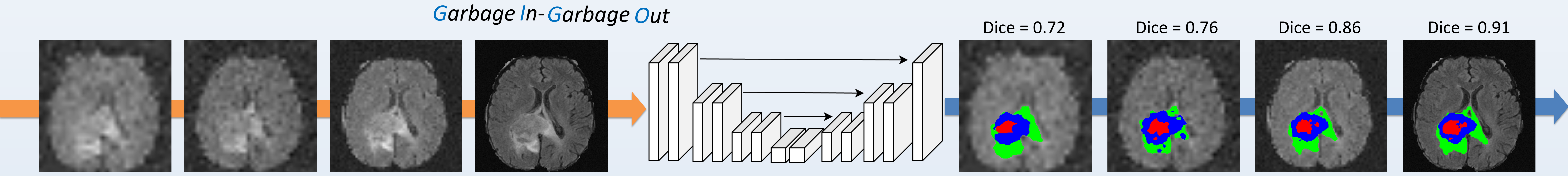
Towards a Global Quality Control for Medical Image segmentation

Benjamin Lambert[1,3], Florence Forbes[2], Senan Doyle[1] , Michel Dojat[2,3]

1. Pixyl, Research and Development Laboratory, Grenoble **2.** Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK **3.** Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut des Neurosciences

The GIGO principle in medical-image analysis

Garbage In- Garbage Out



AI algorithms often yield suboptimal predictions with poor-quality input images or those differing from the training set. Automating Quality Control (QC) is crucial as data volume increases, yet existing literature often separately addresses input and output QC. This study introduces a unified QC model, assessing both image quality and segmentation accuracy simultaneously. Leveraging Mahalanobis distance and Inter-model agreement, our approach categorizes predictions into four regimes: optimal, robust, dysfunctional, or divergent.

Input Quality Control

Goal: detect poor-quality input images

How: compute latent-space Mahalanobis distance [1]

- Feature maps are collected from the penultimate convolution layer. In our U-Net, they have a shape of 32xHxWxD for a 3D medical image x.
- To reduce the dimensionality of the feature map $\phi(x)$, a spatial averaging is performed, resulting in a 32-dimensional latent representation:

$$s(x) = \frac{1}{HWD}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{d=1}^{D}\phi(x)(h,w,d)$$

- From the training dataset, the mean (μ) and covariance matrix (Σ) of the in-distribution latent representations are computed.
- At test-time, we compute the Mahalanobis distance (MD) between the test latent representation z_{test} and the fitted moments:

