergoing division (mitosis) in whole slide images acquired in digital pathology. Despite the different image analysis task and imaging modality, its goal to evaluate the robustness of algorithms when applied to data from different scanners is similar to the FeTS challenge's objective. In addition to the four scanners in the training set, two more scanners were used for testing. As in the FeTS challenge, there was not a single algorithmic design choice that dominated the ranking, and ensembling of model outputs appeared to be beneficial. A third benchmark on out-of-distribution generalization focused on fetal brain tissue segmentation in in-utero magnetic resonance imaging (MRI) scans (Payette et al. 2024). As in the preceding competitions, the test set was small compared to the FeTS challenge, with two in-distribution hospitals (80 cases) and two out-of-distribution hospitals (80 cases). Similarly to the FeTS challenge results, they found that the top teams achieved comparable performance on hospitals seen and unseen during training, respectively. This lack of apparent generalization gap was partially explained by the heterogeneous image quality, which varied both for in-distribution and out-of-distribution samples. Data augmentation and using nnU-Net configurations were reported as an effective strategy in the FeTA challenge, in agreement with the FeTS challenge results. Overall, several trends from related work were confirmed in the FeTS challenge on a significantly larger scale, in terms of test set size and diversity.

Practical hurdles for federated image analysis workflows have rarely been reported or were not based on real-world experiences. Recently, Bujotzek et al. (2024) published a guide to real-world federated learning for radiology, which also describes related practical hurdles. Although federated learning is arguably even more challenging than federated evaluation, as it involves multiple rounds of communication and local model training in addition to regular evaluation runs, parts of the infrastructure and workflows are similar. The experiences from the FeTS challenge confirm several of the hurdles reported in (Bujotzek et al. 2024), such as inconsistencies in annotations, varying duration of federated workflows, and high debugging efforts. Moreover, the competition format of the FeTS Challenge led to the special situation that many participants contributed algorithms, which caused compatibility issues at some institutions. This is less of a concern for federated learning studies that include only a small team of algorithm developers, but a crucial aspect for federated competitions. Overall, the congruence with findings from (Bujotzek et al. 2024) highlights that the practical problems identified in the FeTS challenge are common for federated projects, calling for the development of new solutions to existing organizational, infrastructure, and data quality hurdles.

4.1.3 Limitations and Future Work

As a pioneering project in combining competitions with federated data, the FeTS challenge also has limitations that could be tackled in future work.

The number of original algorithms submitted to the challenge was low (three in 2021 and five in 2022). Although 36 additional algorithms were evaluated in 2022, they were previously published models developed without the multi-centric generalization objective in mind. Therefore, the methodological innovation through the FeTS challenge was small. Two potential avenues for increasing the participation are: (i) make the training data more interesting for researchers; (ii) better convey the challenge's concept to the community. Regarding (i), many participants join a challenge because of the novel datasets published with it. The FeTS challenge re-used the BraTS challenge training data and added institution partitioning files (section 2.1). Although the high number of registrations in FeTS22 (35) suggests that the challenge data stirred interest in the research community, it appears that only a few (6 teams made submissions) found it promising enough to compete in the challenge. Extending the imaging data or meta-data, or balancing the partitioning could provide additional incentives for researchers to engage in the generalization task. Regarding (ii), the FeTS challenge combined the generalization task (Task 2) reported here with a simulated federated learning task (Task 1). As these tasks require different methods, separating the two into independent challenges may increase visibility and help attract more participants. Future challenges with multi-site generalization evaluation should therefore stress the generalization aspect more than the federated aspect in publicly communicated material such as the title or webpage.

The federated test data used for the competition was large and geographically diverse, but meta-data was only available in the form of site-specific statistics. This limited the insights that could be gained from the challenge results. Gathering case-specific meta-data can be valuable for quantifying distribution shifts and dataset diversity in real-world use cases, which helps with designing benchmarks like the FeTS challenge and estimating performance in real-world clinical deployments more accurately. Collecting such meta-data in a challenge is also useful for analyzing the data characteristics that are related to the generalization capabilities of segmentation models. As patient-specific meta-data can contain personal information, privacy-preserving methods may be required for such investigations in the future. Furthermore, qualitative analyses on individual subjects (such as an error analysis) are especially difficult in a federation without direct data access, so tools and agreements for secure remote visual data inspection or limited data sharing would be helpful.

As described in section 3.1.2.6, setting up the federated infrastructure for FeTS and running the federated evaluation required the combined efforts of multiple institutions

over several months or even years, when considering the previous work (Pati et al. 2022a) that enabled the FeTS challenge. This high initial cost is an obstacle for future, similar endeavors. Streamlining federated workflows, for example through enhanced error reporting and debugging tools, could facilitate faster, more efficient implementation of federated challenges. Tools for data curation and quality control are important building blocks in federated projects, too. Quality control is also needed for annotated segmentation masks. In the visual inspection performed for the FeTS challenge (section 3.1.2.2), 10 % of the checked samples were excluded due to annotation quality issues, some of which highlighted inconsistencies that are also present in the training set, calling for future actions to homogenize the annotations for brain tumor segmentation tasks. For about half of the federated test data, quality control was restricted to automatic sanity checks, which should in the future be replaced with visual inspection, possibly guided by automatic quality estimation methods.

4.2 Failure Detection

The benchmark design from section 2.2 unified the evaluation of methods that were previously studied separately but serve the same purpose of failure detection. This enabled the comparison of a wide range of algorithms, based on pixel confidence aggregation or image-level failure detection, on multiple medical imaging datasets (section 3.2).

4.2.1 Interpretation of the Benchmark Results

First, the results obtained in this benchmarking study are interpreted in relation to the three research questions (RQs) posed in the introduction (section 1.4, page 29).

RQ 2.1 asked about best practices for the evaluation of segmentation failure detection. To approach this question systematically, requirements for the evaluation protocol were derived in section 2.2.2.1, based on theoretical considerations and pitfalls observed in related works. These requirements were adapted from Jaeger et al. (2022), to align them with the practical needs of failure detection for medical image segmentation. The proposed evaluation protocol based on the area under the risk-coverage curve (AURC) was identified as the best fit for benchmarks that evaluate a wide spectrum of failure detection methods. Quantitative analyses from section 3.2.3 confirmed that the AURC fulfills requirement R2, while other metrics like Spearman correlation coefficient (SC) violate it in certain cases, which can lead to method recommendations that do not meet practical needs. As different metrics capture different aspects of performance, it is usually a good idea to evaluate methods with multiple metrics. Evaluating Pearson correlation coefficient (PC) and SC in addition to AURC did not yield new insights, however, in this benchmark. Using out-

of-distribution (OOD) detection as an alternative task formulation, in contrast, changed the method rankings significantly, which showed that failure detection is a different task that requires a separate evaluation methodology (requirement R1). Requirement R3 is motivated by the fact that the risk function should be adapted to the specific medical application in which failure detection is used. Evaluation protocols that can incorporate different risk functions are beneficial for benchmarking studies such as this part of the thesis, as they allow investigating how a change in the risk function impacts the results. The analysis in section 3.2.3 showed only small ranking differences when varying risk functions, but in general such analyses could provide insights into which methods perform robustly. The last requirement (R4) demands using realistic failure sources for failure detection evaluation. This thesis implemented failure sources as distribution shifts in the test sets, which resulted in a clear segmentation performance drop, as expected (section 3.2.1.1). While most of the shifts originated from multi-centric data collection, it is difficult to quantify how realistic the setup was. As overall similar failure detection results were obtained for the Brain tumor 2D dataset with artificial shifts, it may also be worth studying realistic, synthetic data as a preliminary testbed for failure detection.

RQ 2.2 is central for all benchmarks, ultimately asking: Which method performs best? Within the scope of this project, the answer is clear. The pairwise DSC method robustly ranked first across all evaluated datasets and metrics when combined with an ensemble of segmentation models. While the ensemble helps to avoid segmentation errors in the first place, the pairwise DSC quantifies the agreement within the ensemble, which proved to be beneficial for detecting failures. Simplicity is a major strength of this method because it can be easily transferred to alternative segmentation model architectures (section 3.2.2.2) or potentially even to different risk functions, by replacing DSC with alternative segmentation metrics. Segmentation quality regression models (based on a deep neural network or on a regression forest) also showed potential on some datasets, but they revealed weaknesses in scenarios with small training sets. Hence, incorporating multiple datasets in the benchmark with different characteristics is crucial to finding generalizable failure detection methods and identifying potential failure modes, thereby improving their practical applicability.

RQ 2.3 focused on the subset of methods that aggregate a pixel confidence map, which have been studied less extensively than image-level failure detection methods. The naive baseline of averaging the confidence scores across all pixels was among the worst-performing methods, as expected from their bias towards object size (Jungo et al. 2020; Kahl et al. 2024). Although methods that proposed remedies to this bias (patch-based and non-boundary aggregation) obtained slightly better failure detection results in this benchmark on average, their performance was unstable across datasets and could not compete with the best methods. Learning to estimate the risk function from patterns in the confidence map using regression forest (RF) turned out to improve AURC scores compared

to the simpler methods above. While the original method by Jungo et al. (2020) used radiomics to extract features from the confidence map, a variant proposed in this thesis relied on simpler features extracted from the confidence map and the predicted segmentation. This simple method achieved consistently higher performance in the benchmark, which suggests that the confidence maps alone from current methods may be suboptimal for failure detection, and that features of the predicted segmentation mask can provide additional useful information. Similar to the overall benchmark results, evaluating confidence aggregation methods on multiple datasets was important, as it exposed performance drops of RF-based methods on datasets with few training samples. Furthermore, none of the aggregation strategies were competitive with image-level failure detection methods across datasets, so developing a robust combination of pixel confidence and aggregation methods remains an open problem.

4.2.2 Comparison to Related Work

Although this benchmarking study is the first to comprehensively compare a diverse set of methods across several datasets, previous results for individual methods and datasets exist, which are compared to the results from section 3.2 in this section.

Regarding the evaluation protocol, a few alternative metrics to AURC have been used previously. Malinin et al. (2022) proposed "error-retention" curves, which essentially only differ from risk-coverage curves in the detail that the selective risk is computed by averaging the actual risk values for samples below the confidence threshold and replace the risk of samples above the threshold with the risk of an oracle prediction (usually zero). In the low-coverage region (for example, 5 %), the selective risk in error-retention curves is typically low due to the majority of oracle predictions (for example, 95 %) being averaged, which leads to high-confidence-high-risk predictions having less impact than for AURC. Since the risk-coverage curve more intuitively handles high-confidence samples and is an established, well-studied evaluation method for selective classification as well as failure detection (El-Yaniv and Wiener 2010; Jaeger et al. 2022), this benchmark opted for AURC as the main metric. Recently, however, Traub et al. (2024) brought forth theoretical arguments that AURC over-emphasizes the importance of high-confidence samples, and proposed a generalized AURC, which handles these samples identically to the error-retention curve. An analysis equivalent to risk-coverage curves was employed by Ng et al. (2023), but they chose a binary risk function defined via manual thresholds on a segmentation metric. For specific applications with a well-defined failure threshold, this is reasonable, but the benchmark in this thesis prefers a continuous risk function as argued in requirement R3 from section 2.2.2.1. Finally, Galil et al. (2023) recommended using the maximum coverage at a predefined, target risk level (or vice versa, minimum risk at a predefined,

target coverage level) as an alternative to AURC. For example, in a specific application, the goal could be to accept as many automatic segmentations as possible (maximize coverage) while limiting the expected risk among the accepted samples to an equivalent of mean DSC ≥ 0.9. Methods with larger coverage for this risk constraint would then be preferable for the application. This benchmark used publicly available datasets that do not come with a natural target risk, so AURC appears advantageous in that it avoids noise through arbitrary constraints.

Pairwise DSC was originally proposed by Roy et al. (2019), who used it in conjunction with MC-Dropout. The combination with ensembles of neural networks has been explored by a few other works. Hoebel et al. (2020) and Hoebel et al. (2022) reported the Pearson correlation coefficient (PC) between pairwise and true DSC on a similar brain tumor dataset, based on different prediction models. The results from Hoebel et al. (2022) agree with this thesis in the finding that pairwise DSC tends to overestimate true DSC scores. Differences were observed in the comparison of prediction models: Ensembles achieved worse scores in their experiments than MC-Dropout, whereas ensembles were slightly superior in this thesis. Although the numbers are not directly comparable due to differences in the experimental setup, such as different test sets, segmentation models and MC-Dropout hyperparameters, this discrepancy is worth examining further. Unfortunately, the referenced paper does not provide public source code, in contrast to this benchmark, making a detailed comparison currently difficult. Results by Ng et al. (2023) on a cardiac MRI dataset indicate that a slight modification of the pairwise DSC in combination with an ensemble performs best, in agreement with this benchmark. While these results from the related work were obtained on individual datasets, this benchmark evaluates failure detection methods across six different CT and MRI datasets, as well as initial results on three other modalities.

Aggregating pixel confidence maps for failure detection has been studied by Jungo et al. (2020) and Kahl et al. (2024). This benchmark confirmed the results of experiments performed by Jungo et al. (2020) on a brain tumor dataset, which showed that advanced aggregation methods clearly outperformed the mean confidence baseline. However, their best-performing method, the RF trained on radiomics features, was inferior in this benchmark's experiments to a simplified method also based on RFs proposed in this thesis. The failure modes of both methods for small training sets are a novel finding from this benchmark, too. Kahl et al. (2024) examined uncertainty for segmentation in a wider sense, instead of focusing on failure detection in medical imaging datasets, so their results are hard to compare to this benchmark. One shared finding is that the performance ranking of simple aggregation methods, such as mean confidence over the whole image, non-boundary region, or sliding window patches, is unstable across datasets.

Quality regression networks are a popular method and have been studied for individual

medical applications, such as cardiac segmentation (Robinson et al. 2018; Li et al. 2022) and brain tumors (Qiu et al. 2023). The different dataset setups prevent a direct comparison to this benchmark's results, in particular the common practice of not specifying a concrete segmentation model whose failures should be detected but rather building a test set containing predictions from different models of varying (often balanced) quality. This style of test set was not used in this thesis for the reasons described in requirement R4 of section 2.2.2.1. However, balancing the *training set* with such techniques could potentially improve the failure detection performance of quality regression networks. It was not implemented in this thesis because no reference implementation was published by the aforementioned papers. An additional finding from this benchmark that was not reported in the related work so far is the potential weakness of quality regression networks with small datasets and/or large distribution shifts, which are possible explanations for the performance drop on the Prostate and Covid datasets in this benchmark.

In this benchmark, results with OOD detection metrics turned out vastly different from those obtained with failure detection metrics, which confirms observations from recent work (Lennartz and Schultz 2023) and generalizes them to a larger set of datasets. The Mahalanobis method proposed by González et al. (2022) showcases this discrepancy in particular, as it excelled in OOD detection but demonstrated limited utility for failure detection in this benchmark. Similar results on the relation of OOD and failure detection were also found recently for medical image classification (Anthony and Kamnitsas 2025). Their analysis also highlighted that failure detection may not be the only goal in practice, because OOD artifacts can in some cases result in correct predictions, although the model used a shortcut to arrive at it. As this is usually not desired, a combination of OOD and failure detection methods can be beneficial.

4.2.3 Limitations and Future Work

The method selection for this benchmark aimed to include a wide variety of approaches to failure detection. Popularity in the related work (as reviewed by Lambert et al. (2024)) was another criterion for selecting baselines. As segmentation uncertainty, quality estimation, and OOD detection are active research fields with numerous publications, it was not possible to include all relevant methods. Integrating more algorithms was also hindered by the lack of public source code and insufficient reporting of experimental setups in the corresponding papers, which made reproducing results from the related work difficult. For example, from recently published works mentioned in the related work section 1.3.2, only two came with a complete codebase (Lennartz and Schultz 2023; Kahl et al. 2024), whereas three published incomplete code (Jungo et al. 2020; González et al. 2022; Qiu et al. 2023), which missed a subset of methods, important steps in the dataset preparation or

data loading pipeline, respectively. The majority of works did not release any reference implementation (Robinson et al. 2018; Liu et al. 2019; Mehrtash et al. 2020; Wang et al. 2020c; Li et al. 2022; Ng et al. 2023). This benchmark study publishes the complete source code for all methods and experiments* so that the research community can expand the large but limited failure detection method collection.

Similar to the method selection, more datasets could be added to the benchmark. While failure detection is also important for in-distribution data, the most challenging scenarios are datasets with realistic distribution shifts between training and test set, so collecting such datasets is a promising future direction. Orthogonal to a larger dataset collection, detailed, dataset-specific analyses are a future perspective. Failure source analysis was limited in this study to qualitative results (fig. 3.14), whereas in-depth, quantitative analyses are necessary to answer the question of why models make errors. These might help to find limitations and potential improvements of current segmentation or failure detection methods. A final limitation of the datasets used in this benchmark is the possibility of annotation shifts between training and test sets. As in all comparisons of segmentation models, such shifts should be avoided, because they can lead to noisy segmentation metric results, which will also affect failure detection scores. Systematic annotation shifts are not expected for the brain tumor, heart, and kidney tumor datasets, which were annotated with standardized protocols and used in international competitions before. In contrast, the Covid and prostate datasets are composed of several independently annotated datasets from different sources, making annotation inconsistencies more likely. The ranking stability analysis (fig. 3.17) suggests that potential noise from annotation shifts did not have an impact on the main insights from this benchmark, but an investigation into annotation consistency for the aforementioned datasets could be valuable.

Failure detection as defined in this thesis is a practically important task, but not the only purpose of segmentation uncertainty estimation. Other downstream tasks include OOD detection, active learning, probability calibration, and ambiguity modeling (Kahl et al. 2024). Even when focusing on failure detection, there are different levels on which it can be applied: While this study evaluated methods that provide one confidence score per case, in some applications one confidence score per class (in a multi-class segmentation task) or one per pixel would be ideal. Fortunately, the evaluation protocol used here can be adapted easily to these alternative confidence granularity levels. For class-level failure detection, the risk function needs to be defined for each class separately, which is often easy, as segmentation metrics are usually defined per class. Not all methods from this benchmark can be trivially adapted to this setting, though. Pixel-level failure detection boils down to the classification failure detection task studied by Jaeger et al. (2022), which can also be evaluated using AURC with a binary risk function. In this task formulation,

*https://github.com/MIC-DKFZ/segmentation_failures_benchmark

none of the image-level methods from section 2.2.5.3 are applicable anymore.

Ultimately, the goal of failure detection methods is to support the integration of automatic segmentation methods in real clinical workflows by reducing their error rate and increasing trust in the predictions. While this study focused on failure detection performance measured through metrics like AURC, other considerations may be equally important for practical implementations. After selecting the best failure detection method for a specific application, it is necessary to determine the optimal operating point for the detector, which was not investigated in this benchmark. How to choose a confidence threshold for rejecting predictions has been studied under the assumption of an in-distribution test set (Geifman and El-Yaniv 2017), but controlling the expected risk in scenarios with distribution shifts, which are common in real-world deployments, is an open problem. Another practically relevant aspect is the efficiency of failure detection methods, which is crucial for time-critical or resource-limited scenarios sometimes encountered during deployment. For extremely resource-limited setups, confidence aggregation methods based on a single network prediction may be the only feasible option, at the cost of significantly lower failure detection performance compared to the best methods from the benchmark. Pairwise DSC appears as the best trade-off, as it achieves the highest failure detection scores and adds only minor computation workload compared to the prediction model. Although it cannot be applied to a single network, running an ensemble of five networks is still feasible for many applications. Future research could try to reduce the inference costs of an ensemble while generating similar predictive distributions.

4.3 Overall Conclusions

This thesis investigated the robustness of deep learning algorithms for medical image segmentation in two benchmarking studies on generalization and failure detection, respectively. The first study evaluated how state-of-the-art brain tumor segmentation algorithms generalize on images from more than 2500 test patients, which were distributed among 32 institutions in an international federation, by sending the models to the data owners instead of collecting the data centrally. The FeTS Challenge was the first competition to use such a large-scale federated evaluation, highlighting its potential as a tool for comprehensive generalization benchmarking in medical image analysis but also the significant practical hurdles in conducting federated experiments. While the average-case segmentation accuracy was high, the competition results also revealed a lack of robustness, in the sense that the worst-case performance at many institutions was still not satisfactory.

This finding established a connection to the second benchmark performed in this thesis, which focused on failure detection methods. Even if a segmentation model makes mistakes, these methods can help to identify the incorrect predictions, allowing them to

be filtered out or corrected by a human expert before the erroneous segmentation affects downstream analyses. The evaluation protocol developed for this study allows a unified comparison of different approaches to failure detection that were previously studied in isolation. Together with the open-source benchmark implementation, which includes a wide variety of method baselines, multiple datasets, and realistic distribution shifts in the test set, it lays the foundation for future method development. In the benchmark results, the pairwise DSC method stood out in particular, as it consistently reduced the remaining error after filtering, across different computed tomography (CT) and MRI datasets. Being simple and easy to adapt to new segmentation problems, it constitutes a strong baseline that can be used in future research and practical applications of failure detection, such as quality control of automated segmentations.

Overall, this thesis introduced innovative benchmarks for assessing the current state of the art in generalization and failure detection methods for medical image segmentation, which are complementary approaches for increasing reliability and trust in automatic segmentations. The thesis also provides a foundation for exploring synergies between the two tasks in the future. Federated data is not only useful for generalization benchmarking competitions but also for investigating how failure detection methods improve robustness in real-world clinical scenarios. Failure detection methods, in turn, can help with annotation quality control in the preparation of a federated evaluation workflow. They also help to identify failure modes of segmentation models, fostering the development of generalization methods that avoid such errors.

5 Summary

Radiology is at the forefront of adopting artificial intelligence (AI) solutions in clinical practice because the steadily increasing need for examinations based on medical imaging exceeds the growth in the human workforce. Semantic segmentation is an important component of image analysis pipelines, including applications in computer-aided diagnosis, radiation therapy planning, and disease monitoring. Nowadays, deep learning (DL) algorithms can perform automatic segmentation of various anatomical structures based on appropriately annotated training datasets. However, these algorithms do not work perfectly and can especially make mistakes when applied to data that has different characteristics than the data the models were trained on. The discrepancy between training and testing data characteristics is called distribution shift and frequently occurs when deploying models in new hospitals. In this thesis, benchmarks were developed for methods that improve the robustness of segmentation methods to such distribution shifts. Two complementary approaches were studied here: Methods that improve the out-of-distribution *generalization* directly or methods that know when they are wrong (*failure detection*).

Generalization methods were benchmarked in this thesis by organizing an international competition, also known as a challenge. Such challenges are the gold standard in medical image analysis for comparing state-of-the-art algorithms, due to their standardized, fair conditions for all participants. While many competitions are organized each year, they usually use research datasets that originate from a small set of institutions and scanners. Therefore, it is unknown how well algorithms generalize to more diverse multicentric data with distribution shifts that arise in the real world. This thesis introduces the idea of using federated data in the competition setting, which lowers the hurdles for contributing data significantly, as the data does not leave the institution where it was acquired. To perform a federated evaluation, the segmentation algorithms are sent to the institutions in the federation, and results on their performance are communicated back for analyzing the robustness. The concept of federated evaluation benchmarks was implemented here in a competition for the task of brain tumor segmentation, the Federated Tumor Segmentation (FeTS) Challenge. As the first federated challenge conducted so far, the FeTS Challenge revealed and partially addressed practical hurdles associated with federated evaluation,

notably the high organizational effort, the increased difficulty of annotation quality control compared to conventional challenges, and the constraints on the challenge analysis due to the lack of direct access to federated data. However, it also highlighted the potential of federated benchmarks to boost the dataset size and diversity considerably, exemplified by the testing dataset of the FeTS Challenge, to which 32 international institutions contributed 2625 cases with multi-parametric magnetic resonance imaging (MRI) scans. Evaluating the 41 segmentation models submitted to the competition on the test data showed that they obtained good average-case generalization, but also a lack of worst-case robustness on 13 of the 32 institutions.

Failure detection is important for the reliability of segmentation methods in practice, so it has been studied from many perspectives, including uncertainty estimation, out-of-distribution detection, and segmentation quality estimation. Progress in method development is currently hindered for two reasons: The evaluation protocols used by the above approaches differ, making cross-comparison of methods towards the same goal of failure detection difficult. Furthermore, novel methods have often been evaluated only in a single segmentation task (anatomical region) or not considering distribution shifts, which leaves questions about their generalizability unanswered. Therefore, the second part of the thesis addresses these shortcomings, by developing an evaluation protocol based on a risk-coverage analysis, which allows comparing all relevant methods in failure detection while avoiding pitfalls in current practice. A benchmark was designed that implemented the proposed evaluation and compared several, diverse failure detection methods in experiments with multiple public datasets that contain realistic distribution shifts. The benchmark results provided insights into how uncertainties on the pixel level can be effectively aggregated into image-level uncertainties for failure detection. Moreover, an existing, simple method was identified as a strong baseline for future research, as it consistently outperformed more complicated algorithms across datasets. Due to its flexibility and efficiency, it can be easily adapted to new segmentation tasks and practical applications.

In conclusion, large-scale benchmarking studies were conducted in this thesis, which test state-of-the-art generalization and failure detection algorithms in scenarios that simulate performance in real-world deployments. The experiments demonstrated how to employ multicentric data in centralized and federated form for evaluating robustness to distribution shifts, revealing common failure sources, and identifying practical algorithms that are able to generalize to new hospitals and abstain from uncertain predictions. The code for both benchmarks is made available to the community to foster meaningful method comparison and progress in robust medical image segmentation algorithms.

6 Zusammenfassung

Bei der Einführung von auf künstlicher Intelligenz basierten Lösungen in der klinischen Praxis hat die Radiologie eine Vorreiterrolle, da der stetig wachsende Bedarf an bildbasierten Untersuchungen nicht von den verfügbaren Radiologen gedeckt werden kann. Die semantische Segmentierung ist eine zentrale Komponente von Bildanalyse-Pipelines und findet unter anderem Anwendung in der computergestützten Diagnose, der Planung von Strahlentherapien und der Überwachung von Krankheitsverläufen. Heutzutage können Deep Learning-Algorithmen verschiedene anatomische Strukturen automatisch segmentieren, mithilfe von entsprechend annotierten Trainingsdatensätzen. Diese Algorithmen können jedoch auch Fehler machen, insbesondere wenn sie auf Daten angewendet werden, die sich in ihren Eigenschaften von den Trainingsdaten unterscheiden. Die Diskrepanz zwischen den Eigenschaften der Trainings- und Testdaten wird als Distribution Shift bezeichnet und tritt häufig auf, wenn Modelle in neuen Krankenhäusern eingesetzt werden. Für diese Doktorarbeit wurden Benchmarks für Methoden entwickelt, die die Robustheit von Segmentierungsverfahren gegenüber solchen Distribution Shifts verbessern. Dabei wurden zwei komplementäre Ansätze untersucht: Methoden, die die Generalisierung auf Daten mit Distribution Shifts verbessern, sowie Methoden, die erkennen können, wann sie falsche Vorhersagen treffen (Fehlererkennung).

Die Benchmarking-Studie zu Generalisierung erfolgte in dieser Arbeit durch die Organisation eines internationalen Wettbewerbs (auch Challenge genannt). Solche Challenges gelten als Goldstandard in der medizinischen Bildanalyse für den Vergleich von Algorithmen, da sie standardisierte und faire Bedingungen für alle Teilnehmenden bieten. Obwohl jedes Jahr zahlreiche Wettbewerbe organisiert werden, basieren sie in der Regel auf Forschungsdatensätzen, die von wenigen Institutionen und Scannern stammen. Daher ist oft unklar, wie gut die Algorithmen auf multizentrische Daten mit größerer Diversität und realistischen Distribution Shifts generalisieren. Diese Arbeit führt das Konzept ein, föderierte Daten in Challenge-Settings zu nutzen. Solche Daten verlassen die Institution, in der sie erhoben wurden, nicht, was die Hürden für die Bereitstellung von Daten erheblich senkt. Für eine föderierte Evaluation werden die Segmentierungsalgorithmen an die Institutionen im Verbund geschickt, und deren Evaluierungsergebnisse werden zurückgemeldet, um

die Robustheit der Modelle zu analysieren. Dieses Konzept wird in einer Challenge zur Segmentierung von Hirntumoren umgesetzt—der Federated Tumor Segmentation (FeTS) Challenge. Als erste ihrer Art offenbart und adressiert die FeTS Challenge einige praktische Herausforderungen der föderierten Evaluation, insbesondere den hohen organisatorischen Aufwand, die erschwerte Qualitätskontrolle von Annotationen im Vergleich zu konventionellen Challenges und die eingeschränkte Analysemöglichkeit aufgrund des fehlenden direkten Zugriffs auf die föderierten Daten. Gleichzeitig zeigt die Challenge aber auch das Potenzial föderierter Benchmarks, die Größe und Vielfalt der Testdatensätze erheblich zu steigern. Dies wird durch die FeTS Challenge exemplarisch demonstriert, bei der 32 internationale Institutionen insgesamt 2625 Fälle mit multiparametrischen Magnetresonanztomographie (MRT)-Scans beisteuerten. Die Evaluierung der 41 in der Challenge eingereichten Segmentierungsmodelle auf diesen Testdaten zeigte, dass die Modelle im Durchschnitt gut generalisierten, aber auf Daten von 13 der 32 beteiligten Institutionen in Einzelfällen Fehler machten, die auf einen Mangel an Robustheit hinweisen.