$$\mathrm{MD}(z_{test};\mu,\Sigma) = (z_{test}-\mu)^T\Sigma^{-1}(z_{test}-\mu)$$

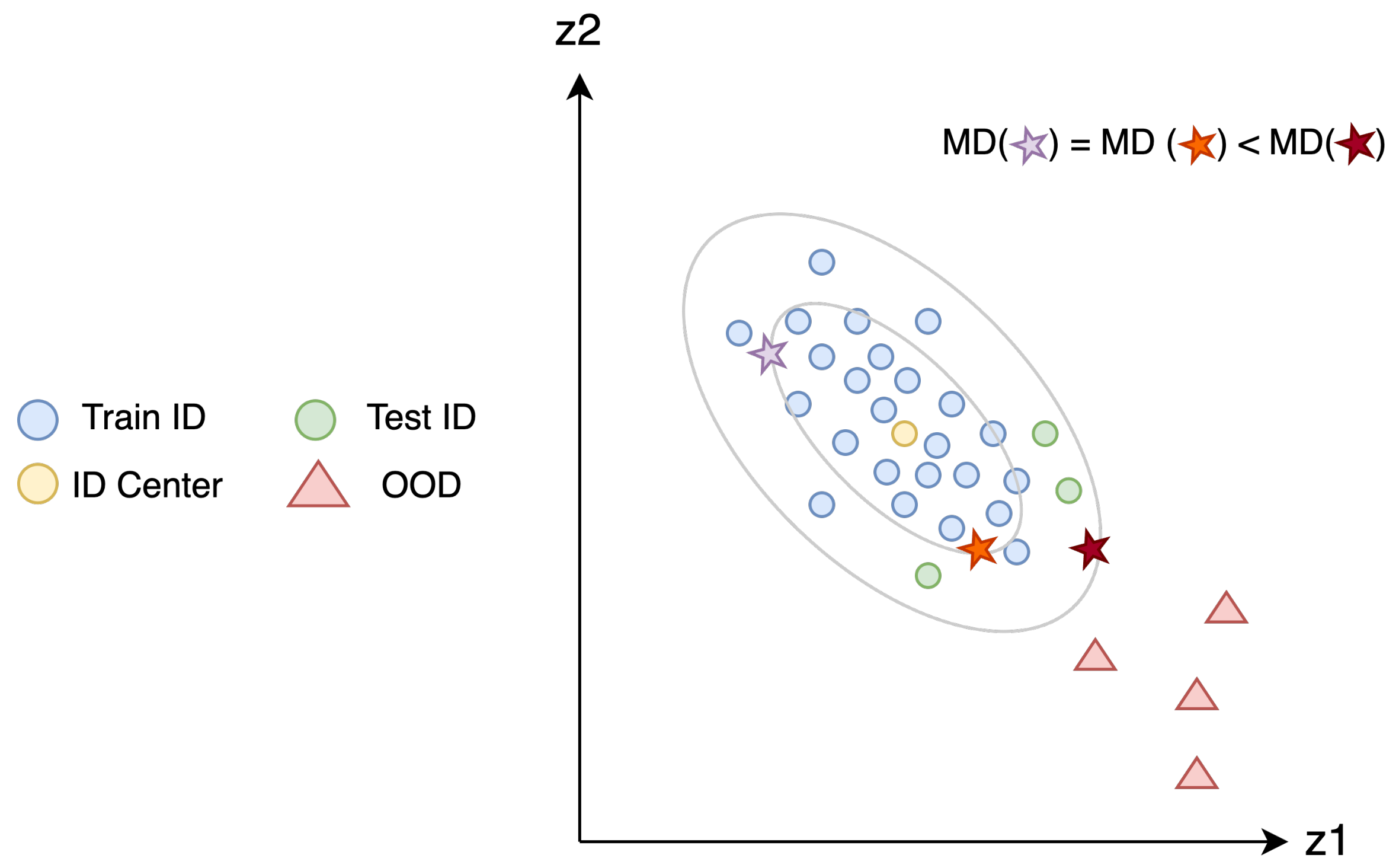


Illustration of the Mahalanobis distance in a two-dimensional setting.

Output Quality Control

Goal : detect poor-quality output segmentations

How: compute inter-model segmentation variability [2]

- An ensemble of 5 individually trained U-Nets is constructed. At test-time, each model produces a segmentation **Sk**, which are aggregated into a Majority Vote segmentation (**MV**).
- We compute the Dice score between each individual segmentation and the Majority Vote. The Ensemble Prediction Agreement (EPA) is then taken as the average of the Dice scores:

$$\mathrm{EPA} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{Dice}(S_k,\mathrm{MV})$$