Die Fehlererkennung ist für die Zuverlässigkeit von Segmentierungsmethoden in der Praxis von großer Bedeutung und wurde aus vielen Perspektiven untersucht, darunter Unsicherheitsabschätzung, Out-of-Distribution-Erkennung und Schätzung der Segmentierungsqualität. Der Fortschritt in diesem Forschungsbereich wird derzeit durch zwei Probleme behindert: Erstens unterscheiden sich die Evaluationsprotokolle der verschiedenen Ansätze, was einen direkten Vergleich der Methoden zur Fehlererkennung erschwert. Zweitens wurden neue Methoden bisher oft nur für ein Segmentierungsproblem (z. B. in einer anatomischen Region) getestet oder nicht hinsichtlich Distribution Shifts evaluiert, sodass ihre Anwendbarkeit auf ein breiteres Aufgabenspektrum unklar bleibt. Der zweite Teil dieser Arbeit adressiert diese Defizite durch die Entwicklung eines Evaluationsprotokolls basierend auf einer Risk-Coverage Analyse, welches den Vergleich aller relevanten Methoden der Fehlererkennung ermöglicht und Schwachstellen bisheriger Ansätze vermeidet. Ein Benchmark wurde entwickelt, der diese Evaluationsstrategie implementiert und verschiedene, diverse Methoden zur Fehlererkennung in Experimenten mit mehreren öffentlichen Datensätzen vergleicht, die realistische Distribution Shifts enthalten. Die Ergebnisse dieser Studie lieferten Erkenntnisse darüber, wie Unsicherheitswerte auf Pixel-Ebene effektiv zu einem Unsicherheitswert auf Bild-Ebene für die Fehlererkennung aggregiert werden können. Zudem wurde eine existierende, einfache Methode als starke Referenz für zukünftige Forschung identifiziert, da sie über mehrere Datensätze hinweg konsistent leistungsfähiger als komplexere Algorithmen war. Dank ihrer Flexibilität und Effizienz kann diese Methode leicht an neue Segmentierungsprobleme und praktische Anwendungen angepasst werden.

Zusammenfassend führte diese Dissertation groß angelegte Benchmarking-Studien durch, die modernste Generalisierungs- und Fehlererkennungsalgorithmen in realitätsna-

hen Szenarien testen. Die Experimente demonstrieren, wie multizentrische Daten sowohl in zentralisierter als auch in föderierter Form genutzt werden können, um die Robustheit gegenüber Distribution Shifts zu evaluieren. Dabei wurden häufige Fehlerquellen aufgedeckt und praxistaugliche Algorithmen identifiziert, die eine gute Generalisierung auf neue Krankenhäuser ermöglichen und außerdem signalisieren können, wenn Segmentierungen potenziell fehlerhaft sind. Der Code für beide Benchmarks wird der wissenschaftlichen Gemeinschaft zur Verfügung gestellt, um eine fundierte Vergleichbarkeit von Methoden zu ermöglichen und den Fortschritt in der robusten medizinischen Bildsegmentierung voranzutreiben.

Bibliography

Adams, J. and S. Y. Elhabian (Aug. 14, 2023). **Benchmarking Scalable Epistemic Uncertainty Quantification in Organ Segmentation**. DOI: 10.48550/arXiv.2308.07506. arXiv: 2308.07506 [cs, eess]. URL: http://arxiv.org/abs/2308.07506 (visited on 10/18/2023). Pre-published (cit. on p. 26).

Adewole, M., J. D. Rudie, A. Gbadamosi, O. Toyobo, C. Raymond, D. Zhang, O. Omidiji, R. Akinola, M. A. Suwaid, A. Emegoakor, N. Ojo, K. Aguh, C. Kalaiwo, G. Babatunde, A. Ogunleye, Y. Gbadamosi, K. Iorpagher, E. Calabrese, M. Aboian, M. Linguraru, J. Albrecht, B. Wiestler, F. Kofler, A. Janas, D. LaBella, A. F. Kzerooni, H. B. Li, J. E. Iglesias, K. Farahani, J. Eddy, T. Bergquist, V. Chung, R. T. Shinohara, W. Wiggins, Z. Reitman, C. Wang, X. Liu, Z. Jiang, A. Familiar, K. V. Leemput, C. Bukas, M. Piraud, G.-M. Conte, E. Johansson, Z. Meier, B. H. Menze, U. Baid, S. Bakas, F. Dako, A. Fatade, and U. C. Anazodo (2023). **The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-africa)**. arXiv: 2305.19369 [eess.IV] (cit. on p. 3).

Aerts, H. J. W. L., E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin (June 3, 2014). **Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach**. In: *Nature Communications* 5.1, p. 4006. ISSN: 2041-1723. DOI: 10.1038/ncomms5006. URL: https://www.nature.com/articles/ncomms5006 (visited on 12/20/2024) (cit. on pp. 1, 4).

Akbar, A. S., C. Fatichah, and N. Suciati (2021). **Unet3D with Multiple Atrous Convolutions Attention Block for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 182–193 (cit. on p. 190).

Alam, S., B. Halandur, P. P. Mana, D. Goplen, A. Lundervold, and A. S. Lundervold (2021). **Brain Tumor Segmentation from Multiparametric MRI Using a Multi-encoder U-Net Architecture**. In: *International MICCAI Brainlesion Workshop*, pp. 289–301 (cit. on p. 190).

AlBadawy, E. A., A. Saha, and M. A. Mazurowski (2018). **Deep Learning for Segmentation of Brain Tumors: Impact of Cross-Institutional Training and Testing**. In: *Medical*

Physics 45.3, pp. 1150–1158. ISSN: 2473-4209. DOI: 10.1002/mp.12752. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12752 (visited on 06/16/2020) (cit. on p. 2).

Alexander, R., S. Waite, M. A. Bruno, E. A. Krupinski, L. Berlin, S. Macknik, and S. Martinez-Conde (Aug. 2022). **Mandating Limits on Workload, Duty, and Speed in Radiology**. In: *Radiology* 304.2, pp. 274–282. ISSN: 0033-8419. DOI: 10.1148/radiol.212631. URL: https://pubs.rsna.org/doi/full/10.1148/radiol.212631 (visited on 05/31/2024) (cit. on p. 1).

An, P., S. Xu, S. A. Harmon, E. B. Turkbey, T. H. Sanford, A. Amalou, M. Kassin, N. Varble, M. Blain, V. Anderson, F. Patella, G. Carrafiello, B. T. Turkbey, and B. J. Wood (2020). **CT Images in COVID-19**. The Cancer Imaging Archive. DOI: 10.7937/TCIA.2020.GQRY-NC81 (cit. on p. 52).

Anthony, H. and K. Kamnitsas (2025). **Evaluating Reliability in Medical DNNs: A Critical Analysis of Feature and Confidence-Based OOD Detection**. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. Ed. by C. H. Sudre, R. Mehta, C. Ouyang, C. Qin, M. Rakic, and W. M. Wells. Cham: Springer Nature Switzerland, pp. 160–170. ISBN: 978-3-031-73158-7. DOI: 10.1007/978-3-031-73158-7_15 (cit. on p. 119).

Antonelli, M., A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A. L. Simpson, L. Maier-Hein, and M. J. Cardoso (July 15, 2022). **The Medical Segmentation Decathlon**. In: *Nature Communications* 13.1 (1), p. 4128. ISSN: 2041-1723. DOI: 10.1038/s41467-022-30695-9. URL: https://www.nature.com/articles/s41467-022-30695-9 (visited on 02/05/2024) (cit. on p. 52).

Aubreville, M., N. Stathonikos, C. A. Bertram, R. Klopfleisch, N. ter Hoeve, F. Ciompi, F. Wilm, C. Marzahl, T. A. Donovan, A. Maier, J. Breen, N. Ravikumar, Y. Chung, J. Park, R. Nateghi, F. Pourakpour, R. H. J. Fick, S. Ben Hadj, M. Jahanifar, A. Shephard, J. Dexl, T. Wittenberg, S. Kondo, M. W. Lafarge, V. H. Koelzer, J. Liang, Y. Wang, X. Long, J. Liu, S. Razavi, A. Khademi, S. Yang, X. Wang, R. Erber, A. Klang, K. Lipnik, P. Bolfa, M. J. Dark, G. Wasinger, M. Veta, and K. Breininger (Feb. 1, 2023). **Mitosis Domain Generalization in Histopathology Images — The MIDOG Challenge**. In: *Medical Image Analysis* 84, p. 102699. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102699.

URL: https://www.sciencedirect.com/science/article/pii/S1361841522003279 (visited on 06/26/2024) (cit. on pp. 2, 19, 20, 27, 113).

Badgeley, M. A., J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley (Apr. 30, 2019). **Deep Learning Predicts Hip Fracture Using Confounding Patient and Healthcare Variables**. In: *npj Digital Medicine* 2.1 (1), pp. 1–10. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0105-1. URL: https://www.nature.com/articles/s41746-019-0105-1 (visited on 07/14/2023) (cit. on p. 2).

Bai, W., H. Suzuki, J. Huang, C. Francis, S. Wang, G. Tarroni, F. Guitton, N. Aung, K. Fung, S. E. Petersen, S. K. Piechnik, S. Neubauer, E. Evangelou, A. Dehghan, D. P. O'Regan, M. R. Wilkins, Y. Guo, P. M. Matthews, and D. Rueckert (Oct. 2020). **A Population-Based Phenome-Wide Association Study of Cardiac and Aortic Structure and Function**. In: *Nature Medicine* 26.10, pp. 1654–1662. ISSN: 1546-170X. DOI: 10.1038/s41591-020-1009-y. URL: https://www.nature.com/articles/s41591-020-1009-y (visited on 02/13/2025) (cit. on p. 1).

Baid, U., S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, L. M. Prevedello, J. D. Rudie, C. Sako, R. T. Shinohara, T. Bergquist, R. Chai, J. Eddy, J. Elliott, W. Reade, T. Schaffter, T. Yu, J. Zheng, A. W. Moawad, L. O. Coelho, O. McDonnell, E. Miller, F. E. Moron, M. C. Oswood, R. Y. Shih, L. Siakallis, Y. Bronstein, J. R. Mason, A. F. Miller, G. Choudhary, A. Agarwal, C. H. Besada, J. J. Derakhshan, M. C. Diogo, D. D. Do-Dai, L. Farage, J. L. Go, M. Hadi, V. B. Hill, M. Iv, D. Joyner, C. Lincoln, E. Lotan, A. Miyakoshi, M. Sanchez-Montano, J. Nath, X. V. Nguyen, M. Nicolas-Jilwan, J. O. Jimenez, K. Ozturk, B. D. Petrovic, C. Shah, L. M. Shah, M. Sharma, O. Simsek, A. K. Singh, S. Soman, V. Statsevych, B. D. Weinberg, R. J. Young, I. Ikuta, A. K. Agarwal, S. C. Cambron, R. Silbergleit, A. Dusoi, A. A. Postma, L. Letourneau-Guillon, G. J. G. Perez-Carrillo, A. Saha, N. Soni, G. Zaharchuk, V. M. Zohrabian, Y. Chen, M. M. Cekic, A. Rahman, J. E. Small, V. Sethi, C. Davatzikos, J. Mongan, C. Hess, S. Cha, J. Villanueva-Meyer, J. B. Freymann, J. S. Kirby, B. Wiestler, P. Crivellaro, R. R. Colen, A. Kotrotsou, D. Marcus, M. Milchenko, A. Nazeri, H. Fathallah-Shaykh, R. Wiest, A. Jakab, M.-A. Weber, A. Mahajan, B. Menze, A. E. Flanders, and S. Bakas (Sept. 12, 2021). **The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification**. DOI: 10.48550/arXiv.2107.02314. arXiv: 2107.02314 [cs]. URL: http://arxiv.org/abs/2107.02314 (visited on 06/28/2023). Pre-published (cit. on pp. 3, 19, 32, 33, 35, 37, 39, 112, 189).

Bakas, S., H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos (Sept. 5, 2017). **Advancing The Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features**. In: *Scientific Data* 4.1 (1), p. 170117. ISSN: 2052-4463. DOI: 10.1038/sdata.2017

.117. URL: https://www.nature.com/articles/sdata2017117 (visited on 12/08/2020) (cit. on pp. 32, 51).

Bakas, S., U. Baid, J. Rudie, E. Calabrese, M. Aboian, U. Anazodo, G. M. Conte, J. Albrecht, H. B. Li, F. Kofler, M. Correia De Verdier, R. Huang, D. LaBella, R. Saluja, L. Gagnon, M. Aboian, A. Abayazeed, K. Farahani, V. Chung, Z. Reitman, J. Kirkpatrick, C. Wang, J. Villanueva-Meyer, A. Flanders, M. Aboian, A. Nada, M. Aboian, A. Abayazeed, P. Lohman, A. Moawad, A. Janas, K. Krantchev, F. Memon, Y. Velichko, E. Schrickel, K. Link, S. Aneja, R. Maresca, A. Nada, P. Vollmuth, V. M. Pérez, M. W. Pease, D. Godfrey, S. Floyd, M. Adewole, F. Dako, O. Toyobo, O. Omidiji, Y. Gbadamosi, A. Ogunleye, N. Ojo, K. Iorpagher, G. Babatunde, K. Aguh, A. Emegoakor, C. Kalaiwo, M. G. Linguraru, A. F. Kazerooni, Z. Jiang, X. Liu, D. Gandhi, N. Khalili, A. Vossough, A. Nabavizadeh, J. B. Ware, B. Menze, E. Johanson, Z. Meier, W. Chen, N. Petrick, B. Sahiner, R. Chai, B. Wiestler, J. E. Iglesias, S. M. Anwar, K. Van Leemput, and M. Piraud (Apr. 2024). **BraTS 2024 Cluster of Challenges (BraTS + Beyond-BraTS)**. DOI: 10.5281/zenodo.10978907. URL: https://doi.org/10.5281/zenodo.10978907 (cit. on p. 112).

Bakas, S., S. Pati, M. Sheller, A. Karargyris, P. Mattson, B. Edwards, U. Baid, Y. Chen, R. (Shinohara, J. Martin, B. Menze, M. Zenk, K. Maier-Hein, R. Floca, A. Reinke, L. Maier-Hein, F. Isensee, D. Zimmerer, and Y. Chen (Mar. 16, 2022). **The Federated Tumor Segmentation (FeTS) Challenge 2022**. In: DOI: 10.5281/zenodo.6622476. URL: https://zenodo.org/records/6622476 (visited on 06/05/2024) (cit. on p. 32).

Bakas, S., M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, M. Prastawa, E. Alberts, J. Lipkova, J. Freymann, J. Kirby, M. Bilello, H. Fathallah-Shaykh, R. Wiest, J. Kirschke, B. Wiestler, R. Colen, A. Kotrotsou, P. Lamontagne, D. Marcus, M. Milchenko, A. Nazeri, M.-A. Weber, A. Mahajan, U. Baid, E. Gerstner, D. Kwon, G. Acharya, M. Agarwal, M. Alam, A. Albiol, A. Albiol, F. J. Albiol, V. Alex, N. Allinson, P. H. A. Amorim, A. Amrutkar, G. Anand, S. Andermatt, T. Arbel, P. Arbelaez, A. Avery, M. Azmat, P. B., W. Bai, S. Banerjee, B. Barth, T. Batchelder, K. Batmanghelich, E. Battistella, A. Beers, M. Belyaev, M. Bendszus, E. Benson, J. Bernal, H. N. Bharath, G. Biros, S. Bisdas, J. Brown, M. Cabezas, S. Cao, J. M. Cardoso, E. N. Carver, A. Casamitjana, L. S. Castillo, M. Catà, P. Cattin, A. Cerigues, V. S. Chagas, S. Chandra, Y.-J. Chang, S. Chang, K. Chang, J. Chazalon, S. Chen, W. Chen, J. W. Chen, Z. Chen, K. Cheng, A. R. Choudhury, R. Chylla, A. Clérigues, S. Colleman, R. G. R. Colmeiro, M. Combalia, A. Costa, X. Cui, Z. Dai, L. Dai, L. A. Daza, E. Deutsch, C. Ding, C. Dong, S. Dong, W. Dudzik, Z. Eaton-Rosen, G. Egan, G. Escudero, T. Estienne, R. Everson, J. Fabrizio, Y. Fan, L. Fang, X. Feng, E. Ferrante, L. Fidon, M. Fischer, A. P. French, N. Fridman, H. Fu, D. Fuentes, Y. Gao, E. Gates, D. Gering, A. Gholami, W. Gierke, B. Glocker, M. Gong, S. González-Villá, T. Grosges, Y. Guan, S. Guo, S.

Gupta, W.-S. Han, I. S. Han, K. Harmuth, H. He, A. Hernández-Sabaté, E. Herrmann, N. Himthani, W. Hsu, C. Hsu, X. Hu, X. Hu, Y. Hu, Y. Hu, R. Hua, T.-Y. Huang, W. Huang, S. Van Huffel, Q. Huo, V. HV, K. M. Iftekharuddin, F. Isensee, M. Islam, A. S. Jackson, S. R. Jambawalikar, A. Jesson, W. Jian, P. Jin, V. J. M. Jose, A. Jungo, B. Kainz, K. Kamnitsas, P.-Y. Kao, A. Karnawat, T. Kellermeier, A. Kermi, K. Keutzer, M. T. Khadir, M. Khened, P. Kickingereder, G. Kim, N. King, H. Knapp, U. Knecht, L. Kohli, D. Kong, X. Kong, S. Koppers, A. Kori, G. Krishnamurthi, E. Krivov, P. Kumar, K. Kushibar, D. Lachinov, T. Lambrou, J. Lee, C. Lee, Y. Lee, M. Lee, S. Lefkovits, L. Lefkovits, J. Levitt, T. Li, H. Li, W. Li, H. Li, X. Li, Y. Li, H. Li, Z. Li, X. Li, Z. Li, X. Li, W. Li, Z.-S. Lin, F. Lin, P. Lio, C. Liu, B. Liu, X. Liu, M. Liu, J. Liu, L. Liu, X. Llado, M. M. Lopez, P. R. Lorenzo, Z. Lu, L. Luo, Z. Luo, J. Ma, K. Ma, T. Mackie, A. Madabushi, I. Mahmoudi, K. H. Maier-Hein, P. Maji, C. P. Mammen, A. Mang, B. S. Manjunath, M. Marcinkiewicz, S. McDonagh, S. McKenna, R. McKinley, M. Mehl, S. Mehta, R. Mehta, R. Meier, C. Meinel, D. Merhof, C. Meyer, R. Miller, S. Mitra, A. Moiyadi, D. Molina-Garcia, M. A. B. Monteiro, G. Mrukwa, A. Myronenko, J. Nalepa, T. Ngo, D. Nie, H. Ning, C. Niu, N. K. Nuechterlein, E. Oermann, A. Oliveira, D. D. C. Oliveira, A. Oliver, A. F. I. Osman, Y.-N. Ou, S. Ourselin, N. Paragios, M. S. Park, B. Paschke, J. G. Pauloski, K. Pawar, N. Pawlowski, L. Pei, S. Peng, S. M. Pereira, J. Perez-Beteta, V. M. Perez-Garcia, S. Pezold, B. Pham, A. Phophalia, G. Piella, G. N. Pillai, M. Piraud, M. Pisov, A. Popli, M. P. Pound, R. Pourreza, P. Prasanna, V. Prkovska, T. P. Pridmore, S. Puch, E. Puybareau, B. Qian, X. Qiao, M. Rajchl, S. Rane, M. Rebsamen, H. Ren, X. Ren, K. Revanuru, M. Rezaei, O. Rippel, L. C. Rivera, C. Robert, B. Rosen, D. Rueckert, M. Safwan, M. Salem, J. Salvi, I. Sanchez, I. Sánchez, H. M. Santos, E. Sartor, D. Schellingerhout, K. Scheufele, M. R. Scott, A. A. Scussel, S. Sedlar, J. P. Serrano-Rubio, N. J. Shah, N. Shah, M. Shaikh, B. U. Shankar, Z. Shboul, H. Shen, D. Shen, L. Shen, H. Shen, V. Shenoy, F. Shi, H. E. Shin, H. Shu, D. Sima, M. Sinclair, O. Smedby, J. M. Snyder, M. Soltaninejad, G. Song, M. Soni, J. Stawiaski, S. Subramanian, L. Sun, R. Sun, J. Sun, K. Sun, Y. Sun, G. Sun, S. Sun, Y. R. Suter, L. Szilagyi, S. Talbar, D. Tao, D. Tao, Z. Teng, S. Thakur, M. H. Thakur, S. Tharakan, P. Tiwari, G. Tochon, T. Tran, Y. M. Tsai, K.-L. Tseng, T. A. Tuan, V. Turlapov, N. Tustison, M. Vakalopoulou, S. Valverde, R. Vanguri, E. Vasiliev, J. Ventura, L. Vera, T. Vercauteren, C. A. Verrastro, L. Vidyaratne, V. Vilaplana, A. Vivekanandan, G. Wang, Q. Wang, C. J. Wang, W. Wang, D. Wang, R. Wang, Y. Wang, C. Wang, G. Wang, N. Wen, X. Wen, L. Weninger, W. Wick, S. Wu, Q. Wu, Y. Wu, Y. Xia, Y. Xu, X. Xu, P. Xu, T.-L. Yang, X. Yang, H.-Y. Yang, J. Yang, H. Yang, G. Yang, H. Yao, X. Ye, C. Yin, B. Young-Moxon, J. Yu, X. Yue, S. Zhang, A. Zhang, K. Zhang, X. Zhang, L. Zhang, X. Zhang, Y. Zhang, L. Zhang, J. Zhang, X. Zhang, T. Zhang, S. Zhao, Y. Zhao, X. Zhao, L. Zhao, Y. Zheng, L. Zhong, C. Zhou, X. Zhou, F. Zhou, H. Zhu, J. Zhu, Y. Zhuge, W. Zong, J. Kalpathy-Cramer, K. Farahani, C. Davatzikos, K. van Leemput, and B. Menze (Apr. 23, 2019).

Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. arXiv: 1811.02629 [cs, stat]. URL: http://arxiv.org/abs/1811.02629 (visited on 12/09/2020) (cit. on pp. 19, 32, 37–39, 51, 78, 112).

Bakas, S., M. Sheller, S. Pati, B. Edwards, G. A. Reina, U. Baid, Y. Chen, R. (Shinohara, J. Martin, B. Menze, S. Albarqouni, M. Bilello, S. Mohan, J. B. Freymann, J. S. Kirb, C. Davatzikos, H. Fathallah-Shaykh, R. Wiest, A. Jakab, R. R. Colen, A. Kotrotsou, D. Marcus, M. Milchenko, A. Nazeri, M.-A. Weber, A. Mahajan, U. Baid, P. Vollmuth, M. Zenk, K. Maier-Hein, D. Zimmerer, A. Reinke, L. Maier-Hein, and J. Kleesiek (Mar. 2, 2021). **Federated Tumor Segmentation**. In: DOI: 10.5281/zenodo.4573128. URL: https://zenodo.org/records/4573128 (visited on 06/05/2024) (cit. on p. 32).

Beede, E., E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis (Apr. 23, 2020). **A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy**. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, pp. 1–12. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376718. URL: https://dl.acm.org/doi/10.1145/3313831.3376718 (visited on 07/14/2023) (cit. on p. 2).

Bell, L. C. and E. Shimron (Jan. 2024). **Sharing Data Is Essential for the Future of AI in Medical Imaging**. In: *Radiology: Artificial Intelligence* 6.1, e230337. DOI: 10.1148/ryai.230337. URL: https://pubs.rsna.org/doi/10.1148/ryai.230337 (visited on 01/08/2025) (cit. on p. 19).

Bernhardt, M., F. D. S. Ribeiro, and B. Glocker (June 16, 2022). **Failure Detection in Medical Image Classification: A Reality Check and Benchmarking Testbed**. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=VBHuLfnOMf (visited on 12/13/2024) (cit. on p. 25).

Bloch, B. N., A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani (2015). **NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures (ISBI-MR-Prostate-2013)**. The Cancer Imaging Archive. DOI: 10.7937/K9/TCIA.2015.ZF0VLOPV (cit. on p. 52).

Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra (May 21, 2015). **Weight Uncertainty in Neural Networks**. In: arXiv. DOI: 10.48550/arXiv.1505.05424. arXiv: 1505.05424 [cs, stat]. URL: http://arxiv.org/abs/1505.05424 (visited on 09/24/2024) (cit. on p. 21).

Bogunović, H., F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller, K. Gopinath, A. K. Gostar, K. Jeon, Z. Ji, S. H. Kang, D. D. Koozekanani, D. Lu, D. Morley, K. K. Parhi, H. S. Park, A. Rashno, M. Sarunic, S. Shaikh, J. Sivaswamy, R. Tennakoon, S. Yadav, S. De Zanet, S. M. Waldstein, B. S.

Gerendas, C. Klaver, C. I. Sánchez, and U. Schmidt-Erfurth (Aug. 2019). **RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge**. In: *IEEE Transactions on Medical Imaging* 38.8, pp. 1858–1874. ISSN: 1558-254X. DOI: 10.1109/TMI.2019.2901398. URL: https://ieeexplore.ieee.org/document/8653407 (visited on 08/01/2024) (cit. on p. 53).

Boone, L., M. Biparva, P. Mojiri Forooshani, J. Ramirez, M. Masellis, R. Bartha, S. Symons, S. Strother, S. E. Black, C. Heyn, A. L. Martel, R. H. Swartz, and M. Goubran (Sept. 1, 2023). **ROOD-MRI: Benchmarking the Robustness of Deep Learning Segmentation Models to out-of-Distribution and Corrupted Data in MRI**. In: *NeuroImage* 278, p. 120289. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2023.120289. URL: https://www.sciencedirect.com/science/article/pii/S1053811923004408 (visited on 07/26/2024) (cit. on p. 19).

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). **Language Models Are Few-Shot Learners**. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (visited on 12/19/2024) (cit. on p. 6).

Bujotzek, M. R., Ü. Akünal, S. Denner, P. Neher, M. Zenk, E. Frodl, A. Jaiswal, M. Kim, N. R. Krekiehn, M. Nickel, R. Ruppel, M. Both, F. Döllinger, M. Opitz, T. Persigehl, J. Kleesiek, T. Penzkofer, K. Maier-Hein, A. Bucher, and R. Braren (Oct. 25, 2024). **Real-World Federated Learning in Radiology: Hurdles to Overcome and Benefits to Gain**. In: *Journal of the American Medical Informatics Association*, ocae259. ISSN: 1527-974X. DOI: 10.1093/jamia/ocae259. URL: https://doi.org/10.1093/jamia/ocae259 (visited on 11/05/2024) (cit. on pp. 20, 113, 173).

Bukhari, S. T. and H. Mohy-ud-Din (2021). **E1D3 U-Net for Brain Tumor Segmentation: Submission to the RSNA-ASNR-MICCAI BraTS 2021 Challenge**. In: *International MICCAI Brainlesion Workshop*, pp. 276–288 (cit. on p. 190).

Bungert, T. J., L. Kobelke, and P. F. Jäger (2023). **Understanding Silent Failures in Medical Image Classification**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor. Cham: Springer Nature Switzerland, pp. 400–410. ISBN: 978-3-031-43898-1. DOI: 10.1007/978-3-031-43898-1_39 (cit. on p. 25).

Campello, V. M., P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, M. Parreño, A. Albiol, F. Kong, S. C. Shadden, J. C. Acero, V. Sundaresan, M. Saber, M. Elattar, H. Li, B. Menze, F. Khader, C. Haarburger,

C. M. Scannell, M. Veta, A. Carscadden, K. Punithakumar, X. Liu, S. A. Tsaftaris, X. Huang, X. Yang, L. Li, X. Zhuang, D. Viladés, M. L. Descalzo, A. Guala, L. L. Mura, M. G. Friedrich, R. Garg, J. Lebel, F. Henriques, M. Karakas, E. Çavuş, S. E. Petersen, S. Escalera, S. Seguí, J. F. Rodríguez-Palomares, and K. Lekadir (Dec. 2021). **Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M Amp;Ms Challenge**. In: *IEEE Transactions on Medical Imaging* 40.12, pp. 3543–3554. ISSN: 1558-254X. DOI: 10.1109/TMI.2021.3090082 (cit. on pp. 2, 19, 20, 27, 52, 112).

Cardoso, M. J., W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, A. Myronenko, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M. Z. Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B. S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P. F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L. A. D. Cooper, H. R. Roth, D. Xu, D. Bericat, R. Floca, S. K. Zhou, H. Shuaib, K. Farahani, K. H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, and A. Feng (Nov. 4, 2022). **MONAI: An Open-Source Framework for Deep Learning in Healthcare**. DOI: 10.48550/arXiv.2211.02701. arXiv: 2211.02701 [cs]. URL: http://arxiv.org/abs/2211.02701 (visited on 02/05/2024). Pre-published (cit. on p. 56).

Carnahan, P., J. Moore, D. Bainbridge, M. Eskandari, E. C. S. Chen, and T. M. Peters (2021). **DeepMitral: Fully Automatic 3D Echocardiography Segmentation for Patient Specific Mitral Valve Modelling**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert. Cham: Springer International Publishing, pp. 459–468. ISBN: 978-3-030-87240-3. DOI: 10.1007/978-3-030-87240-3_44 (cit. on p. 53).

Carré, A., E. Deutsch, and C. Robert (2021). **Automatic Brain Tumor Segmentation with a Bridge-Unet Deeply Supervised Enhanced with Downsampling Pooling Combination, Atrous Spatial Pyramid Pooling, Squeeze-and-Excitation and EvoNorm**. In: *International MICCAI Brainlesion Workshop*, pp. 253–266 (cit. on p. 190).

Castro, D. C., I. Walker, and B. Glocker (July 22, 2020). **Causality Matters in Medical Imaging**. In: *Nature Communications* 11.1 (1), p. 3673. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17478-w. URL: https://www.nature.com/articles/s41467-020-17478-w (visited on 02/25/2021) (cit. on pp. 2, 9, 10).

Chao, P., C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin (2019). **Hardnet: A Low Memory Traffic Network**. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (cit. on p. 189).

Chen, C., K. Hammernik, C. Ouyang, C. Qin, W. Bai, and D. Rueckert (July 2, 2021). **Cooperative Training and Latent Space Data Augmentation for Robust Medical**

Image Segmentation. arXiv: 2107.01079 [cs, q-bio]. URL: http://arxiv.org/abs/2107.01079 (visited on 07/19/2021) (cit. on pp. 18, 19).

Chen, X., K. Men, B. Chen, Y. Tang, T. Zhang, S. Wang, Y. Li, and J. Dai (2020). **CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy**. In: *Frontiers in Oncology* 10. ISSN: 2234-943X. URL: https://www.frontiersin.org/article/10.3389/fonc.2020.00524 (visited on 02/24/2022) (cit. on p. 47).

Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger (2016). **3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation**. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Ed. by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells. Cham: Springer International Publishing, pp. 424–432. ISBN: 978-3-319-46723-8. DOI: 10.1007/978-3-319-46723-8_49 (cit. on p. 6).

Clark, K., B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior (Dec. 1, 2013). **The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository**. In: *Journal of Digital Imaging* 26.6, pp. 1045–1057. ISSN: 1618-727X. DOI: 10.1007/s10278-013-9622-7. URL: https://doi.org/10.1007/s10278-013-9622-7 (visited on 02/05/2024) (cit. on p. 52).

Crum, W., O. Camara, and D. Hill (Nov. 2006). **Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis**. In: *IEEE Transactions on Medical Imaging* 25.11, pp. 1451–1461. ISSN: 1558-254X. DOI: 10.1109/TMI.2006.880587 (cit. on p. 62).

Czolbe, S., K. Arnavaz, O. Krause, and A. Feragen (2021). **Is Segmentation Uncertainty Useful?** In: *Information Processing in Medical Imaging*. Ed. by A. Feragen, S. Sommer, J. Schnabel, and M. Nielsen. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 715–726. ISBN: 978-3-030-78191-0. DOI: 10.1007/978-3-030-78191-0_55 (cit. on p. 23).

Davatzikos, C., S. Rathore, S. Bakas, S. Pati, M. Bergman, R. Kalarot, P. Sridharan, A. Gastounioti, N. Jahani, E. Cohen, H. Akbari, B. Tunc, J. Doshi, D. Parker, M. Hsieh, A. Sotiras, H. Li, Y. Ou, R. K. Doot, M. Bilello, Y. Fan, R. T. Shinohara, P. Yushkevich, R. Verma, and D. Kontos (Jan. 2018). **Cancer Imaging Phenomics Toolkit: Quantitative Imaging Analytics for Precision Diagnostics and Predictive Modeling of Clinical Outcome**. In: *Journal of Medical Imaging* 5.1, p. 011018. ISSN: 2329-4302. DOI: 10.1117/1.JMI.5.1.011018. pmid: 29340286. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5764116/ (visited on 06/06/2024) (cit. on p. 37).

Dayan, I., H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, A. Liu, A. B. Costa, B. J. Wood, C.-S. Tsai, C.-H. Wang, C.-N. Hsu, C. K. Lee, P. Ruan, D. Xu, D. Wu, E. Huang, F. C. Kitamura, G. Lacey, G. C. de Antônio Corradi, G. Nino, H.-H. Shin, H.

Obinata, H. Ren, J. C. Crane, J. Tetreault, J. Guan, J. W. Garrett, J. D. Kaggie, J. G. Park, K. Dreyer, K. Juluru, K. Kersten, M. A. B. C. Rockenbach, M. G. Linguraru, M. A. Haider, M. AbdelMaseeh, N. Rieke, P. F. Damasceno, P. M. C. e Silva, P. Wang, S. Xu, S. Kawano, S. Sriswasdi, S. Y. Park, T. M. Grist, V. Buch, W. Jantarabenjakul, W. Wang, W. Y. Tak, X. Li, X. Lin, Y. J. Kwon, A. Quraini, A. Feng, A. N. Priest, B. Turkbey, B. Glicksberg, B. Bizzo, B. S. Kim, C. Tor-Díez, C.-C. Lee, C.-J. Hsu, C. Lin, C.-L. Lai, C. P. Hess, C. Compas, D. Bhatia, E. K. Oermann, E. Leibovitz, H. Sasaki, H. Mori, I. Yang, J. H. Sohn, K. N. K. Murthy, L.-C. Fu, M. R. F. de Mendonça, M. Fralick, M. K. Kang, M. Adil, N. Gangai, P. Vateekul, P. Elnajjar, S. Hickman, S. Majumdar, S. L. McLeod, S. Reed, S. Gräf, S. Harmon, T. Kodama, T. Puthanakit, T. Mazzulli, V. L. de Lavor, Y. Rakvongthai, Y. R. Lee, Y. Wen, F. J. Gilbert, M. G. Flores, and Q. Li (Oct. 2021). **Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19**. In: *Nature Medicine* 27.10 (10), pp. 1735–1743. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01506-3. URL: https://www.nature.com/articles/s41591-021-01506-3 (visited on 12/07/2021) (cit. on p. 20).

Demoustier, M., I. Khemir, Q. D. Nguyen, L. Martin-Gaffé, and N. Boutry (2021). **Residual 3d U-Net with Localization for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 389–399 (cit. on p. 190).

DeVries, T. and G. W. Taylor (July 2, 2018). **Leveraging Uncertainty Estimates for Predicting Segmentation Quality**. arXiv: 1807.00502 [cs]. URL: http://arxiv.org/abs/1807.00502 (visited on 02/24/2022) (cit. on pp. 24, 47).

Dobko, M., D.-I. Kolinko, O. Viniavskyi, and Y. Yelisieiev (2021). **Combining CNNs with Transformer for Multimodal 3D MRI Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 232–241 (cit. on p. 190).

Dong, X., J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo (Jan. 9, 2022). **CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows**. DOI: 10.48550/arXiv.2107.00652. arXiv: 2107.00652. URL: http://arxiv.org/abs/2107.00652 (visited on 10/30/2024). Pre-published (cit. on p. 73).

Dou, Q., T. Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H. H. C. Lee, K. Yu, Z. Feng, L. Dong, E. Burian, F. Jungmann, R. Braren, M. Makowski, B. Kainz, D. Rueckert, B. Glocker, S. C. H. Yu, and P. A. Heng (Mar. 29, 2021). **Federated Deep Learning for Detecting COVID-19 Lung Abnormalities in CT: A Privacy-Preserving Multinational Validation Study**. In: *npj Digital Medicine* 4.1 (1), pp. 1–11. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00431-6. URL: https://www.nature.com/articles/s41746-021-00431-6 (visited on 04/06/2021) (cit. on p. 20).

Druzhinina, P., E. Kondrateva, A. Bozhenko, V. Yarkin, M. Sharaev, and A. Kurmukov (2021). **BRATS2021: Exploring Each Sequence in Multi-modal Input for Baseline**

U-net Performance. In: *International MICCAI Brainlesion Workshop*, pp. 194–203 (cit. on p. 190).

Dusenberry, M., G. Jerfel, Y. Wen, Y. Ma, J. Snoek, K. Heller, B. Lakshminarayanan, and D. Tran (Nov. 21, 2020). **Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors**. In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 2782–2792. URL: https://proceedings.mlr.press/v119/dusenberry20a.html (visited on 09/24/2024) (cit. on p. 21).

Eisenmann, M., A. Reinke, V. Weru, M. D. Tizabi, F. Isensee, T. J. Adler, S. Ali, V. Andrearczyk, M. Aubreville, U. Baid, S. Bakas, N. Balu, S. Bano, J. Bernal, S. Bodenstedt, A. Casella, V. Cheplygina, M. Daum, M. de Bruijne, A. Depeursinge, R. Dorent, J. Egger, D. G. Ellis, S. Engelhardt, M. Ganz, N. Ghatwary, G. Girard, P. Godau, A. Gupta, L. Hansen, K. Harada, M. P. Heinrich, N. Heller, A. Hering, A. Huaulmé, P. Jannin, A. E. Kavur, O. Kodym, M. Kozubek, J. Li, H. Li, J. Ma, C. Martín-Isla, B. Menze, A. Noble, V. Oreiller, N. Padoy, S. Pati, K. Payette, T. Rädsch, J. Rafael-Patiño, V. S. Bawa, S. Speidel, C. H. Sudre, K. van Wijnen, M. Wagner, D. Wei, A. Yamlahi, M. H. Yap, C. Yuan, M. Zenk, A. Zia, D. Zimmerer, D. B. Aydogan, B. Bhattarai, L. Bloch, R. Brüngel, J. Cho, C. Choi, Q. Dou, I. Ezhov, C. M. Friedrich, C. D. Fuller, R. R. Gaire, A. Galdran, Á. G. Faura, M. Grammatikopoulou, S. Hong, M. Jahanifar, I. Jang, A. Kadkhodamohammadi, I. Kang, F. Kofler, S. Kondo, H. Kuijf, M. Li, M. Luu, T. Martinčič, P. Morais, M. A. Naser, B. Oliveira, D. Owen, S. Pang, J. Park, S.-H. Park, S. Plotka, E. Puybareau, N. Rajpoot, K. Ryu, N. Saeed, A. Shephard, P. Shi, D. Štepec, R. Subedi, G. Tochon, H. R. Torres, H. Urien, J. L. Vilaça, K. A. Wahid, H. Wang, J. Wang, L. Wang, X. Wang, B. Wiestler, M. Wodzinski, F. Xia, J. Xie, Z. Xiong, S. Yang, Y. Yang, Z. Zhao, K. Maier-Hein, P. F. Jäger, A. Kopp-Schneider, and L. Maier-Hein (2023). **Why Is the Winner the Best?** In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19955–19966. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Eisenmann_Why_Is_the_Winner_the_Best_CVPR_2023_paper.html (visited on 01/22/2025) (cit. on p. 172).

Esser, P., R. Rombach, and B. Ommer (June 23, 2021). **Taming Transformers for High-Resolution Image Synthesis**. DOI: 10.48550/arXiv.2012.09841. arXiv: 2012.09841 [cs]. URL: http://arxiv.org/abs/2012.09841 (visited on 09/13/2022). Pre-published (cit. on p. 25).

Feng, X., H. Bai, D. Kim, G. Maragkos, J. Machaj, and R. Kellogg (2021). **Brain Tumor Segmentation with Patch-Based 3D Attention UNet from Multi-parametric MRI**. In: *International MICCAI Brainlesion Workshop*, pp. 90–96 (cit. on p. 190).

Fidon, L., S. Shit, I. Ezhov, J. C. Paetzold, S. Ourselin, and T. Vercauteren (2021). **Generalized Wasserstein Dice Loss, Test-Time Augmentation, and Transformers for the BraTS**

2021 Challenge. In: *International MICCAI Brainlesion Workshop*, pp. 187–196 (cit. on pp. 189, 190).

Full, P. M., F. Isensee, P. F. Jäger, and K. Maier-Hein (Nov. 15, 2020). **Studying Robustness of Semantic Segmentation under Domain Shift in Cardiac MRI**. arXiv: 2011.07592 [cs, eess]. URL: http://arxiv.org/abs/2011.07592 (visited on 03/15/2021) (cit. on pp. 18, 52).

Fumero, F., S. Alayon, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez (June 2011). **RIM-ONE: An Open Retinal Image Database for Optic Nerve Evaluation**. In: *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*. 2011 24th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. DOI: 10.1109/CBMS.2011.5999143. URL: https://ieeexplore.ieee.org/abstract/document/5999143 (visited on 09/25/2024) (cit. on p. 53).

Futrega, M., A. Milesi, M. Marcinkiewicz, and P. Ribalta (2021). **Optimized U-Net for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 15–29 (cit. on p. 190).

Gal, Y. (2016). **Uncertainty in Deep Learning**. University of Cambridge (cit. on pp. 11, 12, 14).

Gal, Y. and Z. Ghahramani (2016). **Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning**. In: p. 10 (cit. on pp. 21, 60).

Galil, I., M. Dabbah, and R. El-Yaniv (Feb. 23, 2023). **What Can We Learn From The Selective Prediction And Uncertainty Estimation Performance Of 523 Imagenet Classifiers**. DOI: 10.48550/arXiv.2302.11874. arXiv: 2302.11874 [cs]. URL: http://arxiv.org/abs/2302.11874 (visited on 01/09/2024). Pre-published (cit. on p. 117).

Geifman, Y. and R. El-Yaniv (May 23, 2017). **Selective Classification for Deep Neural Networks**. In: URL: https://arxiv.org/abs/1705.08500v2 (visited on 02/11/2022) (cit. on pp. 44, 121).

Gillies, R. J., P. E. Kinahan, and H. Hricak (Nov. 18, 2015). **Radiomics: Images Are More than Pictures, They Are Data**. In: *Radiology* 278.2, p. 563. DOI: 10.1148/radiol.2015151169. pmid: 26579733. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4734157/ (visited on 11/15/2024) (cit. on p. 61).

González, C., K. Gotkowski, M. Fuchs, A. Bucher, A. Dadras, R. Fischbach, I. J. Kaltenborn, and A. Mukhopadhyay (Nov. 2022). **Distance-Based Detection of out-of-Distribution Silent Failures for Covid-19 Lung Lesion Segmentation**. In: *Medical Image Analysis* 82, p. 102596. ISSN: 13618415. DOI: 10.1016/j.media.2022.102596. URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841522002298 (visited on 11/02/2022) (cit. on pp. 3, 19, 23, 25, 45, 47, 52, 64, 119).

Graham, M. S., P.-D. Tudosiu, P. Wright, W. H. L. Pinaya, J.-M. U-King-Im, Y. Mah, J. Teo, R. H. Jäger, D. Werring, P. Nachev, S. Ourselin, and M. J. Cardoso (June 22, 2022). **Transformer-Based out-of-Distribution Detection for Clinically Safe Segmentation**. In: Medical Imaging with Deep Learning. URL: https://openreview.net/forum?id=En7660i-CLJ (visited on 08/10/2022) (cit. on pp. 25, 45).

Griethuysen, J. J. M. van, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts (Nov. 1, 2017). **Computational Radiomics System to Decode the Radiographic Phenotype**. In: *Cancer Research* 77.21, e104–e107. ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-17-0339. pmid: 29092951 (cit. on pp. 23, 62).

Gulrajani, I. and D. Lopez-Paz (2021). **In Search of Lost Domain Generalization**. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=lQdXeXDoWtI (cit. on pp. 18, 19).

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger (July 17, 2017). **On Calibration of Modern Neural Networks**. In: *Proceedings of the 34th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html (visited on 01/26/2022) (cit. on p. 21).

Han, S., Y. Liu, S. J. Cai, M. Qian, J. Ding, M. Larion, M. R. Gilbert, and C. Yang (May 2020). **IDH Mutation in Glioma: Molecular Mechanisms and Potential Therapeutic Targets**. In: *British Journal of Cancer* 122.11, pp. 1580–1589. ISSN: 1532-1827. DOI: 10.1038/s41416-020-0814-x. URL: https://www.nature.com/articles/s41416-020-0814-x (visited on 10/30/2024) (cit. on p. 34).

Harari, P. M., S. Song, and W. A. Tomé (2010). **Emphasizing Conformal Avoidance versus Target Definition for IMRT Planning in Head-and-Neck Cancer**. In: *International Journal of Radiation Oncology*Biology*Physics* 77.3, pp. 950–958. ISSN: 0360-3016. DOI: 10.1016/j.ijrobp.2009.09.062. URL: https://www.sciencedirect.com/science/article/pii/S0360301609033975 (cit. on p. 1).

He, K., X. Zhang, S. Ren, and J. Sun (2016). **Deep Residual Learning for Image Recognition**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (cit. on pp. 56, 189).

He, Y., H.-P. Wang, M. Zenk, and M. Fritz (May 1, 2022). **CosSGD: Communication-Efficient Federated Learning with a Simple Cosine-Based Quantization**. DOI: 10.48550/arXiv.2012.08241. arXiv: 2012.08241 [cs]. URL: http://arxiv.org/abs/2012.08241 (visited on 12/14/2022). Pre-published (cit. on p. 172).

He, Y., A. Carass, L. Zuo, B. E. Dewey, and J. L. Prince (Aug. 1, 2021). **Autoencoder Based Self-Supervised Test-Time Adaptation for Medical Image Analysis**. In: *Medical Image Analysis* 72, p. 102136. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102136. URL:

https://www.sciencedirect.com/science/article/pii/S1361841521001821 (visited on 11/10/2021) (cit. on p. 18).

Heimann, T. and H.-P. Meinzer (Aug. 1, 2009). **Statistical Shape Models for 3D Medical Image Segmentation: A Review**. In: *Medical Image Analysis* 13.4, pp. 543–563. ISSN: 1361-8415. DOI: 10.1016/j.media.2009.05.004. URL: https://www.sciencedirect.com/science/article/pii/S1361841509000425 (visited on 12/18/2024) (cit. on p. 5).

Heller, N., F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, G. Yao, Y. Gao, Y. Zhang, Y. Wang, F. Hou, J. Yang, G. Xiong, J. Tian, C. Zhong, J. Ma, J. Rickman, J. Dean, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, H. Kaluzniak, S. Raza, J. Rosenberg, K. Moore, E. Walczak, Z. Rengel, Z. Edgerton, R. Vasdev, M. Peterson, S. McSweeney, S. Peterson, A. Kalapara, N. Sathianathen, N. Papanikolopoulos, and C. Weight (Jan. 1, 2021). **The State of the Art in Kidney and Kidney Tumor Segmentation in Contrast-Enhanced CT Imaging: Results of the KiTS19 Challenge**. In: *Medical Image Analysis* 67, p. 101821. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101821. URL: https://www.sciencedirect.com/science/article/pii/S1361841520301857 (visited on 02/05/2024) (cit. on p. 52).

Heller, N., F. Isensee, D. Trofimova, R. Tejpaul, Z. Zhao, H. Chen, L. Wang, A. Golts, D. Khapun, D. Shats, Y. Shoshan, F. Gilboa-Solomon, Y. George, X. Yang, J. Zhang, J. Zhang, Y. Xia, M. Wu, Z. Liu, E. Walczak, S. McSweeney, R. Vasdev, C. Hornung, R. Solaiman, J. Schoephoerster, B. Abernathy, D. Wu, S. Abdulkadir, B. Byun, J. Spriggs, G. Struyk, A. Austin, B. Simpson, M. Hagstrom, S. Virnig, J. French, N. Venkatesh, S. Chan, K. Moore, A. Jacobsen, S. Austin, M. Austin, S. Regmi, N. Papanikolopoulos, and C. Weight (July 4, 2023). **The KiTS21 Challenge: Automatic Segmentation of Kidneys, Renal Tumors, and Renal Cysts in Corticomedullary-Phase CT**. DOI: 10.48550/arXiv.2307.01984. arXiv: 2307.01984 [cs]. URL: http://arxiv.org/abs/2307.01984 (visited on 02/05/2024). Pre-published (cit. on p. 52).

Hendrycks, D. and T. Dietterich (Mar. 28, 2019). **Benchmarking Neural Network Robustness to Common Corruptions and Perturbations**. arXiv: 1903.12261 [cs, stat]. URL: http://arxiv.org/abs/1903.12261 (visited on 03/02/2021) (cit. on p. 19).

Higgins, I., L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner (Nov. 4, 2016). **Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework**. In: URL: https://openreview.net/forum?id=Sy2fzU9gl (visited on 08/22/2019) (cit. on p. 64).

Hoebel, K., V. Andrearczyk, A. Beers, J. Patel, K. Chang, A. Depeursinge, H. Müller, and J. Kalpathy-Cramer (Mar. 10, 2020). **An Exploration of Uncertainty Information for Segmentation Quality Assessment**. In: *Medical Imaging 2020: Image Processing*. Medical Imaging 2020: Image Processing. Vol. 11313. SPIE, pp. 381–390. DOI: 10.1117/12.2

548722. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11313/113131K/An-exploration-of-uncertainty-information-for-segmentation-quality-assessment/10.1117/12.2548722.full (visited on 11/10/2023) (cit. on pp. 22, 118).

Hoebel, K., C. Bridge, A. Lemay, K. Chang, J. Patel, B. R. M.d, and J. Kalpathy-Cramer (Apr. 4, 2022). **Do I Know This? Segmentation Uncertainty under Domain Shift**. In: *Medical Imaging 2022: Image Processing*. Medical Imaging 2022: Image Processing. Vol. 12032. SPIE, pp. 261–276. DOI: 10.1117/12.2611867. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12032/1203211/Do-I-know-this-segmentation-uncertainty-under-domain-shift/10.1117/12.2611867.full (visited on 06/22/2023) (cit. on pp. 22, 25, 26, 52, 118).

Hsu, C., C. Chang, T. W. Chen, H. Tsai, S. Ma, and W. Wang (2021). **Brain Tumor Segmentation (BraTS) Challenge Short Paper: Improving Three-Dimensional Brain Tumor Segmentation Using SegResNet and Hybrid Boundary-Dice Loss**. In: *International MICCAI Brainlesion Workshop*, pp. 334–344 (cit. on p. 190).

Iglesias, J. E. and M. R. Sabuncu (Aug. 1, 2015). **Multi-Atlas Segmentation of Biomedical Images: A Survey**. In: *Medical Image Analysis* 24.1, pp. 205–219. ISSN: 1361-8415. DOI: 10.1016/j.media.2015.06.012. URL: https://www.sciencedirect.com/science/article/pii/S1361841515000997 (visited on 12/18/2024) (cit. on p. 5).

Isensee, F., P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein (Feb. 2021a). **nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation**. In: *Nature Methods* 18.2 (2), pp. 203–211. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z. URL: https://www.nature.com/articles/s41592-020-01008-z (visited on 03/16/2021) (cit. on pp. 2, 3, 7, 38, 55, 56, 79).

Isensee, F., P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein (2021b). **nnU-Net for Brain Tumor Segmentation**. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*. Springer, pp. 118–132 (cit. on pp. 69, 72, 73, 86).

Isensee, F., T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jäger (2024). **nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel. Cham: Springer Nature Switzerland, pp. 488–498. ISBN: 978-3-031-72114-4. DOI: 10.1007/978-3-031-72114-4_47 (cit. on pp. 6, 7).

Jaeger, P. F., C. T. Lüth, L. Klein, and T. J. Bungert (Sept. 29, 2022). **A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification**. In: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=YnkGMIh0gvX (visited on 03/21/2024) (cit. on pp. 4, 25, 28, 44–46, 48, 50, 115, 117, 120).

Jia, H., C. Bai, W. Cai, H. Huang, and Y. Xia (2021). **HNF-Netv2 for Brain Tumor Segmentation Using Multi-Modal MR Imaging**. In: *International MICCAI Brainlesion Workshop*, pp. 106–115 (cit. on pp. 189, 190).

Jiang, M., H. Yang, X. Zhang, S. Zhang, and Q. Dou (2022). **Efficient Federated Tumor Segmentation via Parameter Distance Weighted Aggregation and Client Pruning**. In: *International MICCAI Brainlesion Workshop*, pp. 161–172 (cit. on p. 72).

Jiang, Z., C. Zhao, X. Liu, and M. G. Linguraru (2021). **Brain Tumor Segmentation in Multi-Parametric Magnetic Resonance Imaging Using Model Ensembling and Super-Resolution**. In: *International MICCAI Brainlesion Workshop*, pp. 125–137 (cit. on p. 190).

Joskowicz, L., D. Cohen, N. Caplan, and J. Sosna (Mar. 1, 2019). **Inter-Observer Variability of Manual Contour Delineation of Structures in CT**. In: *European Radiology* 29.3, pp. 1391–1399. ISSN: 1432-1084. DOI: 10.1007/s00330-018-5695-5. URL: https://doi.org/10.1007/s00330-018-5695-5 (visited on 11/23/2020) (cit. on p. 1).

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis (Aug. 2021). **Highly Accurate Protein Structure Prediction with AlphaFold**. In: *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 12/19/2024) (cit. on p. 5).

Jun, M., G. Cheng, W. Yixin, A. Xingle, G. Jiantao, Y. Ziqi, Z. Minqing, L. Xin, D. Xueyuan, C. Shucheng, W. Hao, M. Sen, Y. Xiaoyu, N. Ziwei, L. Chen, T. Lu, Z. Yuntao, Z. Qiongjie, D. Guoqiang, and H. Jian (Apr. 20, 2020). **COVID-19 CT Lung and Infection Segmentation Dataset**. Version Verson 1.0. Zenodo. DOI: 10.5281/zenodo.3757476. URL: https://zenodo.org/records/3757476 (visited on 02/05/2024) (cit. on p. 52).

Jungo, A., F. Balsiger, and M. Reyes (2020). **Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation**. In: *Frontiers in Neuroscience* 14. DOI: 10.3389/fnins.2020.00282. URL: https://www.readcube.com/articles/10.3389%2Ffnins.2020.00282 (visited on 11/17/2022) (cit. on pp. 22, 23, 25, 26, 46, 47, 61, 62, 116–119).

Kades, K., J. Scherer, M. Zenk, M. Kempf, and K. Maier-Hein (2022). **Towards Real-World Federated Learning in Medical Image Analysis Using Kaapana**. In: *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health*. Ed. by S. Albarqouni, S. Bakas, S. Bano, M. J. Cardoso, B. Khanal, B. Landman, X. Li, C. Qin, I. Rekik, N. Rieke, H. Roth, D. Sheet, and D. Xu. Cham: Springer Nature Switzerland, pp. 130–140. ISBN: 978-3-031-18523-6. DOI: 10.1007/978-3-031-18523-6_13 (cit. on p. 172).

Kahl, K.-C., C. T. Lüth, M. Zenk, K. Maier-Hein, and P. F. Jaeger (Jan. 16, 2024). **ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation**. DOI: 10.48550/arXiv.2401.08501. arXiv: 2401.08501 [cs]. URL: http://arxiv.org/abs/2401.08501 (visited on 01/25/2024). Pre-published (cit. on pp. 14, 22, 23, 26, 58, 60, 61, 116, 118–120, 173).

Kalidindi, S. and S. Gandhi (Aug. 21, 2023). **Workforce Crisis in Radiology in the UK and the Strategies to Deal With It: Is Artificial Intelligence the Saviour?** In: *Cureus* 15.8, e43866. DOI: 10.7759/cureus.43866. pmid: 37608900. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10441819/ (visited on 10/22/2024) (cit. on p. 1).

Kamnitsas, K., C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker (Feb. 1, 2017). **Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation**. In: *Medical Image Analysis* 36, pp. 61–78. ISSN: 1361-8415. DOI: 10.1016/j.media.2016.10.004. URL: http://www.sciencedirect.com/science/article/pii/S1361841516301839 (visited on 09/02/2019) (cit. on p. 38).

Karani, N., E. Erdil, K. Chaitanya, and E. Konukoglu (Feb. 1, 2021). **Test-Time Adaptable Neural Networks for Robust Medical Image Segmentation**. In: *Medical Image Analysis* 68, p. 101907. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101907. URL: https://www.sciencedirect.com/science/article/pii/S1361841520302711 (visited on 05/08/2021) (cit. on p. 18).

Karargyris, A., R. Umeton, M. J. Sheller, A. Aristizabal, J. George, A. Wuest, S. Pati, H. Kassem, M. Zenk, U. Baid, P. Narayana Moorthy, A. Chowdhury, J. Guo, S. Nalawade, J. Rosenthal, D. Kanter, M. Xenochristou, D. J. Beutel, V. Chung, T. Bergquist, J. Eddy, A. Abid, L. Tunstall, O. Sanseviero, D. Dimitriadis, Y. Qian, X. Xu, Y. Liu, R. S. M. Goh, S. Bala, V. Bittorf, S. R. Puchala, B. Ricciuti, S. Samineni, E. Sengupta, A. Chaudhari, C. Coleman, B. Desinghu, G. Diamos, D. Dutta, D. Feddema, G. Fursin, X. Huang, S. Kashyap, N. Lane, I. Mallick, P. Mascagni, V. Mehta, C. F. Moraes, V. Natarajan, N. Nikolov, N. Padoy, G. Pekhimenko, V. J. Reddi, G. A. Reina, P. Ribalta, A. Singh, J. J. Thiagarajan, J. Albrecht, T. Wolf, G. Miller, H. Fu, P. Shah, D. Xu, P. Yadav, D. Talby, M. M. Awad, J. P. Howard, M. Rosenthal, L. Marchionni, M. Loda, J. M. Johnson, S.

Bakas, and P. Mattson (July 17, 2023). **Federated Benchmarking of Medical Artificial Intelligence with MedPerf**. In: *Nature Machine Intelligence*, pp. 1–12. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00652-2. URL: https://www.nature.com/articles/s42256-023-00652-2 (visited on 07/18/2023) (cit. on pp. 3, 20, 42, 170, 172).

Kazerooni, A. F., N. Khalili, X. Liu, D. Haldar, Z. Jiang, S. M. Anwar, J. Albrecht, M. Adewole, U. Anazodo, H. Anderson, S. Bagheri, U. Baid, T. Bergquist, A. J. Borja, E. Calabrese, V. Chung, G.-M. Conte, F. Dako, J. Eddy, I. Ezhov, A. Familiar, K. Farahani, S. Haldar, J. E. Iglesias, A. Janas, E. Johansen, B. V. Jones, F. Kofler, D. LaBella, H. A. Lai, K. V. Leemput, H. B. Li, N. Maleki, A. S. McAllister, Z. Meier, B. Menze, A. W. Moawad, K. K. Nandolia, J. Pavaine, M. Piraud, T. Poussaint, S. P. Prabhu, Z. Reitman, A. Rodriguez, J. D. Rudie, M. Sanchez-Montano, I. S. Shaikh, L. M. Shah, N. Sheth, R. T. Shinohara, W. Tu, K. Viswanathan, C. Wang, J. B. Ware, B. Wiestler, W. Wiggins, A. Zapaishchykova, M. Aboian, M. Bornhorst, P. de Blank, M. Deutsch, M. Fouladi, L. Hoffman, B. Kann, M. Lazow, L. Mikael, A. Nabavizadeh, R. Packer, A. Resnick, B. Rood, A. Vossough, S. Bakas, and M. G. Linguraru (2024). **The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-peds)**. arXiv: 2305.17033 [eess.IV] (cit. on p. 3).

Kendall, A., V. Badrinarayanan, and R. Cipolla (Oct. 10, 2016). **Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding**. DOI: 10.48550/arXiv.1511.02680. arXiv: 1511.02680 [cs]. URL: http://arxiv.org/abs/1511.02680 (visited on 02/07/2023). Pre-published (cit. on p. 56).

Kickingereder, P., F. Isensee, I. Tursunova, J. Petersen, U. Neuberger, D. Bonekamp, G. Brugnara, M. Schell, T. Kessler, M. Foltyn, I. Harting, F. Sahm, M. Prager, M. Nowosielski, A. Wick, M. Nolden, A. Radbruch, J. Debus, H.-P. Schlemmer, S. Heiland, M. Platten, A. von Deimling, M. J. van den Bent, T. Gorlia, W. Wick, M. Bendszus, and K. H. Maier-Hein (May 1, 2019). **Automated Quantitative Tumour Response Assessment of MRI in Neuro-Oncology with Artificial Neural Networks: A Multicentre, Retrospective Study**. In: *The Lancet Oncology* 20.5, pp. 728–740. ISSN: 1470-2045, 1474-5488. DOI: 10.1016/S1470-2045(19)30098-1. pmid: 30952559. URL: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045%2819%2930098-1/fulltext (visited on 10/25/2024) (cit. on pp. 1, 4).

Kingma, D. P. and M. Welling (Dec. 20, 2013). **Auto-Encoding Variational Bayes**. arXiv: 1312.6114 [cs, stat]. URL: http://arxiv.org/abs/1312.6114 (visited on 07/22/2019) (cit. on p. 24).

Kirchhoff, Y., M. R. Rokuss, S. Roy, B. Kovacs, C. Ulrich, T. Wald, M. Zenk, P. Vollmuth, J. Kleesiek, F. Isensee, and K. Maier-Hein (2024). **Skeleton Recall Loss for Connectivity Conserving and Resource Efficient Segmentation of Thin Tubular Structures**. In:

Computer Vision – ECCV 2024. Ed. by A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol. Cham: Springer Nature Switzerland, pp. 218–234. ISBN: 978-3-031-72980-5. DOI: 10.1007/978-3-031-72980-5_13 (cit. on p. 173).

Kirillov, A., E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick (2023). **Segment Anything**. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026. URL: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html (visited on 12/19/2024) (cit. on p. 6).

Koh, P. W., S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang (July 16, 2021). **WILDS: A Benchmark of in-the-Wild Distribution Shifts**. arXiv: 2012.07421 [cs]. URL: http://arxiv.org/abs/2012.07421 (visited on 11/30/2021) (cit. on pp. 9, 19).

Kohl, S. A. A., B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. A. Eslami, D. J. Rezende, and O. Ronneberger (June 13, 2018). **A Probabilistic U-Net for Segmentation of Ambiguous Images**. arXiv: 1806.05034 [cs, stat]. URL: http://arxiv.org/abs/1806.05034 (visited on 08/27/2019) (cit. on p. 22).

Kohlberger, T., V. Singh, C. Alvino, C. Bahlmann, and L. Grady (2012). **Evaluating Segmentation Error without Ground Truth**. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*. Ed. by N. Ayache, H. Delingette, P. Golland, and K. Mori. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 528–536. ISBN: 978-3-642-33415-3. DOI: 10.1007/978-3-642-33415-3_65 (cit. on pp. 23, 46).

Korevaar, S., R. Tennakoon, and A. Bab-Hadiashar (2023). **Failure to Achieve Domain Invariance With Domain Generalization Algorithms: An Analysis in Medical Imaging**. In: *IEEE Access* 11, pp. 39351–39372. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3268704. URL: https://ieeexplore.ieee.org/abstract/document/10105917 (visited on 08/13/2024) (cit. on p. 19).

Kotowski, K., S. Adamski, B. Machura, W. Malara, L. Zarudzki, and J. Nalepa (2022). **Federated Evaluation of nnU-Nets Enhanced with Domain Knowledge for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 218–227 (cit. on p. 72).

Kotowski, K., S. Adamski, B. Machura, L. Zarudzki, and J. Nalepa (2021). **Coupling nnU-nets with Expert Knowledge for Accurate Brain Tumor Segmentation from MRI**. In: *International MICCAI Brainlesion Workshop*, pp. 197–209 (cit. on pp. 72, 190).

Kurtzer, G. M., V. Sochat, and M. W. Bauer (2017). **Singularity: Scientific Containers for Mobility of Compute**. In: *PloS one* 12.5, e0177459 (cit. on p. 42).

Kushibar, K., V. Campello, L. Garrucho, A. Linardos, P. Radeva, and K. Lekadir (2022). **Layer Ensembles: A Single-Pass Uncertainty Estimation in Deep Learning for Segmentation**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li. Vol. 13438. Cham: Springer Nature Switzerland, pp. 514–524. ISBN: 978-3-031-16452-1. DOI: 10.1007/978-3-031-16452-1_49. URL: https://link.springer.com/10.1007/978-3-031-16452-1_49 (visited on 01/02/2023) (cit. on p. 52).

Kwon, Y., J.-H. Won, B. J. Kim, and M. C. Paik (Feb. 1, 2020). **Uncertainty Quantification Using Bayesian Neural Networks in Classification: Application to Biomedical Image Segmentation**. In: *Computational Statistics & Data Analysis* 142, p. 106816. ISSN: 0167-9473. DOI: 10.1016/j.csda.2019.106816. URL: https://www.sciencedirect.com/science/article/pii/S016794731930163X (visited on 06/27/2023) (cit. on p. 22).

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). **Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles**. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html (visited on 02/15/2024) (cit. on pp. 22, 47, 60).

Lambert, B., F. Forbes, S. Doyle, H. Dehaene, and M. Dojat (Apr. 1, 2024). **Trustworthy Clinical AI Solutions: A Unified Review of Uncertainty Quantification in Deep Learning Models for Medical Image Analysis**. In: *Artificial Intelligence in Medicine* 150, p. 102830. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2024.102830. URL: https://www.sciencedirect.com/science/article/pii/S0933365724000721 (visited on 05/29/2024) (cit. on pp. 21, 119).

LeCun, Y., Y. Bengio, and G. Hinton (May 2015). **Deep Learning**. In: *Nature* 521.7553 (7553), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539 (visited on 05/15/2020) (cit. on pp. 1, 5, 6).

Lemaître, G., R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau (May 1, 2015). **Computer-Aided Detection and Diagnosis for Prostate Cancer Based on Mono and Multi-Parametric MRI: A Review**. In: *Computers in Biology and Medicine* 60, pp. 8–31. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2015.02.009. URL: https://www.sciencedirect.com/science/article/pii/S001048251500058X (visited on 02/05/2024) (cit. on p. 52).

Lennartz, J. and T. Schultz (2023). **Segmentation Distortion: Quantifying Segmentation Uncertainty Under Domain Shift via the Effects of Anomalous Activations**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor.

Vol. 14222. Cham: Springer Nature Switzerland, pp. 316–325. ISBN: 978-3-031-43898-1. DOI: 10.1007/978-3-031-43898-1_31. URL: https://link.springer.com/10.1007/978-3-031-43898-1_31 (visited on 10/18/2023) (cit. on p. 119).

Li, K., L. Yu, and P.-A. Heng (May 1, 2022). **Towards Reliable Cardiac Image Segmentation: Assessing Image-Level and Pixel-Level Segmentation Quality via Self-Reflective References**. In: *Medical Image Analysis* 78, p. 102426. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102426. URL: https://www.sciencedirect.com/science/article/pii/S1361841522000779 (visited on 01/04/2023) (cit. on pp. 24, 26, 46, 47, 119, 120).

Li, Z., Z. Shen, J. Wen, T. He, and L. Pan (2021). **Automatic Brain Tumor Segmentation Using Multi-Scale Features and Attention Mechanism**. In: *International MICCAI Brainlesion Workshop*, pp. 216–226 (cit. on p. 190).

Lin, Q., X. Chen, C. Chen, and J. M. Garibaldi (2022). **A Novel Quality Control Algorithm for Medical Image Segmentation Based on Fuzzy Uncertainty**. In: *IEEE Transactions on Fuzzy Systems*, pp. 1–14. ISSN: 1941-0034. DOI: 10.1109/TFUZZ.2022.3228332 (cit. on p. 47).

Lin, W.-W., T. Li, T.-M. Huang, J.-W. Lin, M.-H. Yueh, and S.-T. Yau (2021). **A Two-Phase Optimal Mass Transportation Technique for 3d Brain Tumor Detection and Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 400–409 (cit. on p. 190).

Litjens, G., B. van Ginneken, H. Huisman, W. van de Ven, C. Hoeks, D. Barratt, and A. Madabhushi (June 7, 2023). **PROMISE12: Data from the MICCAI Grand Challenge: Prostate MR Image Segmentation 2012**. Version Updated the zip files, fixed some issues. Zenodo. DOI: 10.5281/zenodo.8026660. URL: https://zenodo.org/records/8026660 (visited on 02/05/2024) (cit. on p. 52).

Liu, F., Y. Xia, D. Yang, A. L. Yuille, and D. Xu (2019). **An Alarm System for Segmentation Algorithm Based on Shape Model**. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10652–10661. URL: https://openaccess.thecvf.com/content_ICCV_2019/html/Liu_An_Alarm_System_for_Segmentation_Algorithm_Based_on_Shape_Model_ICCV_2019_paper.html (visited on 08/16/2022) (cit. on pp. 25, 46, 64, 120).

Liu, Q., Q. Dou, and P.-A. Heng (2020). **Shape-Aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 475–485. ISBN: 978-3-030-59713-9. DOI: 10.1007/978-3-030-59713-9_46 (cit. on pp. 19, 52).

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo (Aug. 17, 2021). **Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows**. DOI: 10.48550/arXiv.2103.14030. arXiv: 2103.14030. URL: http://arxiv.org/abs/2103.14030 (visited on 10/30/2024). Pre-published (cit. on pp. 73, 189).

Loshchilov, I. and F. Hutter (Jan. 4, 2019). **Decoupled Weight Decay Regularization**. arXiv: 1711.05101 [cs, math]. URL: http://arxiv.org/abs/1711.05101 (visited on 03/09/2020) (cit. on p. 63).

Luu, H. M. and S.-H. Park (2021). **Extending Nn-UNet for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 173–186 (cit. on p. 190).

Ma, J. and J. Chen (2021). **Nnunet with Region-Based Training and Loss Ensembles for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 421–430 (cit. on p. 190).

Maier-Hein, L., M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, C. Feldmann, A. F. Frangi, P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B. A. Landman, K. März, O. Maier, K. Maier-Hein, B. H. Menze, H. Müller, P. F. Neher, W. Niessen, N. Rajpoot, G. C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, and A. Kopp-Schneider (Dec. 6, 2018). **Why Rankings of Biomedical Image Analysis Competitions Should Be Interpreted with Care**. In: *Nature Communications* 9.1, pp. 1–13. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07619-7. URL: https://www.nature.com/articles/s41467-018-07619-7 (visited on 09/19/2019) (cit. on pp. 2, 15, 16, 31, 40, 41, 53).

Maier-Hein, L., A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, M. Wiesenfarth, A. E. Kavur, C. H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. B. Blaschko, M. J. Cardoso, V. Cheplygina, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. I. Sánchez, S. Shetty, M. van Smeden, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. Van Calster, G. Varoquaux, and P. F. Jäger (Feb. 2024). **Metrics Reloaded: Recommendations for Image Analysis Validation**. In: *Nature Methods* 21.2, pp. 195–212. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02151-z. URL: https://www.nature.com/articles/s41592-023-02151-z (visited on 06/06/2024) (cit. on pp. 9, 40).

Maier-Hein, L., A. Reinke, M. Kozubek, A. L. Martel, T. Arbel, M. Eisenmann, A. Hanbury, P. Jannin, H. Müller, S. Onogur, J. Saez-Rodriguez, B. van Ginneken, A. Kopp-Schneider, and B. A. Landman (Dec. 1, 2020). **BIAS: Transparent Reporting of Biomedical Image Analysis Challenges**. In: *Medical Image Analysis* 66, p. 101796. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101796. URL: https://www.sciencedirect.com/science/article/pii/S1361841520301602 (visited on 05/08/2021) (cit. on pp. 15, 32).

Malinin, A., N. Band, Ganshin, Alexander, G. Chesnokov, Y. Gal, M. J. F. Gales, A. Noskov, A. Ploskonosov, L. Prokhorenkova, I. Provilkov, V. Raina, V. Raina, Roginskiy, Denis, M. Shmatova, P. Tigas, and B. Yangel (Feb. 11, 2022). **Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks**. DOI: 10.48550/arXiv.2107.07455. arXiv: 2107.07455 [cs, stat]. URL: http://arxiv.org/abs/2107.07455 (visited on 04/04/2024). Pre-published (cit. on pp. 20, 26, 27, 117).

Malinin, A. and M. Gales (2018). **Predictive Uncertainty Estimation via Prior Networks**. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. URL: https://papers.nips.cc/paper_files/paper/2018/hash/3ea2db50e62ceefceaf70a9d9a56a6f4-Abstract.html (visited on 09/24/2024) (cit. on p. 22).

Maurya, S., V. Kumar Yadav, S. Agarwal, and A. Singh (2021). **Brain Tumor Segmentation in mpMRI Scans (BraTS-2021) Using Models Based on U-net Architecture**. In: *International MICCAI Brainlesion Workshop*, pp. 312–323 (cit. on p. 190).

McKinley, R., R. Meier, and R. Wiest (2018). **Ensembles of Densely-Connected CNNs with Label-Uncertainty for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 456–465 (cit. on p. 38).

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). **Communication-Efficient Learning of Deep Networks from Decentralized Data**. In: *Artificial Intelligence and Statistics*. PMLR, pp. 1273–1282 (cit. on p. 68).

Mehrtash, A., W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur (Dec. 2020). **Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation**. In: *IEEE Transactions on Medical Imaging* 39.12, pp. 3868–3878. ISSN: 1558-254X. DOI: 10.1109/TMI.2020.3006437 (cit. on pp. 3, 22, 23, 25, 26, 47, 60, 120).

Mehta, R., A. Filos, U. Baid, C. Sako, R. McKinley, M. Rebsamen, K. Dätwyler, R. Meier, P. Radojewski, G. K. Murugesan, S. Nalawade, C. Ganesh, B. Wagner, F. F. Yu, B. Fei, A. J. Madhuranthakam, J. A. Maldjian, L. Daza, C. Gómez, P. Arbeláez, C. Dai, S. Wang, H. Reynaud, Y. Mo, E. Angelini, Y. Guo, W. Bai, S. Banerjee, L. Pei, M. Ak, S. Rosas-González, I. Zemmoura, C. Tauber, M. H. Vu, T. Nyholm, T. Löfstedt, L. M. Ballestar, V. Vilaplana, H. McHugh, G. Maso Talou, A. Wang, J. Patel, K. Chang, K. Hoebel, M.

Gidwani, N. Arun, S. Gupta, M. Aggarwal, P. Singh, E. R. Gerstner, J. Kalpathy-Cramer, N. Boutry, A. Huard, L. Vidyaratne, M. M. Rahman, K. M. Iftekharuddin, J. Chazalon, E. Puybareau, G. Tochon, J. Ma, M. Cabezas, X. Llado, A. Oliver, L. Valencia, S. Valverde, M. Amian, M. Soltaninejad, A. Myronenko, A. Hatamizadeh, X. Feng, Q. Dou, N. Tustison, C. Meyer, N. A. Shah, S. Talbar, M.-A. Weber, A. Mahajan, A. Jakab, R. Wiest, H. M. Fathallah-Shaykh, A. Nazeri, M. Milchenko, D. Marcus, A. Kotrotsou, R. Colen, J. Freymann, J. Kirby, C. Davatzikos, B. Menze, S. Bakas, Y. Gal, and T. Arbel (Aug. 26, 2022). **QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation – Analysis of Ranking Scores and Benchmarking Results**. In: *Machine Learning for Biomedical Imaging* 1 (August 2022 issue), pp. 1–54. ISSN: 2766-905X. DOI: 10.59275/j.melba.2022-354b. URL: https://www.melba-journal.org/papers/2022:026.html (visited on 12/13/2024) (cit. on pp. 26, 27).

Menze, B. H., A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput (Oct. 2015). **The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)**. In: *IEEE Transactions on Medical Imaging* 34.10, pp. 1993–2024. ISSN: 1558-254X. DOI: 10.1109/TMI.2014.2377694 (cit. on pp. 1, 19, 32, 51, 74, 110, 112).

Merkel, D. (Mar. 1, 2014). **Docker: Lightweight Linux Containers for Consistent Development and Deployment**. In: *Linux Journal* 2014.239, 2:2. ISSN: 1075-3583 (cit. on p. 42).

Milesi, A., M. Futrega, M. Marcinkiewicz, and P. Ribalta (2021). **Brain Tumor Segmentation Using Neural Network Topology Search**. In: *International MICCAI Brainlesion Workshop*, pp. 366–376 (cit. on p. 190).

Miyato, T., S.-i. Maeda, M. Koyama, and S. Ishii (2018). **Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning**. In: *IEEE transactions on pattern analysis and machine intelligence* 41.8, pp. 1979–1993 (cit. on p. 189).

Moawad, A. W., A. Janas, U. Baid, D. Ramakrishnan, R. Saluja, N. Ashraf, N. Maleki, L. Jekel, N. Yordanov, P. Fehringer, A. Gkampenis, R. Amiruddin, A. Manteghinejad, M. Adewole, J. Albrecht, U. Anazodo, S. Aneja, S. M. Anwar, T. Bergquist, V. Chiang, V. Chung, G. M. Conte, F. Dako, J. Eddy, I. Ezhov, N. Khalili, K. Farahani, J. E. Iglesias,

Z. Jiang, E. Johanson, A. F. Kazerooni, F. Kofler, K. Krantchev, D. LaBella, K. V. Leemput, H. B. Li, M. G. Linguraru, X. Liu, Z. Meier, B. H. Menze, H. Moy, K. Osenberg, M. Piraud, Z. Reitman, R. T. Shinohara, C. Wang, B. Wiestler, W. Wiggins, U. Shafique, K. Willms, A. Avesta, K. Bousabarah, S. Chakrabarty, N. Gennaro, W. Holler, M. Kaur, P. LaMontagne, M. Lin, J. Lost, D. S. Marcus, R. Maresca, S. Merkaj, G. C. Pedersen, M. von Reppert, A. Sotiras, O. Teytelboym, N. Tillmans, M. Westerhoff, A. Youssef, D. Godfrey, S. Floyd, A. Rauschecker, J. Villanueva-Meyer, I. Pfluger, J. Cho, M. Bendszus, G. Brugnara, J. Cramer, G. J. G. Perez-Carillo, D. R. Johnson, A. Kam, B. Y. M. Kwan, L. Lai, N. U. Lall, F. Memon, M. Krycia, S. N. Patro, B. Petrovic, T. Y. So, G. Thompson, L. Wu, E. B. Schrickel, A. Bansal, F. Barkhof, C. Besada, S. Chu, J. Druzgal, A. Dusoi, L. Farage, F. Feltrin, A. Fong, S. H. Fung, R. I. Gray, I. Ikuta, M. Iv, A. A. Postma, A. Mahajan, D. Joyner, C. Krumpelman, L. Letourneau-Guillon, C. M. Lincoln, M. E. Maros, E. Miller, F. Moron, E. A. Nimchinsky, O. Ozsarlak, U. Patel, S. Rohatgi, A. Saha, A. Sayah, E. D. Schwartz, R. Shih, M. S. Shiroishi, J. E. Small, M. Tanwar, J. Valerie, B. D. Weinberg, M. L. White, R. Young, V. M. Zohrabian, A. Azizova, M. M. T. Bruseler, M. Ghonim, M. Ghonim, A. Okar, L. Pasquini, Y. Sharifi, G. Singh, N. Sollmann, T. Soumala, M. Taherzadeh, P. Vollmuth, M. Foltyn-Dumitru, A. Malhotra, A. H. Abayazeed, F. Dellepiane, P. Lohmann, V. M. Perez-Garcia, H. Elhalawani, M. C. de Verdier, S. Al-Rubaiey, R. D. Armindo, K. Ashraf, M. M. Asla, M. Badawy, J. Bisschop, N. B. Lomer, J. Bukatz, J. Chen, P. Cimflova, F. Corr, A. Crawley, L. Deptula, T. Elakhdar, I. H. Shawali, S. Faghani, A. Frick, V. Gulati, M. A. Haider, F. Hierro, R. H. Dahl, S. M. Jacobs, K.-c. J. Hsieh, S. G. Kandemirli, K. Kersting, L. Kida, S. Kollia, I. Koukoulithras, X. Li, A. Abouelatta, A. Mansour, R.-C. Maria-Zamfirescu, M. Marsiglia, Y. S. Mateo-Camacho, M. McArthur, O. McDonnell, M. McHugh, M. Moassefi, S. M. Morsi, A. Munteanu, K. K. Nandolia, S. R. Naqvi, Y. Nikanpour, M. Alnoury, A. M. A. Nouh, F. Pappafava, M. D. Patel, S. Petrucci, E. Rawie, S. Raymond, B. Roohani, S. Sabouhi, L. M. Sanchez-Garcia, Z. Shaked, P. P. Suthar, T. Altes, E. Isufi, Y. Dhemesh, J. Gass, J. Thacker, A. R. Tarabishy, B. Turner, S. Vacca, G. K. Vilanilam, D. Warren, D. Weiss, F. Worede, S. Yousry, W. Lerebo, A. Aristizabal, A. Karargyris, H. Kassem, S. Pati, M. Sheller, K. E. Link, E. Calabrese, N. hoda Tahon, A. Nada, Y. S. Velichko, S. Bakas, J. D. Rudie, and M. Aboian (2024). **The Brain Tumor Segmentation (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-Treatment MRI**. arXiv: 2306.00838 [q-bio.OT] (cit. on p. 3).

Monteiro, M., L. L. Folgoc, D. C. de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, M. van der Wilk, and B. Glocker (June 10, 2020). **Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty**. arXiv: 2006.06015 [cs]. URL: http://arxiv.org/abs/2006.06015 (visited on 10/07/2020) (cit. on p. 22).

Mora, A. M., M. Baumgartner, G. Brugnara, M. Zenk, Y. Kirchhoff, A. Rastogi, A. Radbruch, M. Bendszus, C. I. Sánchez, P. Vollmuth, and K. Maier-Hein (Apr. 27, 2024). **Curriculum-**

Learning for Vessel Occlusion Detection in Multi-site Brain CT Angiographies. In: Medical Imaging with Deep Learning. URL: https://openreview.net/forum?id=6TrjwzbBko (visited on 01/22/2025) (cit. on p. 173).

Moreno-Torres, J. G., T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera (Jan. 1, 2012). **A Unifying View on Dataset Shift in Classification**. In: *Pattern Recognition* 45.1, pp. 521–530. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2011.06.019. URL: https://www.sciencedirect.com/science/article/pii/S0031320311002901 (visited on 03/01/2021) (cit. on pp. 9, 10).

Morozov, S. P., A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Y. Chernina (May 13, 2020). **MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset**. DOI: 10.48550/arXiv.2005.06465. arXiv: 2005.06465 [cs, eess]. URL: http://arxiv.org/abs/2005.06465 (visited on 02/05/2024). Pre-published (cit. on p. 52).

Nair, T., D. Precup, D. L. Arnold, and T. Arbel (Jan. 1, 2020). **Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation**. In: *Medical Image Analysis* 59, p. 101557. ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101557. URL: https://www.sciencedirect.com/science/article/pii/S1361841519300994 (visited on 08/10/2022) (cit. on pp. 22, 23).

Nalawade, S., C. Ganesh, B. Wagner, D. Reddy, Y. Das, F. F. Yu, B. Fei, A. J. Madhuranthakam, and J. A. Maldjian (2021). **Federated Learning for Brain Tumor Segmentation Using MRI and Transformers**. In: *International MICCAI Brainlesion Workshop*, pp. 444–454 (cit. on p. 68).

Nelms, B. E., W. A. Tomé, G. Robinson, and J. Wheeler (2012). **Variations in the Contouring of Organs at Risk: Test Case from a Patient with Oropharyngeal Cancer**. In: *International Journal of Radiation Oncology* Biology* Physics* 82.1, pp. 368–378 (cit. on p. 1).

Ng, M., F. Guo, L. Biswas, S. E. Petersen, S. K. Piechnik, S. Neubauer, and G. Wright (June 2023). **Estimating Uncertainty in Neural Networks for Cardiac MRI Segmentation: A Benchmark Study**. In: *IEEE Transactions on Biomedical Engineering* 70.6, pp. 1955–1966. ISSN: 1558-2531. DOI: 10.1109/TBME.2022.3232730 (cit. on pp. 19, 26, 47, 117, 118, 120).

Nguyen-Truong, H. and Q.-D. Pham (2021). **Dice Focal Loss with ResNet-like Encoder-Decoder Architecture in 3D Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 97–105 (cit. on p. 190).

NHS, E. (Nov. 23, 2023). **Diagnostic Imaging Dataset Annual Statistical Release 2022/23**. NHS. URL: https://www.england.nhs.uk/statistics/statistical-work

-areas/diagnostic-imaging-dataset/diagnostic-imaging-dataset-2022-23-data/ (visited on 10/25/2024) (cit. on p. 1).

Nikolov, S., S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, and B. Romera-Paredes (2021). **Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study**. In: *Journal of medical Internet research* 23.7, e26151. URL: https://www.jmir.org/2021/7/e26151/span%5B (visited on 10/29/2024) (cit. on pp. 1, 4, 8, 48).

Ogier du Terrail, J., S.-S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, B. Muzellec, C. Philippenko, S. Silva, M. Teleńczuk, S. Albarqouni, S. Avestimehr, A. Bellet, A. Dieuleveut, M. Jaggi, S. P. Karimireddy, M. Lorenzi, G. Neglia, M. Tommasi, and M. Andreux (Dec. 6, 2022). **FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings**. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 5315–5334. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/232eee8ef411a0a316efa298d7be3c2b-Abstract-Datasets_and_Benchmarks.html (visited on 07/26/2024) (cit. on p. 19).

Ogier du Terrail, J., A. Leopold, C. Joly, C. Béguier, M. Andreux, C. Maussion, B. Schmauch, E. W. Tramel, E. Bendjebbar, M. Zaslavskiy, G. Wainrib, M. Milder, J. Gervasoni, J. Guerin, T. Durand, A. Livartowski, K. Moutet, C. Gautier, I. Djafar, A.-L. Moisson, C. Marini, M. Galtier, F. Balazard, R. Dubois, J. Moreira, A. Simon, D. Drubay, M. Lacroix-Triki, C. Franchet, G. Bataillon, and P.-E. Heudel (Jan. 2023). **Federated Learning for Predicting Histological Response to Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer**. In: *Nature Medicine* 29.1, pp. 135–146. ISSN: 1546-170X. DOI: 10.1038/s41591-022-02155-w. URL: https://www.nature.com/articles/s41591-022-02155-w (visited on 04/24/2024) (cit. on p. 20).

Orlando, J. I., H. Fu, J. Barbosa Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, J. Lee, X. Li, P. Liu, S. Lu, B. Murugesan, V. Naranjo, S. S. R. Phaye, S. M. Shankaranarayana, A. Sikka, J. Son, A. van den Hengel, S. Wang, J. Wu, Z. Wu, G. Xu, Y. Xu, P. Yin, F. Li, X. Zhang, Y. Xu, and H. Bogunović (Jan. 1, 2020). **REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs**. In: *Medical Image Analysis* 59, p. 101570. ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101570. URL: https://www.sciencedirect.com/science/article/pii/S1361841519301100 (visited on 09/25/2024) (cit. on p. 53).

Ouyang, C., C. Chen, S. Li, Z. Li, C. Qin, W. Bai, and D. Rueckert (Apr. 2023). **Causality-Inspired Single-Source Domain Generalization for Medical Image Segmentation**. In: *IEEE Transactions on Medical Imaging* 42.4, pp. 1095–1106. ISSN: 1558-254X. DOI: 10

.1109/TMI.2022.3224067. URL: https://ieeexplore.ieee.org/abstract/document/9961940 (visited on 07/26/2024) (cit. on p. 18).

Ovadia, Y., E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek (Dec. 17, 2019). **Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift**. arXiv: 1906.02530 [cs, stat]. URL: http://arxiv.org/abs/1906.02530 (visited on 01/12/2022) (cit. on p. 22).

Parampottupadam, S., R. Floca, D. Bounias, B. Hamm, S. Roy, S. Sav, M. Zenk, and K. Maier-Hein (2024). **Client Security Alone Fails in Federated Learning: 2D and 3D Attack Insights**. In: *MICCAI Student Board EMERGE Workshop: Empowering Medical Image Computing & Research through Early-Career Expertise* (cit. on p. 173).

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala (2019). **PyTorch: An Imperative Style, High-Performance Deep Learning Library**. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf (cit. on p. 58).

Pati, S., U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G. A. Reina, P. Foley, A. Gruzdev, D. Karkada, C. Davatzikos, C. Sako, S. Ghodasara, M. Bilello, S. Mohan, P. Vollmuth, G. Brugnara, C. J. Preetha, F. Sahm, K. Maier-Hein, M. Zenk, M. Bendszus, W. Wick, E. Calabrese, J. Rudie, J. Villanueva-Meyer, S. Cha, M. Ingalhalikar, M. Jadhav, U. Pandey, J. Saini, J. Garrett, M. Larson, R. Jeraj, S. Currie, R. Frood, K. Fatania, R. Y. Huang, K. Chang, C. B. Quintero, J. Capellades, J. Puig, J. Trenkler, J. Pichler, G. Necker, A. Haunschmidt, S. Meckel, G. Shukla, S. Liem, G. S. Alexander, J. Lombardo, J. D. Palmer, A. E. Flanders, A. P. Dicker, H. I. Sair, C. K. Jones, A. Venkataraman, M. Jiang, T. Y. So, C. Chen, P. A. Heng, Q. Dou, M. Kozubek, F. Lux, J. Michálek, P. Matula, M. Keřkovský, T. Kopřivová, M. Dostál, V. Vybíhal, M. A. Vogelbaum, J. R. Mitchell, J. Farinhas, J. A. Maldjian, C. G. B. Yogananda, M. C. Pinho, D. Reddy, J. Holcomb, B. C. Wagner, B. M. Ellingson, T. F. Cloughesy, C. Raymond, T. Oughourlian, A. Hagiwara, C. Wang, M.-S. To, S. Bhardwaj, C. Chong, M. Agzarian, A. X. Falcão, S. B. Martins, B. C. A. Teixeira, F. Sprenger, D. Menotti, D. R. Lucio, P. LaMontagne, D. Marcus, B. Wiestler, F. Kofler, I. Ezhov, M. Metz, R. Jain, M. Lee, Y. W. Lui, R. McKinley, J. Slotboom, P. Radojewski, R. Meier, R. Wiest, D. Murcia, E. Fu, R. Haas, J. Thompson, D. R. Ormond, C. Badve, A. E. Sloan, V. Vadmal, K. Waite, R. R. Colen, L. Pei, M. Ak, A. Srinivasan, J. R. Bapuraj, A. Rao, N. Wang, O. Yoshiaki, T. Moritani, S. Turk, J. Lee, S. Prabhudesai, F. Morón, J. Mandel, K. Kamnitsas, B. Glocker, L. V. M. Dixon, M. Williams, P. Zampakis, V.

Panagiotopoulos, P. Tsiganos, S. Alexiou, I. Haliassos, E. I. Zacharaki, K. Moustakas, C. Kalogeropoulou, D. M. Kardamakis, Y. S. Choi, S.-K. Lee, J. H. Chang, S. S. Ahn, B. Luo, L. Poisson, N. Wen, P. Tiwari, R. Verma, R. Bareja, I. Yadav, J. Chen, N. Kumar, M. Smits, S. R. van der Voort, A. Alafandi, F. Incekara, M. M. J. Wijnenga, G. Kapsas, R. Gahrmann, J. W. Schouten, H. J. Dubbink, A. J. P. E. Vincent, M. J. van den Bent, P. J. French, S. Klein, Y. Yuan, S. Sharma, T.-C. Tseng, S. Adabi, S. P. Niclou, O. Keunen, A.-C. Hau, M. Vallières, D. Fortin, M. Lepage, B. Landman, K. Ramadass, K. Xu, S. Chotai, L. B. Chambless, A. Mistry, R. C. Thompson, Y. Gusev, K. Bhuvaneshwar, A. Sayah, C. Bencheqroun, A. Belouali, S. Madhavan, T. C. Booth, A. Chelliah, M. Modat, H. Shuaib, C. Dragos, A. Abayazeed, K. Kolodziej, M. Hill, A. Abbassy, S. Gamal, M. Mekhaimar, M. Qayati, M. Reyes, J. E. Park, J. Yun, H. S. Kim, A. Mahajan, M. Muzi, S. Benson, R. G. H. Beets-Tan, J. Teuwen, A. Herrera-Trujillo, M. Trujillo, W. Escobar, A. Abello, J. Bernal, J. Gómez, J. Choi, S. Baek, Y. Kim, H. Ismael, B. Allen, J. M. Buatti, A. Kotrotsou, H. Li, T. Weiss, M. Weller, A. Bink, B. Pouymayou, H. F. Shaykh, J. Saltz, P. Prasanna, S. Shrestha, K. M. Mani, D. Payne, T. Kurc, E. Pelaez, H. Franco-Maldonado, F. Loayza, S. Quevedo, P. Guevara, E. Torche, C. Mendoza, F. Vera, E. Ríos, E. López, S. A. Velastin, G. Ogbole, M. Soneye, D. Oyekunle, O. Odafe-Oyibotha, B. Osobu, M. Shu'aibu, A. Dorcas, F. Dako, A. L. Simpson, M. Hamghalam, J. J. Peoples, R. Hu, A. Tran, D. Cutler, F. Y. Moraes, M. A. Boss, J. Gimpel, D. K. Veettil, K. Schmidt, B. Bialecki, S. Marella, C. Price, L. Cimino, C. Apgar, P. Shah, B. Menze, J. S. Barnholtz-Sloan, J. Martin, and S. Bakas (Dec. 5, 2022a). **Federated Learning Enables Big Data for Rare Cancer Boundary Detection**. In: *Nature Communications* 13.1 (1), p. 7346. ISSN: 2041-1723. DOI: 10.1038/s41467-022-33407-5. URL: https://www.nature.com/articles/s41467-022-33407-5 (visited on 12/14/2022) (cit. on pp. 20, 32, 38, 39, 42, 115, 172).

Pati, S., U. Baid, B. Edwards, M. J. Sheller, P. Foley, G. Anthony Reina, S. Thakur, C. Sako, M. Bilello, C. Davatzikos, J. Martin, P. Shah, B. Menze, and S. Bakas (Oct. 12, 2022b). **The Federated Tumor Segmentation (FeTS) Tool: An Open-Source Solution to Further Solid Tumor Research**. In: *Physics in Medicine and Biology* 67.20. ISSN: 1361-6560. DOI: 10.1088/1361-6560/ac9449. pmid: 36137534 (cit. on p. 37).

Pati, S., U. Baid, M. Zenk, B. Edwards, M. Sheller, G. A. Reina, P. Foley, A. Gruzdev, J. Martin, S. Albarqouni, Y. Chen, R. T. Shinohara, A. Reinke, D. Zimmerer, J. B. Freymann, J. S. Kirby, C. Davatzikos, R. R. Colen, A. Kotrotsou, D. Marcus, M. Milchenko, A. Nazeri, H. Fathallah-Shaykh, R. Wiest, A. Jakab, M.-A. Weber, A. Mahajan, L. Maier-Hein, J. Kleesiek, B. Menze, K. Maier-Hein, and S. Bakas (May 13, 2021). **The Federated Tumor Segmentation (FeTS) Challenge**. arXiv: 2105.05874 [cs, eess]. URL: http://arxiv.org/abs/2105.05874 (visited on 09/28/2021) (cit. on pp. 31, 171).

Paulik, M., M. Seigel, H. Mason, D. Telaar, J. Kluivers, R. van Dalen, C. W. Lau, L. Carlson, F. Granqvist, C. Vandevelde, S. Agarwal, J. Freudiger, A. Byde, A. Bhowmick, G. Kapoor, S. Beaumont, Á. Cahill, D. Hughes, O. Javidbakht, F. Dong, R. Rishi, and S. Hung (Feb. 16, 2021). **Federated Evaluation and Tuning for On-Device Personalization: System Design & Applications**. arXiv: 2102.08503 [cs]. URL: http://arxiv.org/abs/2102.08503 (visited on 12/08/2021) (cit. on p. 20).

Pawar, K., S. Zhong, Z. Chen, and G. Egan (2021a). **Brain Tumor Segmentation Using Two-Stage Convolutional Neural Network for Federated Evaluation**. In: *International MICCAI Brainlesion Workshop*, pp. 494–505 (cit. on p. 69).

Pawar, K., S. Zhong, D. S. Goonatillake, G. Egan, and Z. Chen (2021b). **Orthogonal-Nets: A Large Ensemble of 2D Neural Networks for 3D Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 54–67 (cit. on p. 190).

Payette, K., C. Steger, R. Licandro, P. de Dumast, H. B. Li, M. Barkovich, L. Li, M. Dannecker, C. Chen, C. Ouyang, N. McConnell, A. Miron, Y. Li, A. Uus, I. Grigorescu, P. R. Gilliland, M. M. R. Siddiquee, D. Xu, A. Myronenko, H. Wang, Z. Huang, J. Ye, M. Alenyà, V. Comte, O. Camara, J.-B. Masson, A. Nilsson, C. Godard, M. Mazher, A. Qayyum, Y. Gao, H. Zhou, S. Gao, J. Fu, G. Dong, G. Wang, Z. Rieu, H. Yang, M. Lee, S. Płotka, M. K. Grzeszczyk, A. Sitek, L. V. Daza, S. Usma, P. Arbelaez, W. Lu, W. Zhang, J. Liang, R. Valabregue, A. A. Joshi, K. N. Nayak, R. M. Leahy, L. Wilhelmi, A. Dändliker, H. Ji, A. G. Gennari, A. Jakovčić, M. Klaić, A. Adžić, P. Marković, G. Grabarić, G. Kasprian, G. Dovjak, M. Rados, L. Vasung, M. B. Cuadra, and A. Jakab (Feb. 8, 2024). **Multi-Center Fetal Brain Tissue Annotation (FeTA) Challenge 2022 Results**. DOI: 10.48550/arXiv.2402.09463. arXiv: 2402.09463 [eess]. URL: http://arxiv.org/abs/2402.09463 (visited on 07/26/2024). Pre-published (cit. on pp. 2, 19, 20, 27, 113).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). **Scikit-Learn: Machine Learning in Python**. In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 04/10/2024) (cit. on pp. 13, 62).

Peiris, H., M. Hayat, Z. Chen, G. Egan, and M. Harandi (2022). **Hybrid Window Attention Based Transformer Architecture for Brain Tumor Segmentation**. eprint: 2209.07704 (cit. on p. 73).

Pérez-García, F., R. Sparks, and S. Ourselin (Sept. 1, 2021). **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* 208, p. 106236. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2021.106236. URL: https:

//www.sciencedirect.com/science/article/pii/S0169260721003102 (visited on 02/05/2024) (cit. on pp. 51, 57).

Pnev, S., V. Groza, B. Tuchinov, E. Amelina, E. Pavlovskiy, N. Tolstokulakov, M. Amelin, S. Golushko, and A. Letyagin (2021). **Brain Tumor Segmentation with Self-supervised Enhance Region Post-processing**. In: *International MICCAI Brainlesion Workshop*, pp. 267–275 (cit. on p. 190).

Qiu, P., S. Chakrabarty, P. Nguyen, S. S. Ghosh, and A. Sotiras (2023). **QCResUNet: Joint Subject-Level and Voxel-Level Prediction of Segmentation Quality**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor. Vol. 14223. Cham: Springer Nature Switzerland, pp. 173–182. ISBN: 978-3-031-43901-8. DOI: 10.1007/978-3-031-43901-8_17. URL: https://link.springer.com/10.1007/978-3-031-43901-8_17 (visited on 10/18/2023) (cit. on pp. 23, 46, 47, 119).

Quiñonero-Candela, J., M. Sugiyama, A. Schwaighofer, and N. D. Lawrence (2022). **Dataset Shift in Machine Learning**. Mit Press (cit. on p. 9).

RCR (2023). **Rcr-Census-Clinical-Radiology-Workforce-Census-2023**. The Royal College of Radiologists. URL: https://www.rcr.ac.uk/media/5befglss/rcr-census-clinical-radiology-workforce-census-2023.pdf (visited on 10/22/2024) (cit. on p. 1).

Reinke, A., M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina, K. Farahani, B. Glocker, D. Heckmann-Nötzel, F. Isensee, P. Jannin, C. E. Kahn, J. Kleesiek, T. Kurc, M. Kozubek, B. A. Landman, G. Litjens, K. Maier-Hein, B. Menze, H. Müller, J. Petersen, M. Reyes, N. Rieke, B. Stieltjes, R. M. Summers, S. A. Tsaftaris, B. van Ginneken, A. Kopp-Schneider, P. Jäger, and L. Maier-Hein (Apr. 13, 2021). **Common Limitations of Image Processing Metrics: A Picture Story**. arXiv: 2104.05642 [cs, eess]. URL: http://arxiv.org/abs/2104.05642 (visited on 04/21/2021) (cit. on pp. 8, 78).

Ren, J., W. Zhang, N. An, Q. Hu, Y. Zhang, and Y. Zhou (2021). **Ensemble Outperforms Single Models in Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 451–462 (cit. on p. 73).

Robinson, R., O. Oktay, W. Bai, V. Valindria, M. Sanghvi, N. Aung, J. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. Lee, V. Carapella, Y. J. Kim, B. Kainz, S. Piechnik, S. Neubauer, S. Petersen, C. Page, D. Rueckert, and B. Glocker (June 16, 2018). **Real-Time Prediction of Segmentation Quality**. DOI: 10.48550/arXiv.1806.06244. arXiv: 1806.06244 [cs]. URL: http://arxiv.org/abs/1806.06244 (visited on 09/05/2022). Prepublished (cit. on pp. 23, 46, 47, 63, 119, 120).

Rohlfing, T., N. M. Zahr, E. V. Sullivan, and A. Pfefferbaum (May 2010). **The SRI24 Multichannel Atlas of Normal Adult Human Brain Structure**. In: *Human Brain Mapping* 31.5, pp. 798–819. ISSN: 1097-0193. DOI: 10.1002/hbm.20906. pmid: 20017133 (cit. on p. 37).

Rokuss, M., Y. Kirchhoff, S. Roy, B. Kovacs, C. Ulrich, T. Wald, M. Zenk, S. Denner, F. Isensee, P. Vollmuth, J. Kleesiek, and K. Maier-Hein (Sept. 20, 2024). **Longitudinal Segmentation of MS Lesions via Temporal Difference Weighting**. DOI: 10.48550/arXiv.2409.13416. arXiv: 2409.13416 [eess]. URL: http://arxiv.org/abs/2409.13416 (visited on 01/22/2025). Pre-published (cit. on p. 173).

Ronneberger, O., P. Fischer, and T. Brox (2015). **U-Net: Convolutional Networks for Biomedical Image Segmentation**. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241 (cit. on pp. 6, 7, 56, 73, 189).

Roß, T., P. Bruno, A. Reinke, M. Wiesenfarth, L. Koeppel, P. M. Full, B. Pekdemir, P. Godau, D. Trofimova, F. Isensee, S. Moccia, F. Calimeri, B. P. Müller-Stich, A. Kopp-Schneider, and L. Maier-Hein (June 17, 2021). **How Can We Learn (More) from Challenges? A Statistical Approach to Driving Future Algorithm Development**. arXiv: 2106.09302 [cs]. URL: http://arxiv.org/abs/2106.09302 (visited on 06/22/2021) (cit. on p. 111).

Roth, H. R., Z. Xu, C. Tor-Díez, R. Sanchez Jacob, J. Zember, J. Molto, W. Li, S. Xu, B. Turkbey, E. Turkbey, D. Yang, A. Harouni, N. Rieke, S. Hu, F. Isensee, C. Tang, Q. Yu, J. Sölter, T. Zheng, V. Liauchuk, Z. Zhou, J. H. Moltz, B. Oliveira, Y. Xia, K. H. Maier-Hein, Q. Li, A. Husch, L. Zhang, V. Kovalev, L. Kang, A. Hering, J. L. Vilaça, M. Flores, D. Xu, B. Wood, and M. G. Linguraru (Nov. 1, 2022). **Rapid Artificial Intelligence Solutions in a Pandemic—The COVID-19-20 Lung CT Lesion Segmentation Challenge**. In: *Medical Image Analysis* 82, p. 102605. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102605. URL: https://www.sciencedirect.com/science/article/pii/S1361841522002353 (visited on 02/05/2024) (cit. on p. 52).

Roth, J., J. Keller, S. Franke, T. Neumuth, and D. Schneider (2021). **Multi-Plane UNet++ Ensemble for Glioblastoma Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 285–294 (cit. on p. 190).

Roy, A. G., S. Conjeti, N. Navab, and C. Wachinger (July 15, 2019). **Bayesian QuickNAT: Model Uncertainty in Deep Whole-Brain Segmentation for Structure-Wise Quality Control**. In: *NeuroImage* 195, pp. 11–22. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2019.03.042. URL: https://www.sciencedirect.com/science/article/pii/S1053811919302319 (visited on 11/10/2023) (cit. on pp. 22, 24, 63, 118).

Salehi, M., H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou (July 11, 2022). **A Unified Survey on Anomaly, Novelty, Open-Set, and Out of-Distribution**

Detection: Solutions and Future Challenges. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=aRtjVZvbpK&referrer=%5BTMLR%5D(%2Fgroup%3Fid%3DTMLR) (visited on 02/15/2024) (cit. on p. 25).

Saueressig, C., A. Berkley, R. Munbodh, and R. Singh (2021). **A Joint Graph and Image Convolution Network for Automatic Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 356–365 (cit. on p. 190).

Sensoy, M., L. Kaplan, and M. Kandemir (2018). **Evidential Deep Learning to Quantify Classification Uncertainty**. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html (visited on 09/24/2024) (cit. on p. 22).

Shah, D., A. Biswas, P. Sonpatki, S. Chakravarty, and N. Shah (2021). **Neural Network Based Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 324–333 (cit. on p. 190).

Shi, Y., H. Gao, S. Avestimehr, and Y. Yan (2022). **Experimenting FedML and NVFLARE for Federated Tumor Segmentation Challenge**. In: *International MICCAI Brainlesion Workshop*, pp. 228–240 (cit. on p. 73).

Simpson, A. L., M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. Gollub, J. Golia-Pernicka, S. H. Heckers, W. R. Jarnagin, M. K. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso (Feb. 24, 2019). **A Large Annotated Medical Image Dataset for the Development and Evaluation of Segmentation Algorithms**. arXiv: 1902.09063 [cs, eess]. URL: http://arxiv.org/abs/1902.09063 (visited on 12/04/2020) (cit. on p. 52).

Singh, H. S. (2021). **Brain Tumor Segmentation Using Attention Activated U-Net with Positive Mining**. In: *International MICCAI Brainlesion Workshop*, pp. 431–440 (cit. on p. 190).

Sivaswamy, J., S. Krishnadas, A. Chakravarty, G. Joshi, and A. S. Tabish (2015). **A Comprehensive Retinal Image Dataset for the Assessment of Glaucoma from the Optic Nerve Head Analysis**. In: *JSM Biomedical Imaging Data Papers* 2.1 (cit. on p. 53).

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). **Dropout: A Simple Way to Prevent Neural Networks from Overfitting**. In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958. ISSN: 1533-7928. URL: http://jmlr.org/papers/v15/srivastava14a.html (visited on 12/12/2024) (cit. on p. 22).

Sun, K., Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang (2019). **High-Resolution Representations for Labeling Pixels and Regions**. arXiv: 1904.04514 [cs.CV] (cit. on p. 189).

Thakur, S., J. Doshi, S. Pati, S. Rathore, C. Sako, M. Bilello, S. M. Ha, G. Shukla, A. Flanders, A. Kotrotsou, M. Milchenko, S. Liem, G. S. Alexander, J. Lombardo, J. D. Palmer, P. LaMontagne, A. Nazeri, S. Talbar, U. Kulkarni, D. Marcus, R. Colen, C. Davatzikos, G. Erus, and S. Bakas (2020). **Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-institutional Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training**. In: *NeuroImage* 220, p. 117081. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2020.117081. URL: https://www.sciencedirect.com/science/article/pii/S105381192030567X (cit. on p. 37).

Toennies, K. D. (2017). **Guide to Medical Image Analysis**. Springer (cit. on p. 4).

Traub, J., T. J. Bungert, C. T. Lüth, M. Baumgartner, K. H. Maier-Hein, L. Maier-Hein, and P. F. Jaeger (Oct. 19, 2024). **Overcoming Common Flaws in the Evaluation of Selective Classification Systems**. DOI: 10.48550/arXiv.2407.01032. arXiv: 2407.01032 [cs]. URL: http://arxiv.org/abs/2407.01032 (visited on 01/29/2025). Pre-published (cit. on p. 117).

Ulrich, C., F. Isensee, T. Wald, M. Zenk, M. Baumgartner, and K. H. Maier-Hein (2023). **MultiTalent: A Multi-dataset Approach to Medical Image Segmentation**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor. Cham: Springer Nature Switzerland, pp. 648–658. ISBN: 978-3-031-43898-1. DOI: 10.1007/978-3-031-43898-1_62 (cit. on p. 172).

Ulrich, C., C. Knobloch, J. C. Holzschuh, T. Wald, M. R. Rokuss, M. Zenk, M. Fischer, M. Baumgartner, F. Isensee, and K. H. Maier-Hein (2025). **Mitigating False Predictions in Unreasonable Body Regions**. In: *Machine Learning in Medical Imaging*. Ed. by X. Xu, Z. Cui, I. Rekik, X. Ouyang, and K. Sun. Cham: Springer Nature Switzerland, pp. 22–31. ISBN: 978-3-031-73290-4. DOI: 10.1007/978-3-031-73290-4_3 (cit. on p. 173).

Ulyanov, D., A. Vedaldi, and V. Lempitsky (Nov. 6, 2017). **Instance Normalization: The Missing Ingredient for Fast Stylization**. arXiv: 1607.08022 [cs]. URL: http://arxiv.org/abs/1607.08022 (visited on 11/25/2019) (cit. on p. 56).

Valindria, V. V., I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker (Aug. 2017). **Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth**. In: *IEEE Transactions on Medical Imaging* 36.8, pp. 1597–1606. ISSN: 1558-254X. DOI: 10.1109/TMI.2017.2665165 (cit. on pp. 3, 23, 24, 46).

Vasiliuk, A., D. Frolova, M. Belyaev, and B. Shirokikh (Aug. 7, 2023). **Redesigning Out-of-Distribution Detection on 3D Medical Images**. DOI: 10.48550/arXiv.2308.07324. arXiv: 2308.07324 [cs, eess]. URL: http://arxiv.org/abs/2308.07324 (visited on 10/18/2023). Pre-published (cit. on p. 26).

Wang, D., E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell (Sept. 28, 2020a). **Tent: Fully Test-Time Adaptation by Entropy Minimization**. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=uXl3bZLkr3c (visited on 04/23/2021) (cit. on p. 69).

Wang, G., W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren (Apr. 21, 2019a). **Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks**. In: *Neurocomputing* 338, pp. 34–45. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2019.01.103. URL: https://www.sciencedirect.com/science/article/pii/S0925231219301961 (visited on 01/11/2023) (cit. on p. 22).

Wang, J., K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao (Oct. 2021). **Deep High-Resolution Representation Learning for Visual Recognition**. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10, pp. 3349–3364. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2020.2983686. URL: https://ieeexplore.ieee.org/abstract/document/9052469 (visited on 10/30/2024) (cit. on p. 73).

Wang, K., R. Mathews, C. Kiddon, H. Eichner, F. Beaufays, and D. Ramage (Oct. 22, 2019b). **Federated Evaluation of On-device Personalization**. arXiv: 1910.10252 [cs, stat]. URL: http://arxiv.org/abs/1910.10252 (visited on 05/07/2020) (cit. on p. 20).

Wang, S., L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng (Dec. 2020b). **DoFE: Domain-Oriented Feature Embedding for Generalizable Fundus Image Segmentation on Unseen Datasets**. In: *IEEE Transactions on Medical Imaging* 39.12, pp. 4237–4248. ISSN: 1558-254X. DOI: 10.1109/TMI.2020.3015224. URL: https://ieeexplore.ieee.org/document/9163289 (visited on 09/25/2024) (cit. on pp. 19, 53).

Wang, S., G. Tarroni, C. Qin, Y. Mo, C. Dai, C. Chen, B. Glocker, Y. Guo, D. Rueckert, and W. Bai (2020c). **Deep Generative Model-based Quality Control for Cardiac MRI Segmentation**. In: vol. 12264, pp. 88–97. DOI: 10.1007/978-3-030-59719-1_9. arXiv: 2006.13379 [cs, eess]. URL: http://arxiv.org/abs/2006.13379 (visited on 06/06/2023) (cit. on pp. 24, 120).

Warfield, S. K., K. H. Zou, and W. M. Wells (July 2004). **Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation**. In: *IEEE transactions on medical imaging* 23.7, pp. 903–921. ISSN: 0278-0062. DOI: 10.1109/TMI.2004.828354. pmid: 15250643 (cit. on p. 38).

Wiesenfarth, M., A. Reinke, B. A. Landman, M. Eisenmann, L. A. Saiz, M. J. Cardoso, L. Maier-Hein, and A. Kopp-Schneider (Jan. 27, 2021). **Methods and Open-Source Toolkit for Analyzing and Visualizing Challenge Results**. In: *Scientific Reports* 11.1 (1), p. 2369. ISSN: 2045-2322. DOI: 10.1038/s41598-021-82017-6. URL: https:

//www.nature.com/articles/s41598-021-82017-6 (visited on 03/25/2021) (cit. on pp. 16, 40, 41).

Wu, H.-Y. and Y.-L. Lin (2021). **HarDNet-BTS: A Harmonic Shortcut Network for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 261–271 (cit. on p. 190).

Xia, Y., Y. Zhang, F. Liu, W. Shen, and A. Yuille (Sept. 7, 2020). **Synthesize Then Compare: Detecting Failures and Anomalies for Semantic Segmentation**. arXiv: 2003.08440 [cs]. URL: http://arxiv.org/abs/2003.08440 (visited on 02/14/2022) (cit. on p. 24).

Xie, Y., J. Zhang, C. Shen, and Y. Xia (2021). **CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert. Cham: Springer International Publishing, pp. 171–180. ISBN: 978-3-030-87199-4. DOI: 10.1007/978-3-030-87199-4_16 (cit. on pp. 73, 189).

Yan, B. B., Y. Wei, J. M. M. Jagtap, M. Moassefi, D. V. V. Garcia, Y. Singh, S. Vahdati, S. Faghani, B. J. Erickson, and G. M. Conte (2021). **Mri Brain Tumor Segmentation Using Deep Encoder-Decoder Convolutional Neural Networks**. In: *International MICCAI Brainlesion Workshop*, pp. 80–89 (cit. on p. 190).

Yang, H., Z. Shen, Z. Li, J. Liu, and J. Xiao (2021a). **Combining Global Information with Topological Prior for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 204–215 (cit. on p. 190).

Yang, Y., S. Wei, D. Zhang, Q. Yan, S. Zhao, and J. Han (2021b). **Hierarchical and Global Modality Interaction for Brain Tumor Segmentation**. In: *International MICCAI Brainlesion Workshop*, pp. 441–450 (cit. on p. 190).

El-Yaniv, R. and Y. Wiener (Aug. 1, 2010). **On the Foundations of Noise-free Selective Classification**. In: *The Journal of Machine Learning Research* 11, pp. 1605–1641. ISSN: 1532-4435 (cit. on pp. 48, 117).

Yin, Y., H. Yang, Q. Liu, M. Jiang, C. Chen, Q. Dou, and P.-A. Heng (2021). **Efficient Federated Tumor Segmentation via Normalized Tensor Aggregation and Client Pruning**. In: *International MICCAI Brainlesion Workshop*, pp. 433–443 (cit. on p. 69).

Yoon, J. S., K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk (Feb. 15, 2024). **Domain Generalization for Medical Image Analysis: A Survey**. DOI: 10.48550/arXiv.2310.08598. arXiv: 2310.08598. URL: http://arxiv.org/abs/2310.08598 (visited on 10/23/2024). Pre-published (cit. on pp. 2, 18).

Yuan, Y. (2021). **Evaluating Scale Attention Network for Automatic Brain Tumor Segmentation with Large Multi-Parametric MRI Database**. In: *International MICCAI Brainlesion Workshop*, pp. 42–53 (cit. on p. 190).

Yushkevich, P. A., J. Pluta, H. Wang, L. E. Wisse, S. Das, and D. Wolk (2016). **Fast Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe Subregions in 3 Tesla and 7 Tesla T2-weighted MRI**. In: *Alzheimer's & Dementia* 7.12 (cit. on p. 37).

Zech, J. R., M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann (Nov. 6, 2018). **Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study**. In: *PLOS Medicine* 15.11, e1002683. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002683. URL: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683 (visited on 12/03/2020) (cit. on p. 2).

Zeineldin, R. A., M. E. Karar, F. Mathis-Ullrich, and O. Burgert (2021). **Ensemble CNN Networks for GBM Tumors Segmentation Using Multi-Parametric MRI**. In: *International MICCAI Brainlesion Workshop*, pp. 473–483 (cit. on p. 190).

Zenk, M., U. Baid, S. Pati, A. Linardos, B. Edwards, M. Sheller, P. Foley, A. Aristizabal, D. Zimmerer, A. Gruzdev, J. Martin, R. T. Shinohara, A. Reinke, F. Isensee, S. Parampottupadam, K. Parekh, R. Floca, H. Kassem, B. Baheti, S. Thakur, V. Chung, K. Kushibar, K. Lekadir, M. Jiang, Y. Yin, H. Yang, Q. Liu, C. Chen, Q. Dou, P.-A. Heng, X. Zhang, S. Zhang, M. I. Khan, M. A. Azeem, M. Jafaritadi, E. Alhoniemi, E. Kontio, S. A. Khan, L. Mächler, I. Ezhov, F. Kofler, S. Shit, J. C. Paetzold, T. Loehr, B. Wiestler, H. Peiris, K. Pawar, S. Zhong, Z. Chen, M. Hayat, G. Egan, M. Harandi, E. I. Polat, G. Polat, A. Kocyigit, A. Temizel, A. Tuladhar, L. Tyagi, R. Souza, N. D. Forkert, P. Mouches, M. Wilms, V. Shambhat, A. Maurya, S. S. Danannavar, R. Kalla, V. K. Anand, G. Krishnamurthi, S. Nalawade, C. Ganesh, B. Wagner, D. Reddy, Y. Das, F. F. Yu, B. Fei, A. J. Madhuranthakam, J. Maldjian, G. Singh, J. Ren, W. Zhang, N. An, Q. Hu, Y. Zhang, Y. Zhou, V. Siomos, A. Rawat, G. Zizzo, S. Kadhe, J. P. Epperlein, S. Braghin, Y. Wang, R. Kanagavelu, Q. Wei, Y. Yang, Y. Liu, K. Kotowski, S. Adamski, B. Machura, W. Malara, L. Zarudzki, J. Nalepa, Y. Shi, H. Gao, S. Avestimehr, Y. Yan, A. S. Akbar, E. Kondrateva, H. Yang, Z. Li, H.-Y. Wu, J. Roth, C. Saueressig, A. Milesi, Q. D. Nguyen, N. J. Gruenhagen, T.-M. Huang, J. Ma, H. S. H. Singh, N.-Y. Pan, D. Zhang, R. A. Zeineldin, M. Futrega, Y. Yuan, G. M. Conte, X. Feng, Q. D. Pham, Y. Xia, Z. Jiang, H. M. Luu, M. Dobko, A. Carré, B. Tuchinov, H. Mohy-ud-Din, S. Alam, A. Singh, N. Shah, W. Wang, C. Sako, M. Bilello, S. Ghodasara, S. Mohan, C. Davatzikos, E. Calabrese, J. Rudie, J. Villanueva-Meyer, S. Cha, C. Hess, J. Mongan, M. Ingalhalikar, M. Jadhav, U. Pandey, J. Saini, R. Y. Huang, K. Chang, M.-S. To, S. Bhardwaj, C. Chong, M. Agzarian, M. Kozubek, F. Lux, J. Michálek, P. Matula, M. Keřkovský, T. Kopřivová, M. Dostál, V. Vybíhal, M. C. Pinho, J. Holcomb, M. Metz, R. Jain, M. Lee, Y. W. Lui, P. Tiwari, R. Verma, R. Bareja, I. Yadav, J. Chen, N. Kumar, Y. Gusev, K. Bhuvaneshwar, A. Sayah, C. Bencheqroun, A. Belouali, S. Madhavan, R. R. Colen, A. Kotrotsou, P. Vollmuth, G. Brugnara, C. J. Preetha, F. Sahm, M. Bendszus, W. Wick, A. Mahajan, C. Balaña Quintero, J. Capellades, J. Puig,

Y. S. Choi, S.-K. Lee, J. H. Chang, S. S. Ahn, H. F. Shaykh, A. Herrera-Trujillo, M. Trujillo, W. Escobar, A. Abello, J. Bernal, J. Gómez, P. LaMontagne, D. Marcus, M. Milchenko, A. Nazeri, B. Landman, K. Ramadass, K. Xu, S. Chotai, L. B. Chambless, A. Mistry, R. C. Thompson, A. Srinivasan, J. R. Bapuraj, A. Rao, N. Wang, O. Yoshiaki, T. Moritani, S. Turk, J. Lee, S. Prabhudesai, J. Garrett, M. Larson, R. Jeraj, H. Li, T. Weiss, M. Weller, A. Bink, B. Pouymayou, S. Sharma, T.-C. Tseng, S. Adabi, A. X. Falcão, S. B. Martins, B. C. A. Teixeira, F. Sprenger, D. Menotti, D. R. Lucio, S. P. Niclou, O. Keunen, A.-C. Hau, E. Pelaez, H. Franco-Maldonado, F. Loayza, S. Quevedo, R. McKinley, J. Slotboom, P. Radojewski, R. Meier, R. Wiest, J. Trenkler, J. Pichler, G. Necker, A. Haunschmidt, S. Meckel, P. Guevara, E. Torche, C. Mendoza, F. Vera, E. Ríos, E. López, S. A. Velastin, J. Choi, S. Baek, Y. Kim, H. Ismael, B. Allen, J. M. Buatti, P. Zampakis, V. Panagiotopoulos, P. Tsiganos, S. Alexiou, I. Haliassos, E. I. Zacharaki, K. Moustakas, C. Kalogeropoulou, D. M. Kardamakis, B. Luo, L. Poisson, N. Wen, M. Vallières, M. A. L. Loutfi, D. Fortin, M. Lepage, F. Morón, J. Mandel, G. Shukla, S. Liem, G. S. Alexandre, J. Lombardo, J. D. Palmer, A. E. Flanders, A. P. Dicker, G. Ogbole, D. Oyekunle, O. Odafe-Oyibotha, B. Osobu, M. S. Hikima, M. Soneye, F. Dako, A. Dorcas, D. Murcia, E. Fu, R. Haas, J. Thompson, D. R. Ormond, S. Currie, K. Fatania, R. Frood, A. L. Simpson, J. J. Peoples, R. Hu, D. Cutler, F. Y. Moraes, A. Tran, M. Hamghalam, M. A. Boss, J. Gimpel, D. Kattil Veettil, K. Schmidt, L. Cimino, C. Price, B. Bialecki, S. Marella, C. Apgar, A. Jakab, M.-A. Weber, E. Colak, J. Kleesiek, J. B. Freymann, J. S. Kirby, L. Maier-Hein, J. Albrecht, P. Mattson, A. Karargyris, P. Shah, B. Menze, K. Maier-Hein, and S. Bakas (2025a). **Towards Fair Decentralized Benchmarking of Healthcare AI Algorithms: The Federated Tumor Segmentation (FeTS) Challenge**. Manuscript accepted for publication. (cit. on pp. 5, 17, 20, 31, 35, 36, 51, 67, 68, 71, 72, 76, 77, 80, 81, 83, 109, 171, 176, 177, 189, 190).

Zenk, M., D. Zimmerer, F. Isensee, P. F. Jäger, J. Wasserthal, and K. Maier-Hein (2022). **Realistic Evaluation of FixMatch on Imbalanced Medical Image Classification Tasks**. In: *Bildverarbeitung Für Die Medizin 2022*. Ed. by K. Maier-Hein, T. M. Deserno, H. Handels, A. Maier, C. Palm, and T. Tolxdorff. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 291–296. ISBN: 978-3-658-36932-3 (cit. on p. 171).

Zenk, M., D. Zimmerer, F. Isensee, J. Traub, T. Norajitra, P. F. Jäger, and K. Maier-Hein (Apr. 1, 2025b). **Comparative Benchmarking of Failure Detection Methods in Medical Image Segmentation: Unveiling the Role of Confidence Aggregation**. In: *Medical Image Analysis* 101, p. 103392. ISSN: 1361-8415. DOI: 10.1016/j.media.2024.103392. URL: https://www.sciencedirect.com/science/article/pii/S1361841524003177 (visited on 01/07/2025) (cit. on pp. 17, 26, 31, 44, 45, 48–51, 59, 62, 67, 89, 91, 95, 98, 100, 101, 103, 104, 107–109, 171, 179–188).

Zhang, L., X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, and Z. Xu (July 2020). **Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation**. In: *IEEE Transactions on Medical Imaging* 39.7, pp. 2531–2540. ISSN: 1558-254X. DOI: 10.1109/TMI.2020.2973595 (cit. on p. 18).

Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy (Apr. 2023). **Domain Generalization: A Survey**. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4, pp. 4396–4415. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2022.3195549 (cit. on pp. 2, 17, 18).

Zhou, Z., M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang (2018). **Unet++: A Nested u-Net Architecture for Medical Image Segmentation**. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3–11 (cit. on p. 189).

Zou, K., X. Yuan, X. Shen, M. Wang, and H. Fu (2022). **TBraTS: Trusted Brain Tumor Segmentation**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li. Cham: Springer Nature Switzerland, pp. 503–513. ISBN: 978-3-031-16452-1. DOI: 10.1007/978-3-031-16452-1_48 (cit. on p. 22).

Own Contributions

This chapter gives an overview of my contributions in distinction to team efforts.

Own share in data acquisition and data analysis

This thesis was written in the division of Medical Image Computing (MIC) at the German Cancer Research Center (DKFZ), headed by Prof. Dr. Klaus Maier-Hein, who is the primary supervisor for this thesis. Throughout the entire time of my thesis, I was closely collaborating with members of Prof. Dr. Klaus Maier-Hein's group. The thesis is divided into two parts:

1. Benchmarking the generalizability of brain tumor segmentation algorithms in the Federated Tumor Segmentation (FeTS) Challenge

2. Benchmarking failure detection methods for segmentation

Part 1 was conducted in an international cooperation with the Department of Pathology & Laboratory Medicine at the University of Indiana, led by Prof. Dr. Spyridon Bakas (part of the work was done while he was affiliated with the Center for Biomedical Image Computing & Analytics at the University of Pennsylvania). Part 2 (failure detection) was an DKFZ-internal project, in collaboration with Dr. Paul F. Jäger, who headed the interactive machine learning (IML) group at DKFZ. In the following, details on my contribution to the data acquisition and the data analysis are provided separately for each of the two parts.

Own share in Data Acquisition

1. The magnetic resonance imaging (MRI) images used in the FeTS Challenge were acquired independently of this thesis by several international institutions. Prof. Dr. Spyridon Bakas' team contacted the data owners and organized the preprocessing as well as the annotation of reference segmentations. I performed the annotation quality control for a subset of the cases, and Dr. Gianluca Brugnara as well as Dr. Julius Holzschuh answered my questions on brain tumor MRI interpretation. The FeTS Challenge used this imaging data to evaluate different algorithms, resulting in the *challenge results data*, which consisted of metric values, algorithm descriptions, and meta-data for the test datasets. Two separate algorithmic tasks were investigated in the challenge. For Task 1, which is not reported in this thesis, the challenge results data was collected by Prof. Dr. Spyridon Bakas' team. For Task 2, I was the main organizer and coordinated the acquisition of the challenge results data, which was a collaborative, multicentric evaluation effort with the data-contributing institutions.

2. For the failure detection benchmark, I did not acquire imaging data myself but relied on publicly available datasets.

Own share in Data Analysis

1. The FeTS Challenge evaluated segmentation algorithms on brain MRI datasets with segmentation annotations, resulting in the challenge results data described above. The segmentation algorithms were contributed by researchers around the world who participated in the challenge. My task was to specify requirements for submissions, provide templates to the participants for creating inference applications that can be run across many institutions, collect the final submissions and check their functionality. Apart from coordinating the federated evaluation, I also prepared the technical infrastructure with technical support from the MedPerf (Karargyris et al. 2023) team, most notably Micah Sheller and Alejandro Aristazabal. Analysis of the acquired challenge results data for the FeTS Challenge task 2 was performed entirely by me, with regular advisory input from colleagues, in particular Dr. David Zimmerer, Dr. Fabian Isensee and Prof. Dr. Klaus Maier-Hein.

2. All methods for the failure detection benchmark were implemented by me, building upon established code libraries as necessary. Since this benchmark compared mostly previously published methods, in two cases existing method implementations were adapted. Computer experiments and their analysis were also performed entirely by me, with regular advisory input from colleagues, in particular Dr. David Zimmerer, Dr. Fabian Isensee, Dr. Paul F. Jäger and Prof. Dr. Klaus Maier-Hein.

Own Publications

This section lists all publications that I was a part of and contributed to during my PhD. It is subdivided into *First Authorships* and *Co-Authorships*. For all first authorships, I describe my contributions to the publication. Throughout the work on all publications, I received advisory input from my supervisor and colleagues at MIC, in particular Dr. David Zimmerer, Dr. Fabian Isensee, Dr. Paul F. Jäger and Prof. Dr. Klaus Maier-Hein.

First Authorships

1. M. Zenk et al. (2025a). **Towards Fair Decentralized Benchmarking of Healthcare AI Algorithms: The Federated Tumor Segmentation (FeTS) Challenge**. Manuscript accepted for publication.

2. M. Zenk et al. (Apr. 1, 2025b). **Comparative Benchmarking of Failure Detection Methods in Medical Image Segmentation: Unveiling the Role of Confidence Aggregation**. In: *Medical Image Analysis* 101, p. 103392. ISSN: 1361-8415. DOI: 10.1016/j.media.2024.103392. URL: https://www.sciencedirect.com/science/article/pii/S1361841524003177 (visited on 01/07/2025)

3. M. Zenk et al. (2022). **Realistic Evaluation of FixMatch on Imbalanced Medical Image Classification Tasks**. In: *Bildverarbeitung Für Die Medizin 2022*. Ed. by K. Maier-Hein et al. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 291–296. ISBN: 978-3-658-36932-3

Publication 1 summarizes the FeTS Challenge and is based on sections 1.3.1, 2.1, 3.1 and 4.1. As the challenge consisted of two *tasks*, the first authorship of this publication is shared between the Task 1 main organizer (Dr. Ujjwal Baid) and the Task 2 organizer (me). I contributed to both tasks in this publication. For Task 1, I performed the analysis based on the challenge results data. For Task 2, I coordinated the challenge from start to end, beginning with the challenge design, the preparation, the execution (including the development of the technical infrastructure), and the analysis. The original draft of the manuscript was written by Dr. Ujjwal Baid, Sarthak Pati and me, with sections on Task 2 exclusively written by me and the other sections written collaboratively. Some parts were adapted from the preprint (Pati et al. 2021) describing the design of the 2021 challenge, which was written with the same contributions. Paper revisions were performed by me, taking into account feedback from the other team members. This thesis only contains a description of Task 2.
Publication 2 is based on sections 1.3.2, 2.2, 3.2 and 4.2. I designed and implemented the benchmarking study, ran experiments, performed the analyses and wrote the original

manuscript draft.
Publication 3 is not described in this thesis. I was responsible for the method implementations, experiments, and analyses for this publication, as well as writing the original manuscript draft.

Co-Authorships

These publications are not described in this thesis.

- Y. He et al. (May 1, 2022). **CosSGD: Communication-Efficient Federated Learning with a Simple Cosine-Based Quantization**. DOI: 10.48550/arXiv.2012.08241. arXiv: 2012.08241 [cs]. URL: http://arxiv.org/abs/2012.08241 (visited on 12/14/2022). Pre-published

- K. Kades et al. (2022). **Towards Real-World Federated Learning in Medical Image Analysis Using Kaapana**. In: *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health*. Ed. by S. Albarqouni et al. Cham: Springer Nature Switzerland, pp. 130–140. ISBN: 978-3-031-18523-6. DOI: 10.1007/978-3-031-18523-6_13

- S. Pati et al. (Dec. 5, 2022a). **Federated Learning Enables Big Data for Rare Cancer Boundary Detection**. In: *Nature Communications* 13.1 (1), p. 7346. ISSN: 2041-1723. DOI: 10.1038/s41467-022-33407-5. URL: https://www.nature.com/articles/s41467-022-33407-5 (visited on 12/14/2022)

- M. Eisenmann et al. (2023). **Why Is the Winner the Best?** In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19955–19966. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Eisenmann_Why_Is_the_Winner_the_Best_CVPR_2023_paper.html (visited on 01/22/2025)

- A. Karargyris et al. (July 17, 2023). **Federated Benchmarking of Medical Artificial Intelligence with MedPerf**. In: *Nature Machine Intelligence*, pp. 1–12. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00652-2. URL: https://www.nature.com/articles/s42256-023-00652-2 (visited on 07/18/2023)

- C. Ulrich et al. (2023). **MultiTalent: A Multi-dataset Approach to Medical Image Segmentation**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan et al. Cham: Springer Nature Switzerland, pp. 648–658. ISBN: 978-3-031-43898-1. DOI: 10.1007/978-3-031-43898-1_62

- M. R. Bujotzek et al. (Oct. 25, 2024). **Real-World Federated Learning in Radiology: Hurdles to Overcome and Benefits to Gain**. In: *Journal of the American Medical Informatics Association*, ocae259. ISSN: 1527-974X. DOI: 10.1093/jamia/ocae259. URL: https://doi.org/10.1093/jamia/ocae259 (visited on 11/05/2024)

- K.-C. Kahl et al. (Jan. 16, 2024). **ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation**. DOI: 10.48550/arXiv.2401.08501. arXiv: 2401.08501 [cs]. URL: http://arxiv.org/abs/2401.08501 (visited on 01/25/2024). Pre-published

- Y. Kirchhoff et al. (2024). **Skeleton Recall Loss for Connectivity Conserving and Resource Efficient Segmentation of Thin Tubular Structures**. In: *Computer Vision – ECCV 2024*. Ed. by A. Leonardis et al. Cham: Springer Nature Switzerland, pp. 218–234. ISBN: 978-3-031-72980-5. DOI: 10.1007/978-3-031-72980-5_13

- A. M. Mora et al. (Apr. 27, 2024). **Curriculum-Learning for Vessel Occlusion Detection in Multi-site Brain CT Angiographies**. In: Medical Imaging with Deep Learning. URL: https://openreview.net/forum?id=6TrjwzbBko (visited on 01/22/2025)

- S. Parampottupadam et al. (2024). **Client Security Alone Fails in Federated Learning: 2D and 3D Attack Insights**. In: *MICCAI Student Board EMERGE Workshop: Empowering Medical Image Computing & Research through Early-Career Expertise*

- M. Rokuss et al. (Sept. 20, 2024). **Longitudinal Segmentation of MS Lesions via Temporal Difference Weighting**. DOI: 10.48550/arXiv.2409.13416. arXiv: 2409.13416 [eess]. URL: http://arxiv.org/abs/2409.13416 (visited on 01/22/2025). Pre-published

- C. Ulrich et al. (2025). **Mitigating False Predictions in Unreasonable Body Regions**. In: *Machine Learning in Medical Imaging*. Ed. by X. Xu et al. Cham: Springer Nature Switzerland, pp. 22–31. ISBN: 978-3-031-73290-4. DOI: 10.1007/978-3-031-73290-4_3

A Appendix

A.1 Additional Results

A.1.1 Generalization

This section contains additional details on the Federated Tumor Segmentation (FeTS) Challenge 2022 results. Figure A.1 shows a heatmap for the mean Hausdorff distance (HD) instead of Dice similarity coefficient (DSC). Figure A.2 reports results for individual tumor regions (whole tumor (WT), tumor core (TC) and enhancing tumor (ET)) of the best model, which extends the lower diagram in fig. 3.4. To supplement the interpretation of the annotation quality control results in section 3.1.2.2, figs. A.3 and A.4 compare how the ranking and per-case DSC distributions would change if quality control was not performed.

A.1.2 Failure Detection

As a comprehensive comparison of all evaluated methods, table A.1 shows mean and standard deviation of the area under the risk-coverage curve (AURC) metric on the computed tomography (CT) and magnetic resonance imaging (MRI) datasets used in the benchmark. It extends table 3.3 with more combinations of prediction model and pixel/image-level failure detection methods. Figures A.5 to A.7 complement section 3.2.3 by providing a similar plot as fig. 3.12 but with more methods and alternative metrics, namely Spearman correlation coefficient (SC) and Pearson correlation coefficient (PC).

A.2 Additional Image Samples from the Failure Detection Benchmark

Figures A.8 to A.13 show additional samples from the test sets of each CT/MRI dataset used for the segmentation failure detection benchmark (section 2.2.3). Each dataset usually consists of several domains, some of which were seen during training and some are new in the test set. These are indicated above the images.

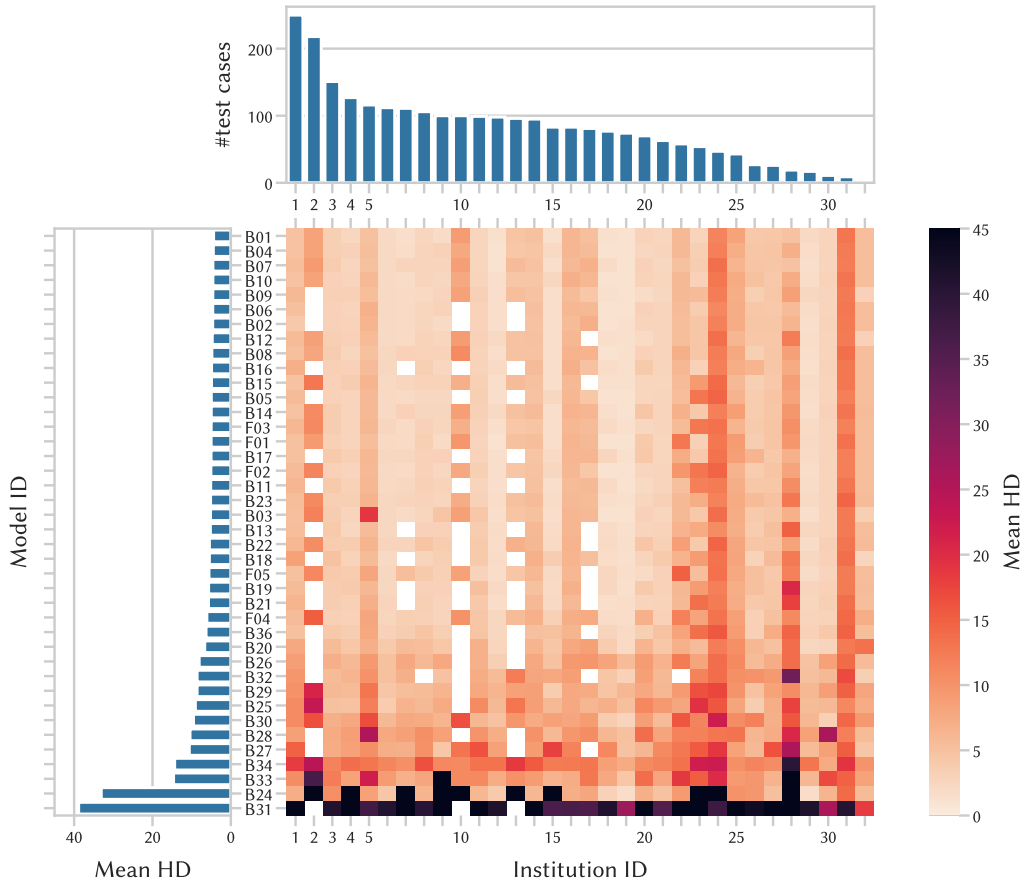

Figure A.1: Aggregated challenge results of the FeTS22 challenge for each evaluated model and institution. Each tile in the heatmaps represents the HD value of a single model, averaged over all test cases and tumor regions of one institution. The values were clipped at 0.5 and white tiles indicate evaluation runs that failed due to technical issues. Models are sorted by mean HD (bar plot on the left) and institutions by their test set size (bar plot at the top). The best models achieve similar performances within each institution, which is apparent from the vertical structures in the heatmap. However, on some institutions the performance of all models drops, indicating a lack of robustness. Figure adapted from (Zenk et al. 2025a).

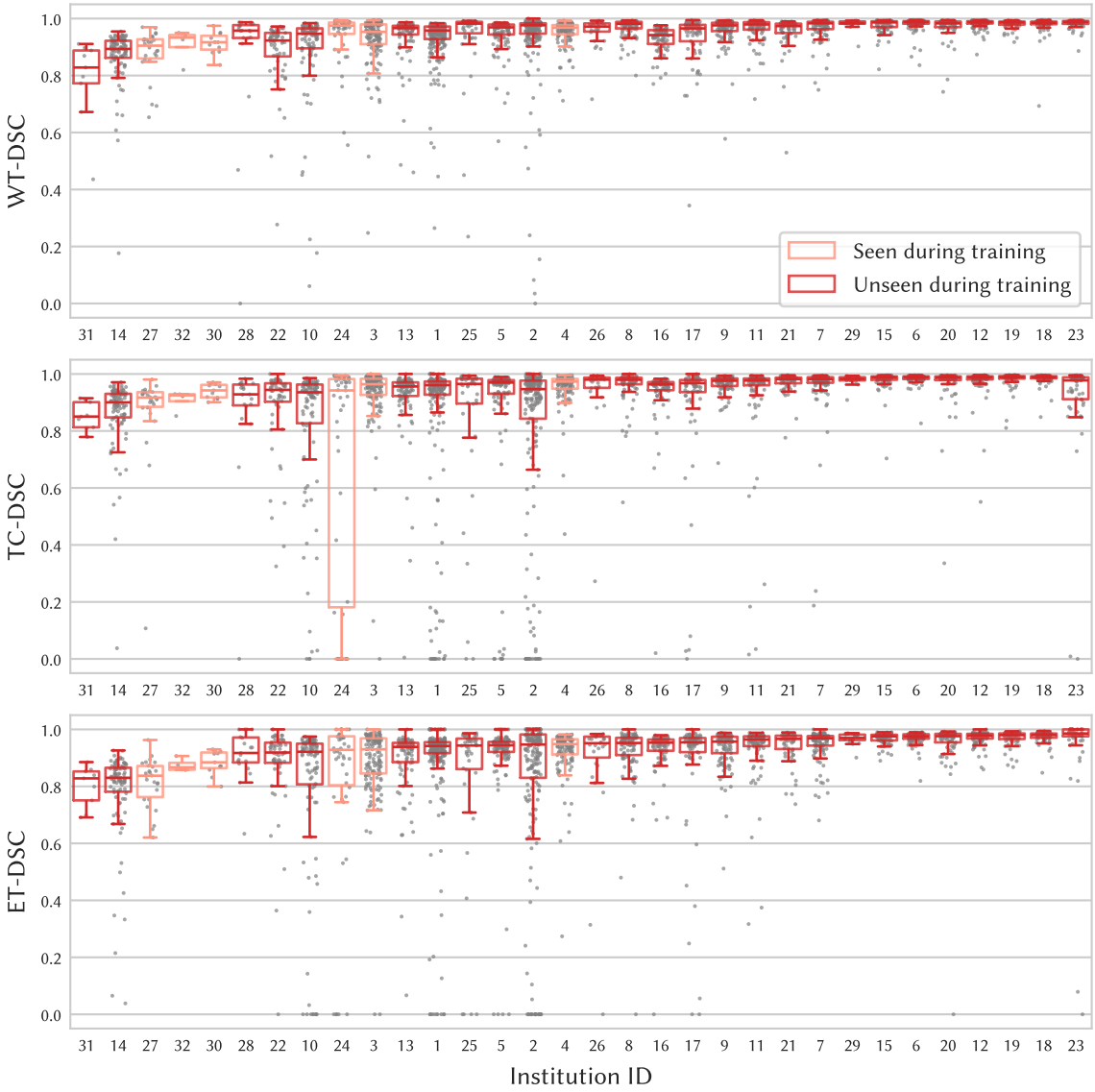

Figure A.2: Case-level challenge results of the top-ranked algorithm (ID: B01) for each institution of the test set. Some Institutions also contributed (different) cases to the training set ("seen during training"), while others were not seen during training. Each gray dot represents the DSC score for a single test case. While the median performance is high for most institutions, there are also often individual cases with reduced performance, even for institutions seen during training. Comparing the three tumor regions, for TC and ET there are more outliers, but even the usually large WT region is not always segmented accurately. Figure adapted from (Zenk et al. 2025a).

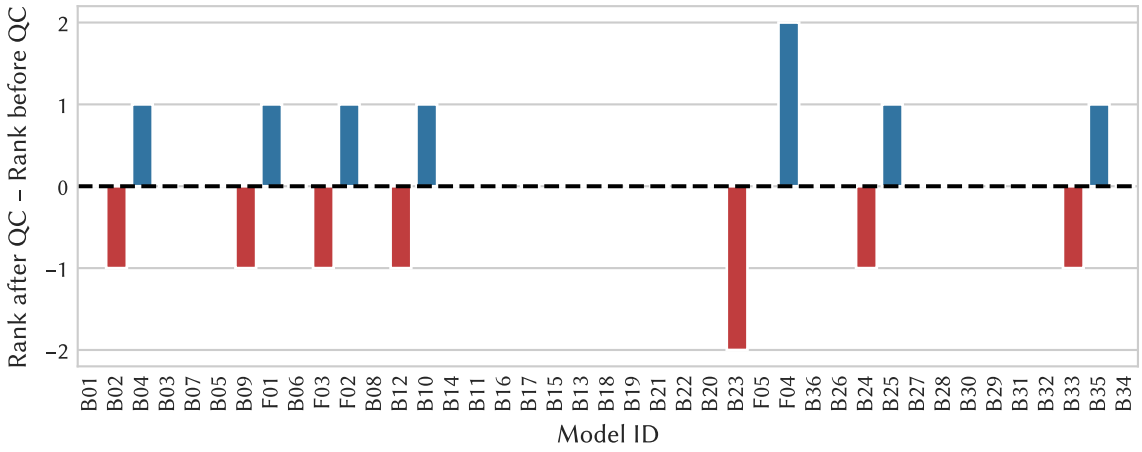

Figure A.3: Effect of annotation quality control (QC) on rankings. Ranking is only compared between the subsets that were quality controlled (1201 cases from which 125 were excluded in the QC). There are only minor changes in the ranking of up to two positions.

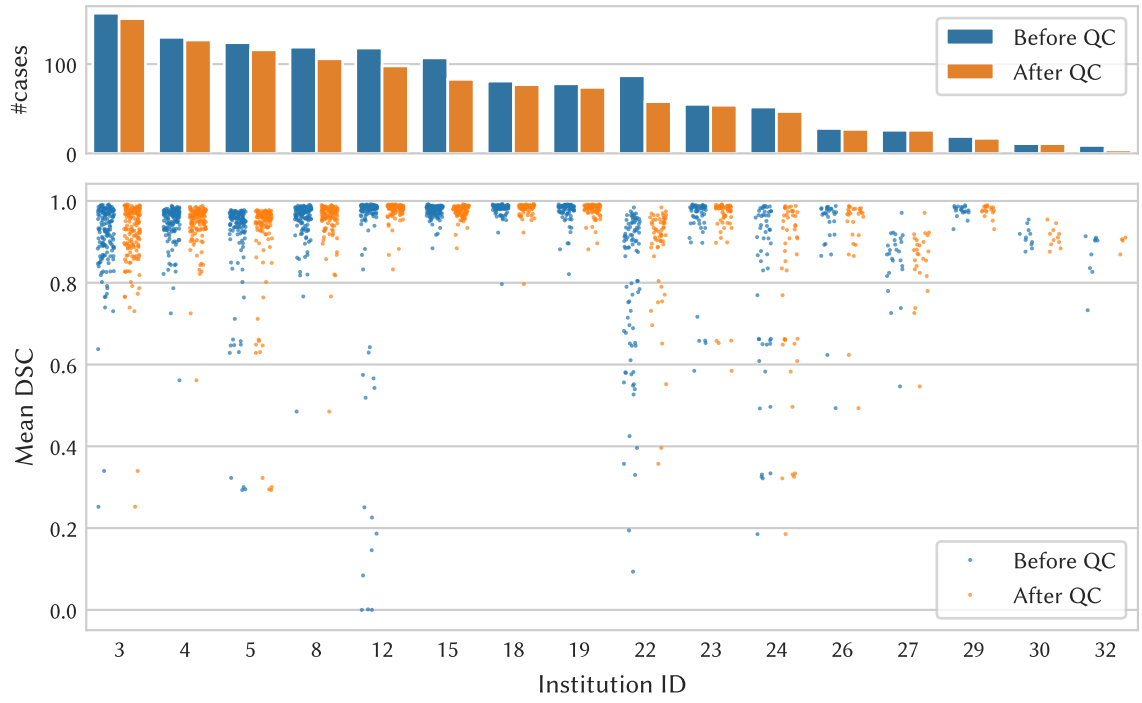

Figure A.4: Effect of annotation quality control (QC) on the mean DSC distributions for the best-performing model (ID: B01). Only results from the institutions that were quality controlled are shown (1201 from 16 institutions, reduced by 125 cases through QC). Although for a few institutions (IDs: 12, 22) the low-quality samples coincided with low-DSC, the DSC distributions before and after QC look similar.

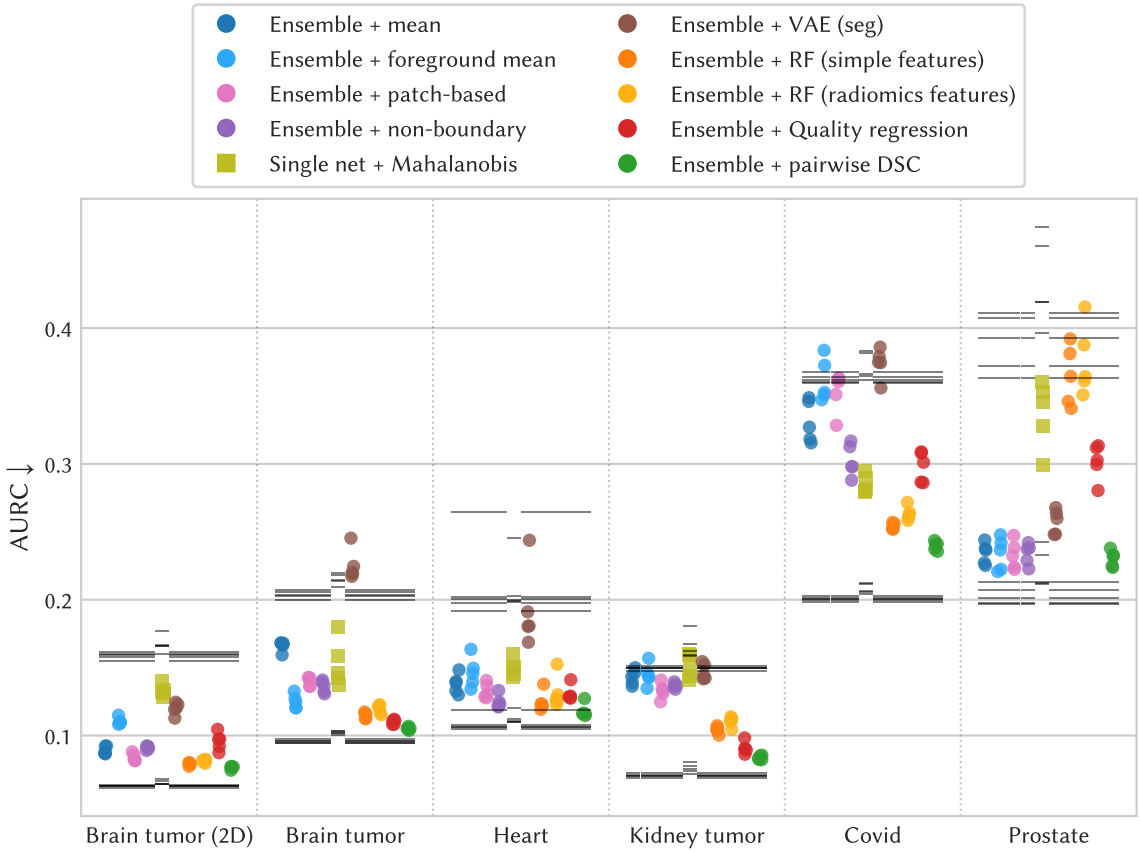

Figure A.5: AURC scores for all datasets and methods in the failure detection benchmark. Each colored marker corresponds to a single experiment, meaning a prediction model + aggregation method, trained on one fold of the training data. The marker shape represents differences in prediction models. Gray "–" markers visualize values for the random/optimal confidence baselines, which differ between the models trained on different folds due to their varying segmentation performance (section 2.2.2.2). Pairwise DSC is among the best methods across all datasets, but the absolute correlation varies between datasets. Quality regression networks are usually the next-best option, but they show a performance drop on the Covid, Prostate and Brain tumor (2D) datasets. Figure adapted from (Zenk et al. 2025b).

Table A.1: AURC scores (×100) on the test sets for all compared failure detection methods based on three types of prediction models. Mean and standard deviation (std) are computed across five prediction models trained on different folds. A color map is applied on each 'mean' column, ranging from light yellow (worse) to dark green (best). PE was used for pixel uncertainty. Abbreviations: Ens.: Ensemble, MCD: MC-Dropout Single: Single network. Table adapted from (Zenk et al. 2025b).

	Brain-2d		Brain		Heart		Kidney		Covid		Prostate	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Ens. + Quality regression	9.6	0.6	11.0	0.2	13.1	0.6	9.1	0.4	29.8	1.1	30.2	1.3
Ens. + RF (radiomics features)	8.1	0.1	11.9	0.3	13.2	1.2	11.1	0.4	26.3	0.5	37.6	2.6
Ens. + RF (simple features)	7.9	0.1	11.5	0.2	12.5	0.7	10.4	0.3	25.4	0.2	36.5	2.2
Ens. + VAE (seg)	12.0	0.5	22.5	1.2	19.3	2.9	14.7	0.6	37.4	1.1	25.8	0.9
Ens. + foreground mean	11.1	0.3	12.5	0.5	14.6	1.1	14.5	0.8	36.1	1.6	23.4	1.2
Ens. + mean	8.9	0.3	16.6	0.4	13.8	0.7	14.3	0.5	33.1	1.5	23.4	0.8
Ens. + non-boundary	9.1	0.1	13.6	0.4	12.5	0.5	13.7	0.2	30.3	1.2	23.4	0.8
Ens. + pairwise DSC	7.6	0.1	10.5	0.1	11.8	0.5	8.4	0.2	24.0	0.3	23.0	0.6
Ens. + patch-based	8.4	0.3	14.0	0.3	13.3	0.6	13.3	0.6	35.3	1.5	23.3	1.0
MCD + foreground mean	11.8	0.6	15.3	0.6	16.4	0.9	15.8	1.4	37.7	1.2	28.8	2.2
MCD + mean	10.1	0.6	20.1	1.2	14.7	0.5	15.8	0.9	36.3	1.1	26.6	2.0
MCD + non-boundary	9.6	0.2	15.6	0.4	12.9	0.4	14.6	0.3	34.9	1.5	26.4	2.1
MCD + pairwise DSC	8.1	0.2	11.3	0.1	12.8	0.4	8.9	0.4	25.1	0.9	24.8	1.2
MCD + patch-based	9.1	0.5	16.4	1.0	14.1	0.3	13.8	0.6	38.2	1.2	26.4	1.9
Single + Mahalanobis	13.3	0.4	15.3	1.7	15.0	0.6	15.1	0.9	28.5	0.7	33.7	2.4
Single + Quality regression	9.7	0.5	11.5	0.2	13.4	0.4	9.8	0.5	30.5	1.9	31.9	1.4
Single + RF (radiomics features)	11.5	0.6	12.5	0.3	13.2	0.6	12.0	0.7	26.3	0.8	40.9	5.0
Single + RF (simple features)	10.7	0.3	12.3	0.3	13.1	0.3	11.5	0.9	27.2	1.3	38.8	4.5
Single + foreground mean	15.2	1.0	15.4	0.6	16.8	1.0	15.5	1.1	37.7	1.2	30.5	2.9
Single + mean	15.1	1.3	20.1	1.3	14.8	0.5	16.1	0.8	36.5	1.1	27.3	2.5
Single + non-boundary	14.8	0.8	15.6	0.5	12.9	0.4	14.8	0.3	35.0	1.4	26.9	2.3
Single + patch-based	14.1	1.3	16.4	1.1	14.3	0.5	14.0	0.8	38.4	1.2	27.7	2.6

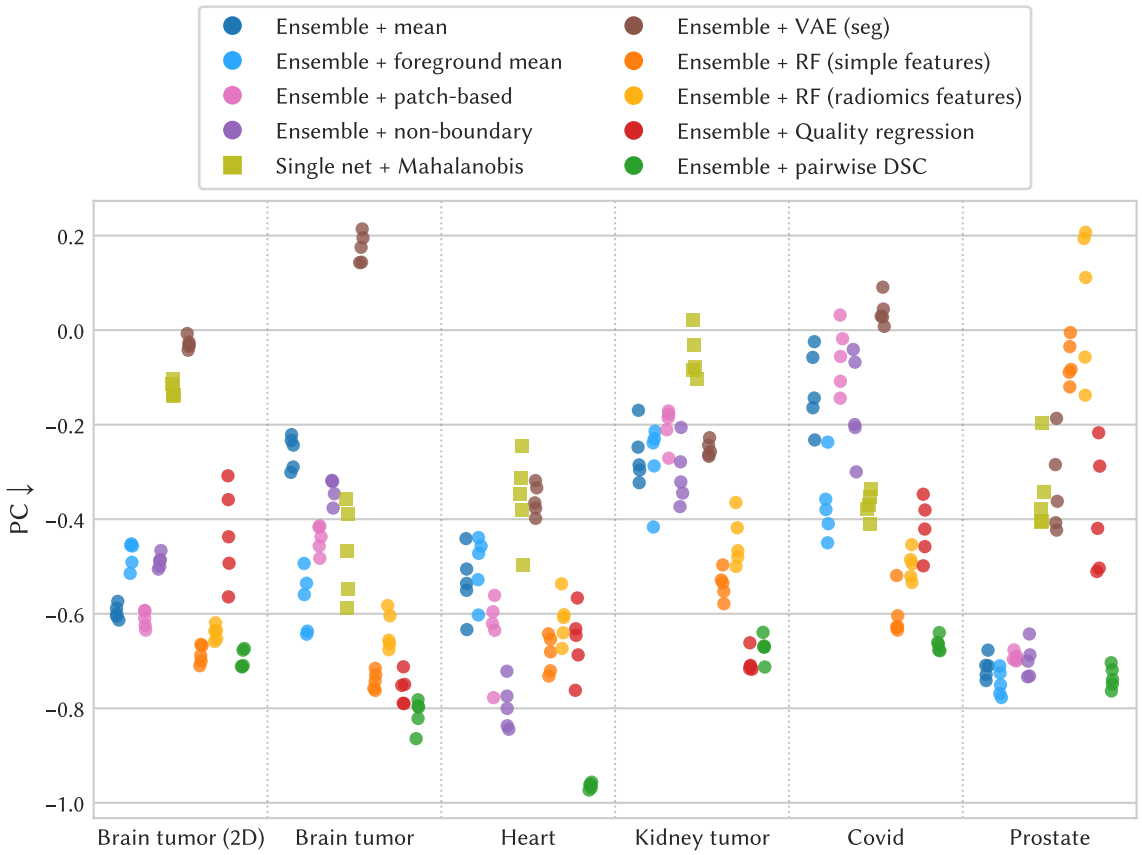

Figure A.6: Pearson correlation coefficient (PC) for all datasets and methods in the failure detection benchmark. Each colored marker corresponds to a single experiment, meaning a prediction model + aggregation method, trained on one fold of the training data. The marker shape represents differences in prediction models. Gray "–" markers visualize values for the random/optimal confidence baselines, which differ between the models trained on different folds due to their varying segmentation performance (section 2.2.2.2). Pairwise DSC performs clearly best across all datasets, often achieving close to optimal AURC. The results are overall in line with the observations under the AURC metric in fig. A.5. Figure adapted from (Zenk et al. 2025b).

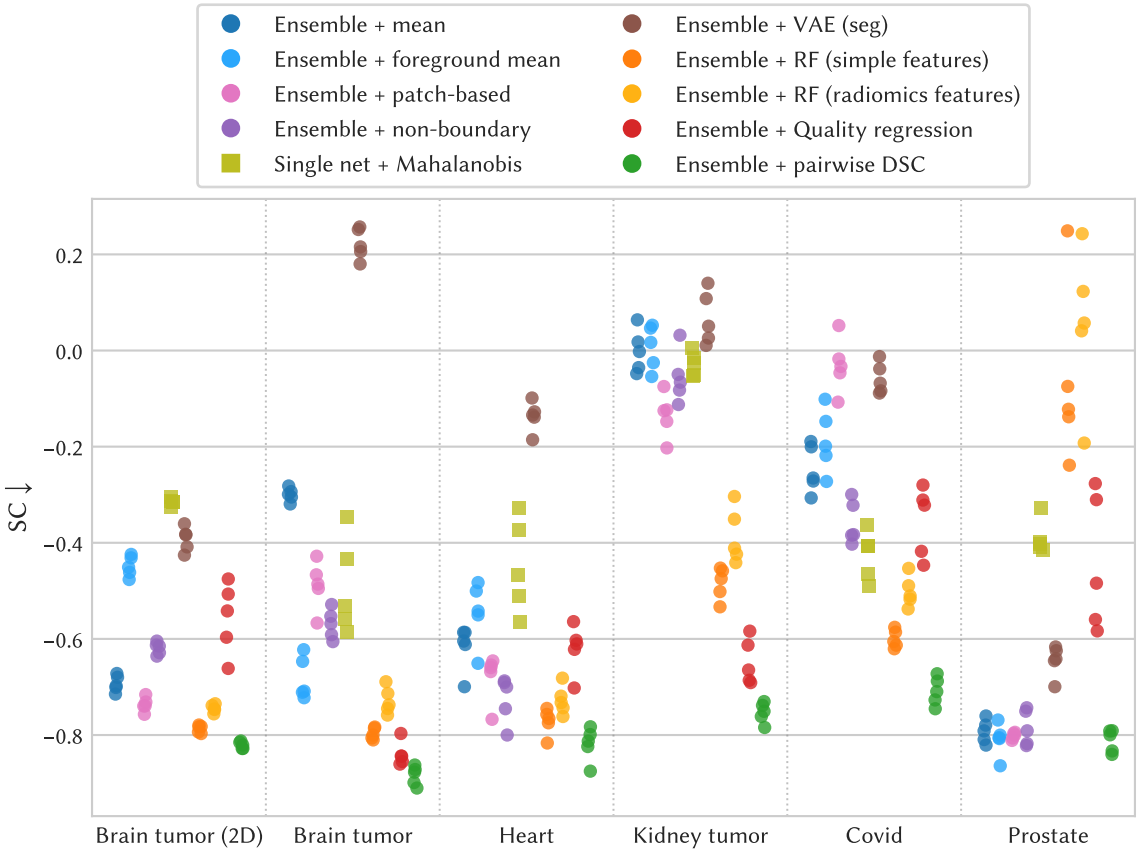

Figure A.7: Spearman correlation coefficient (SC) for all datasets and methods in the failure detection benchmark. Each colored marker corresponds to a single experiment, meaning a prediction model + aggregation method, trained on one fold of the training data. The marker shape represents differences in prediction models. Gray "–" markers visualize values for the random/optimal confidence baselines, which differ between the models trained on different folds due to their varying segmentation performance (section 2.2.2.2). Pairwise DSC performs clearly best across all datasets, achieving SCs of -0.7 and higher. The results are overall in line with the observations under the AURC metric in fig. A.5. Figure adapted from (Zenk et al. 2025b).

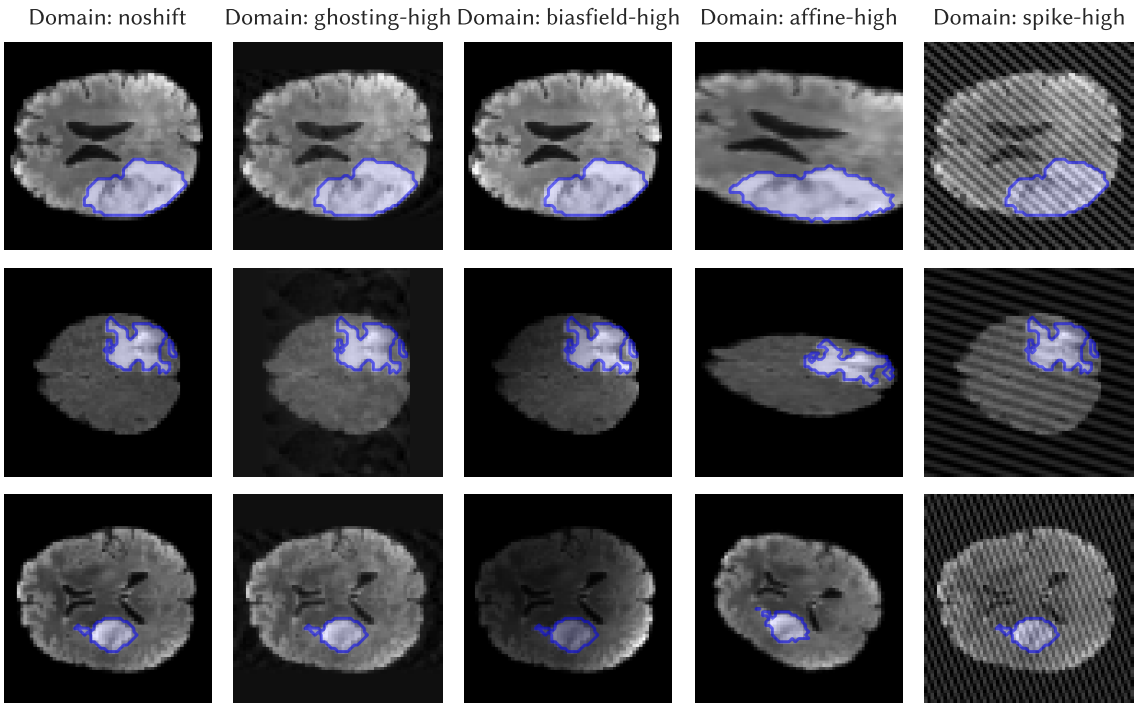

Figure A.8: Samples from the test set of the 2D brain (toy) dataset. Each column shows samples from a different "domain", which corresponds to an artificial corruption for this dataset. Figure adapted from (Zenk et al. 2025b).

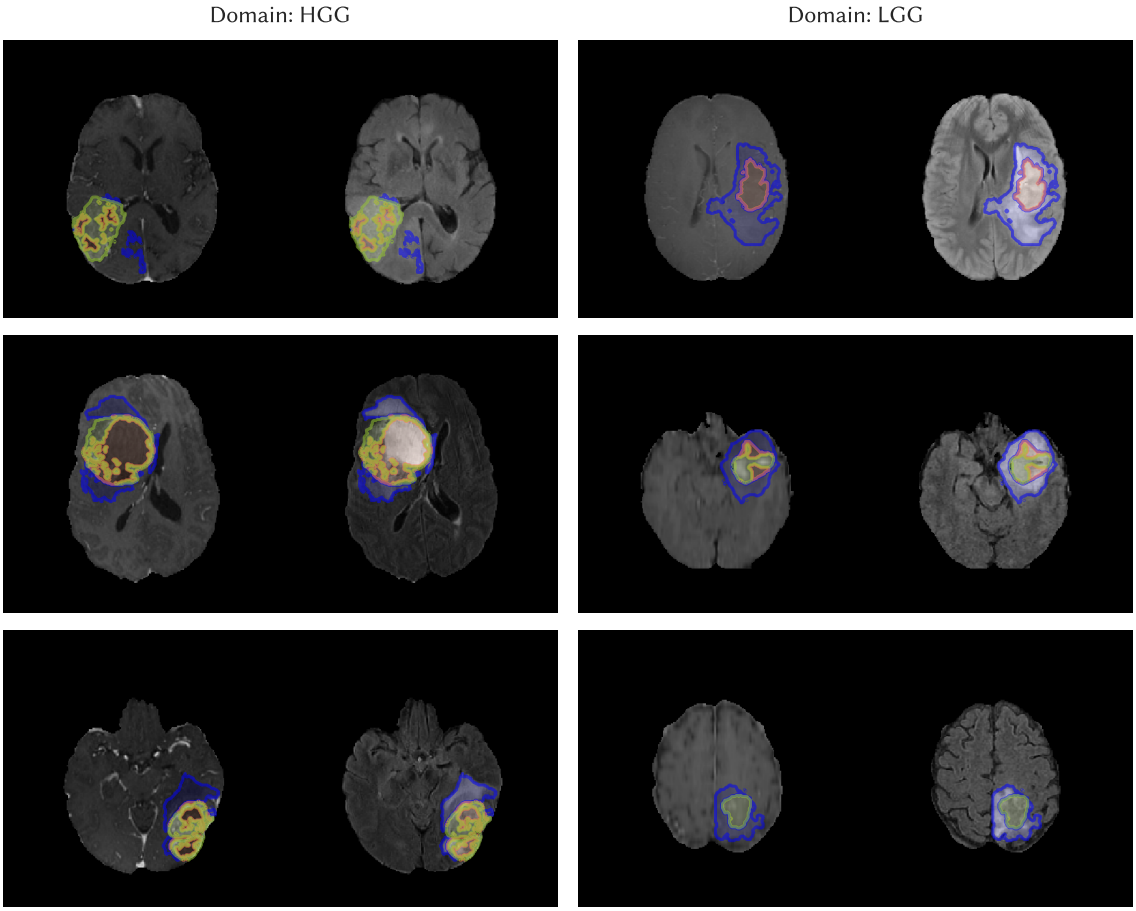

Figure A.9: Samples from the test set of the brain tumor dataset. Each column shows samples from a different "domain", which corresponds to an artificial corruption for this dataset. Figure adapted from (Zenk et al. 2025b).

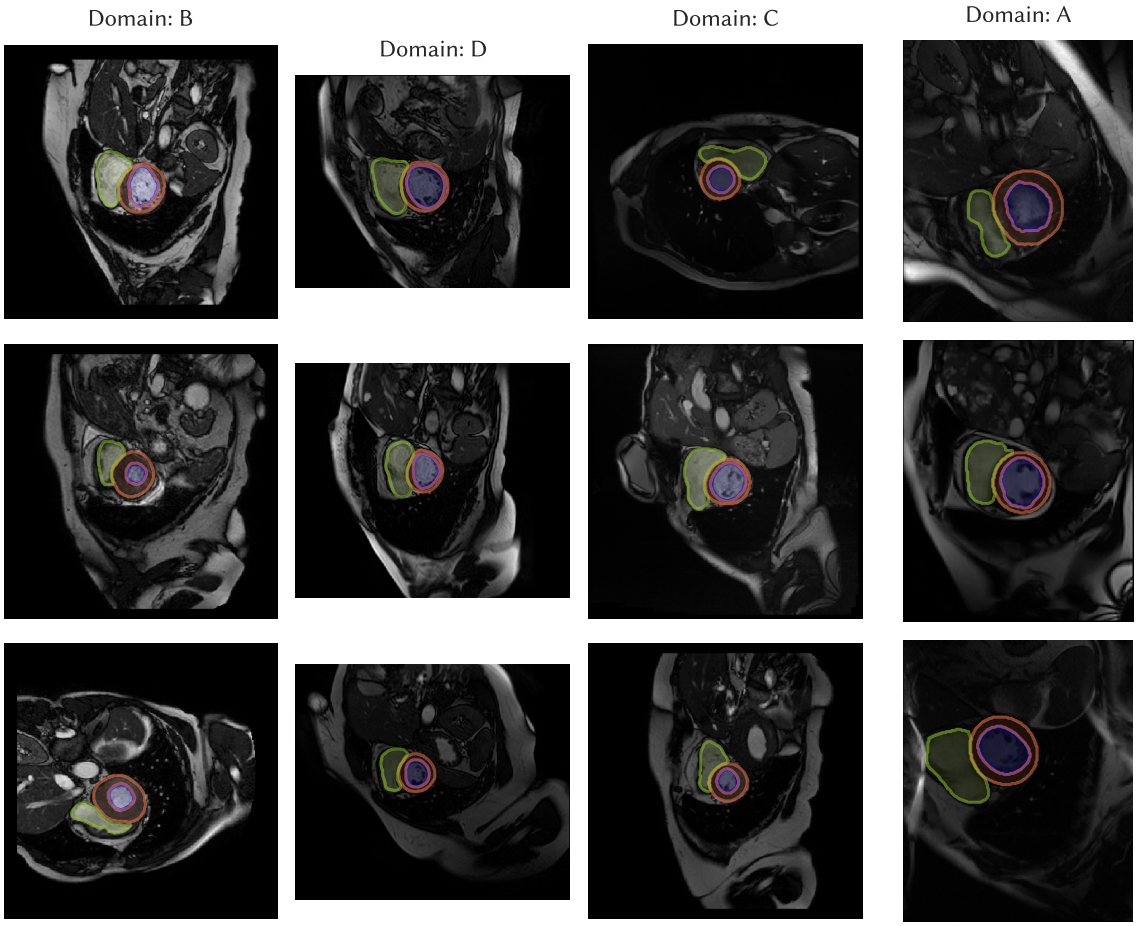

Figure A.10: Samples from the test set of the heart dataset. Each column shows samples from a different "domain", which corresponds to an artificial corruption for this dataset. Figure adapted from (Zenk et al. 2025b).

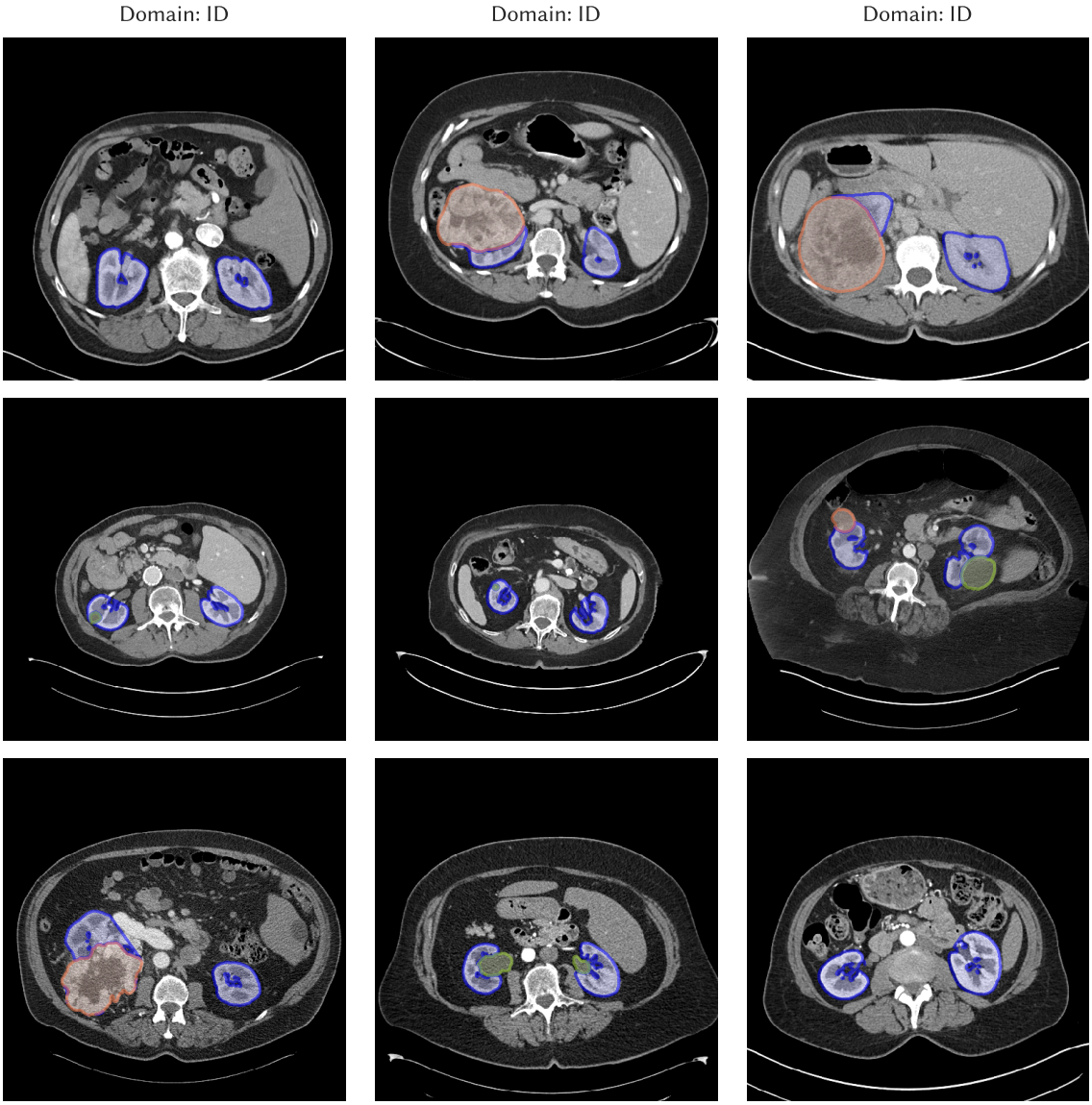

Figure A.11: Samples from the test set of the kidney tumor dataset. There is only one in-distribution domain. Figure adapted from (Zenk et al. 2025b).

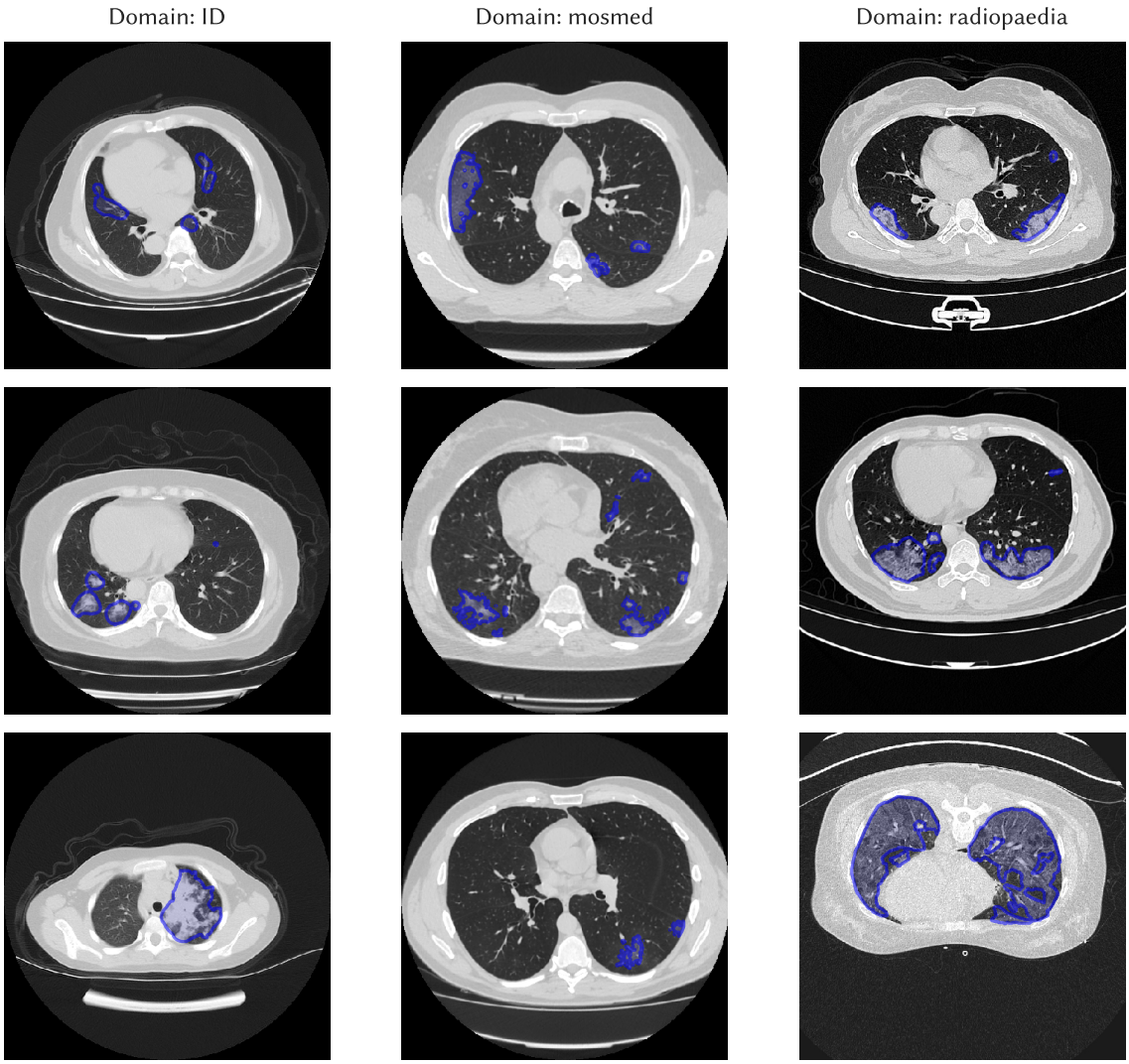

Figure A.12: Samples from the test set of the prostate dataset. Each column shows samples from a different "domain", which corresponds to an artificial corruption for this dataset. Figure adapted from (Zenk et al. 2025b).

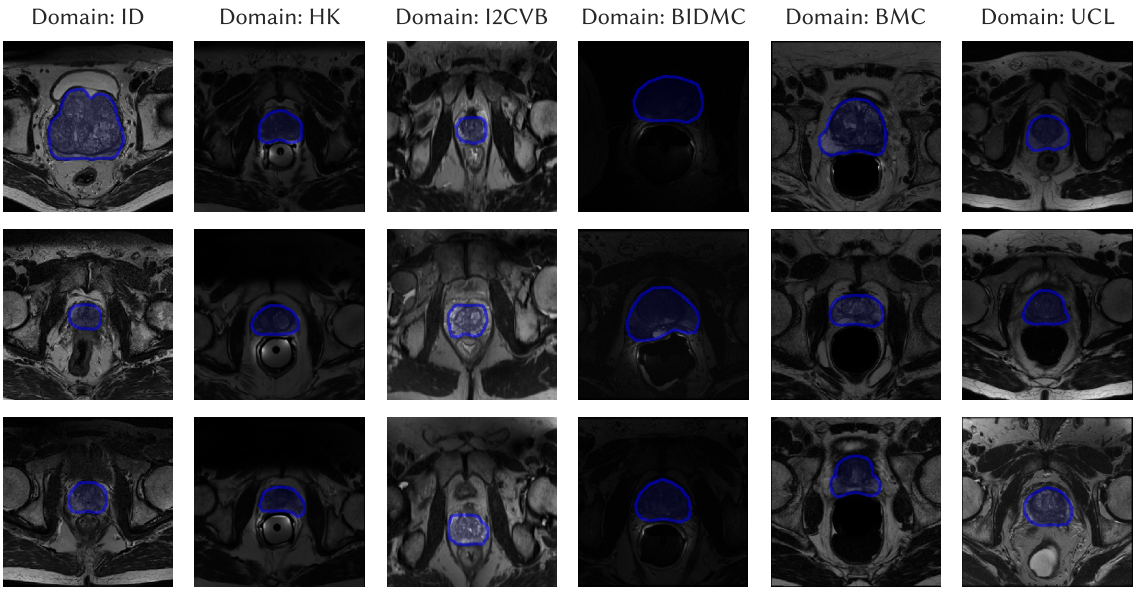

Figure A.13: Samples from the test set of the Covid dataset. Each column shows samples from a different "domain", which corresponds to an artificial corruption for this dataset. Figure adapted from (Zenk et al. 2025b).

A.3 Details on the FeTS Challenge Submissions (Algorithm Characteristics)

The FeTS challenge 2022 evaluated 41 models in total, from which 5 were official submissions and 36 were originally contributed to the Brain Tumor Segmentation (BraTS) challenge 2021 (Baid et al. 2021). The five official submissions are described in section 3.1.2.1, while the 36 BraTS algorithms were described in scientific publications previously (except three teams that did not submit a paper). Table A.2 lists the corresponding references, which can provide additional details beyond table 3.1. The following paragraphs elaborate on the entries for architecture, loss and post-processing, by providing references where necessary and clarifying certain characteristics. They are taken without modification from Zenk et al. (2025a):

Architecture The most common backbone used by the submissions was U-Net (Ronneberger et al. 2015). Several variations to the basic U-Net were introduced by the teams: Some used larger encoders, with more filters per convolution or more convolutional blocks per stage. Adding residual connections to convolutional blocks (He et al. 2016) was also common. Several algorithms extended the U-Net with different kinds of attention modules. Examples include inserting a transformer in the bottleneck of the U-Net or re-weighting feature maps with attention restricted to the channel/spatial dimensions. Some participants used other convolutional neural networks (CNNs) than U-Net, for instance HR-Net (Sun et al. 2019), HNF-Net (Jia et al. 2021), U-Net++ (Zhou et al. 2018) and HarDNet (Chao et al. 2019). Recent hybrid CNN/transformer networks like CoTr (Xie et al. 2021), Swin transformer (Liu et al. 2021) were incorporated in some submissions. Finally, a few teams utilized skip connection blocks that combined features from multiple stages or explored splitting the segmentation task into two stages, first segmenting a coarse whole tumor region and then refining the segmentation of this cropped region.

Loss The most common loss functions were Dice (computed either per sample or per batch) and cross-entropy. Similar to the Dice loss, some teams optimized differentiable versions of segmentation metrics (Jaccard index, generalized Dice, boundary distance and the generalized Wasserstein Dice loss (Fidon et al. 2021)). Two less common loss functions were TopK loss, which considers only the K pixels with the highest loss, and the focal loss, which down-weights the loss for pixels that are classified correctly with high softmax scores. Finally, one team used virtual adversarial training (Miyato et al. 2018) as an auxiliary, regularizing loss term. Most losses can be calculated either region-based (for each of WT, TC, ET) or for the exclusive labels (edema (ED), necrotic core/necrocyst (NCR), ET).

Table A.2: Mapping from model ID to scientific publication for the subset of BraTS 2021 models evaluated within FeTS22. Submissions are listed in ranking order of the BraTS 2021 challenge (top 10 are marked bold). Note that the top 10 here are not identical with the official BraTS challenge winners, because some top models could not be evaluated in the FeTS challenge for technical reasons. Table adapted from (Zenk et al. 2025a).

ID	Reference	ID	Reference
B01	Luu and Park (2021)	**B02**	Yuan (2021)
B03	Futrega et al. (2021)	**B04**	Ma and Chen (2021)
B05	Kotowski et al. (2021)	**B06**	Jia et al. (2021)
B07	Dobko et al. (2021)	**B08**	Alam et al. (2021)
B09	Nguyen-Truong and Pham (2021)	**B10**	Fidon et al. (2021)
B11	Yang et al. (2021b)	B12	Jiang et al. (2021)
B13	Wu and Lin (2021)	B14	Zeineldin et al. (2021)
B15	n/a (team *tigerduck*)	B16	Carré et al. (2021)
B17	Pnev et al. (2021)	B18	Feng et al. (2021)
B19	Singh (2021)	B20	Pawar et al. (2021b)
B21	n/a (team *younet*)	B22	Bukhari and Mohy-ud-Din (2021)
B23	Milesi et al. (2021)	B24	Demoustier et al. (2021)
B25	Shah et al. (2021)	B26	Li et al. (2021)
B27	Akbar et al. (2021)	B28	Yang et al. (2021a)
B29	Maurya et al. (2021)	B30	Lin et al. (2021)
B31	Roth et al. (2021)	B32	Saueressig et al. (2021)
B33	Yan et al. (2021)	B34	n/a (team *Team Two*)
B35	Hsu et al. (2021)	B36	Druzhinina et al. (2021)

Post-processing Techniques that refine a model's segmentation output based on prior knowledge specific to the three brain tumor regions were popular in the challenge. Dropping small connected components from the final mask (or replacing them with neighboring predictions) can help to reduce false positives. Morphological operations like closing or hole filling were also applied by some teams. Since TC is usually a compact core within WT, post-processing methods enforced this property, by removing TC parts that extend beyond WT or filling holes inside TC. Finally, potential confusion between ET and NCR was counteracted by converting ET output regions to NCR if they are very small (or for one team, if an auxiliary network suggests this).

A.4 Dokumentation der verwendeten KI-Hilfsmittel

KI-basierte Hilfsmittel wurden in der vorliegenden Arbeit für folgende Ziele eingesetzt:

- Optimierung von Sprache und Stil
- Beschleunigung von Textüberarbeitung
- Unterstützung bei der Formatierung mit LaTeX

Allgemein wurden bei Verwendung der Hilfsmittel die erzeugten Inhalte genau geprüft und, falls nötig, angepasst. Im Folgenden ist angegeben, für welchen Abschnitt der Arbeit KI-basierte Hilfsmittel verwendet wurden und wie, um oben genannte Ziele zu erreichen:

Abschnitt 1.2	Mithilfe von GitHub Copilot wurden LaTeX Tabellen basierend auf meinen Befehlen umformatiert und Formeln vervollständigt.
Abschnitte 1.3.2, 2.1 und 3.1	Textausschnitte wurden ChatGPT mitgeteilt, mit dem Auftrag, meine detaillierten Stichpunkte auszuformulieren, Freitext sprachlich zu verbessern oder alternative Formulierungen für komplexe Sätze vorzuschlagen. Beim Einarbeiten einzelner Textpassagen aus eigenen Veröffentlichungen in die Dissertation wurden Vorschläge von ChatGPT zu alternativen Formulierungen berücksichtigt, um den bestehenden Text zu überarbeiten und auf die Struktur und den Stil dieser Arbeit anzupassen. LaTeX-Tabellen wurden optisch aufgewertet mit ChatGPT und GitHub Copilot.
Abschnitte 1.3.1, 2.2, 3.1 und 3.2	Textausschnitte wurden ChatGPT mitgeteilt, mit dem Auftrag, meine detaillierten Stichpunkte auszuformulieren, Freitext sprachlich zu verbessern oder alternative Formulierungen für komplexe Sätze vorzuschlagen. Beim Einarbeiten einzelner Textpassagen aus eigenen Veröffentlichungen in die Dissertation wurden Vorschläge von ChatGPT zu alternativen Formulierungen berücksichtigt, um den bestehenden Text zu überarbeiten und auf die Struktur und den Stil dieser Arbeit anzupassen.
Abschnitt 4	ChatGPT wurde benutzt, um sprachliche Fehler mit minimalen Änderungen zu korrigieren.
Abschnitt 6	Die Übersetzung der englischen Zusammenfassung ins Deutsche von ChatGPT wurde überarbeitet, um diesen Abschnitt zu schreiben. Die englische Zusammenfassung habe ich ohne Hilfsmittel geschrieben.

Acknowledgments

After several years of research in the division of Medical Image Computing (MIC) at the German Cancer Research Center (DKFZ), I am happy to write these words today and would like to express my gratitude to the people who were essential in reaching this point.

First and foremost, I sincerely thank Prof. Dr. Klaus Maier-Hein for giving me the opportunity to dive into the field of medical image analysis and for his invaluable supervision, especially when a new direction was needed. One of the many changes in the division during my PhD was the introduction of internal thesis advisory committees. The feedback I received from Dr. David Zimmerer, Dr. Fabian Isensee, and Dr. Paul Jäger in my committee helped me a lot, and I am sincerely grateful to all of them. Furthermore, I thank Prof. Dr. Philipp Vollmuth for serving as an external advisory member.

A significant part of this thesis was a collaboration with the group of Prof. Dr. Spyridon Bakas, including Sarthak Pati and Dr. Ujjwal Baid, as well as the MedPerf team, most notably Micah Sheller and Alejandro Aristizábal. I am deeply grateful to Prof. Bakas and all those involved in the FeTS Challenges for their support—this project would not have been possible without them!

My colleagues in the MIC division (and the extended SYMIC group) have been immensely helpful to me, not only in solving problems and expanding my knowledge but also in creating a great working environment with plenty of fun at social events. Special thanks go to: My office mates Alex, Balint, Ole, Shuhan, and Stefan; my thesis proofreaders Andrés, David, Markus, and Michael; and finally to my "so wird das nichts" colleagues from the very beginning, Jan, Michael, and Silvia.

Challenges and failures were not only research topics of my PhD but also personal experiences throughout this time. Dealing with them and ultimately succeeding would have been impossible without my family and friends. I cannot thank them enough for their unconditional support and the joy they bring into my life. Among these wonderful people, the one who is closest to me is Jana. Thank you for being by my side every day—I am truly happy to be with you.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema ***Robustness of Medical Image Segmentation Algorithms in the Context of Federated Data*** handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Heidelberg, 5. März 2025

Maximilian Armin Zenk

Angaben zu verwendeten KI-basierten Elektronischen Hilfsmitteln

Zur Dokumentation der verwendeten Hilfsmittel ist der schriftlichen Ausarbeitung ein besonderer Anhang (Anhang A.4) hinzugefügt, der eine Liste und Beschreibung aller verwendeter KI-basierter Hilfsmittel enthält. Der besondere Anhang zur Dokumentation der verwendeten Hilfsmittel erfüllt folgende Kriterien:

1. Auflistung der Ziele, für die die KI-basierten Hilfsmittel in der vorliegenden Arbeit eingesetzt wurden.
2. Dokumentation der Verwendungsweise der KI-basierten Hilfsmittel.
3. Nennung der Kapitel und Abschnitte der vorliegenden Arbeit, in denen die KI-basierten Hilfsmittel eingesetzt wurden, um Inhalte zu erzeugen.

Der Gebrauch dieser Hilfsmittel inklusive Art, Ziel und Umfang des Gebrauchs wurde mit meinem offiziellen Betreuer Prof. Dr. rer. nat. Klaus Maier-Hein abgesprochen.

Mir ist bewusst, dass insbesondere der Versuch einer nicht dokumentierten Nutzung KI- basierter Hilfsmittel als Täuschungsversuch zu werten ist:

Gem. § 16 Abs. 2 der Promotionsordnung „Dr. med. / dent.“: „Ergibt sich vor Aushändigung der Promotionsurkunde, dass der Kandidat/die Kandidatin bei einer Promotionsleistung getäuscht hat, so können einzelne oder alle Promotionsleistungen für ungültig erklärt werden. In schweren Fällen kann die Zulassung zum Promotionsverfahren zurückgenommen werden.“

Und § 16 Abs. 2 der Promotionsordnung „Dr. sc. hum.“: „Ergibt sich vor Aushändigung der Promotionsurkunde, dass der Doktorand / Doktorandin bei einer Promotionsleistung getäuscht hat, so kann der Promotionsausschuss diese Promotionsleistung oder alle bisher erbrachten Promotionsleistungen für ungültig erklären. In besonders schweren Fällen kann der Promotionsausschuss die Annahme als Doktorand / Doktorandin endgültig widerrufen.“

Heidelberg, 5. März 2025

Maximilian Armin Zenk