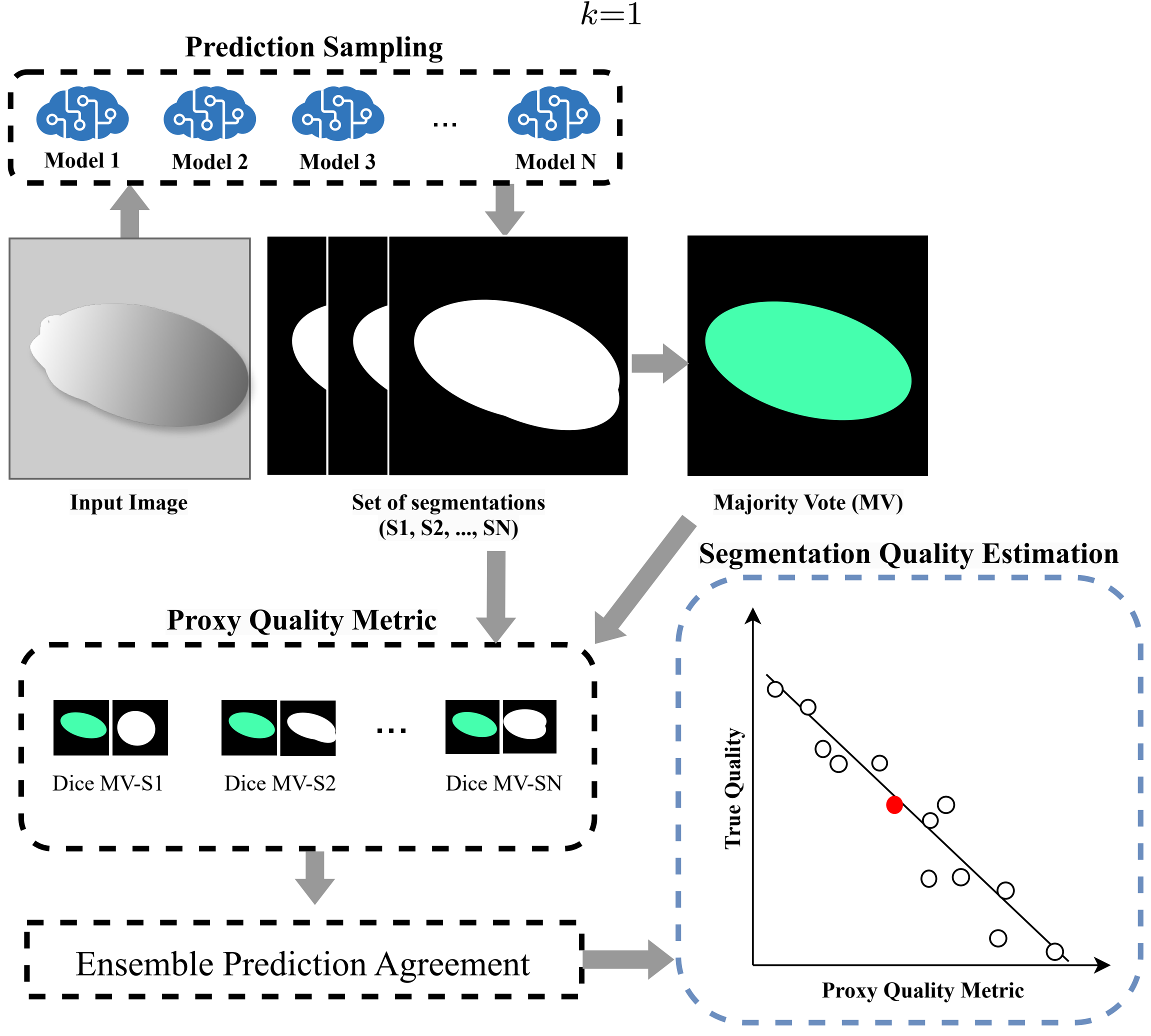
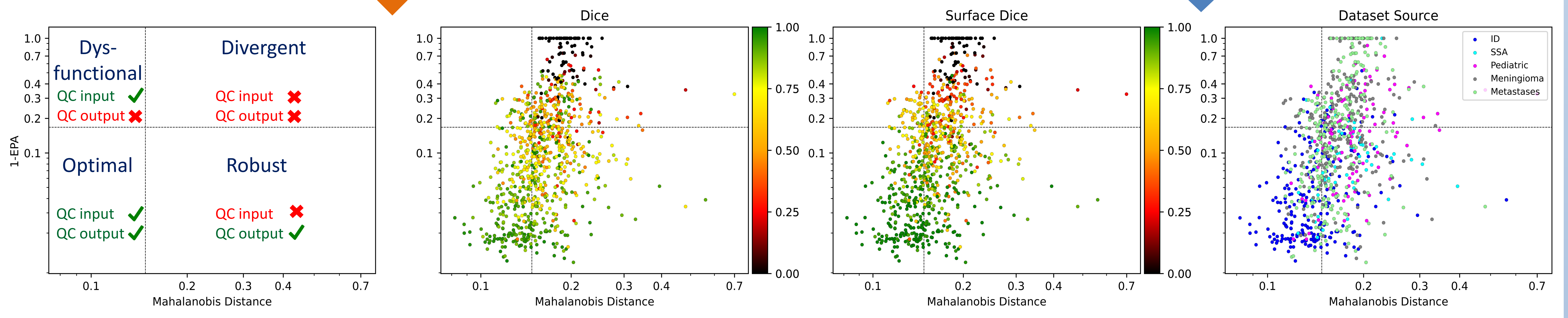


Illustration of the Ensemble Prediction Agreement computation.



Regime	Proportion	Dice	Surface Dice
Optimal	264/874 (30.20%)	0.828 ± 0.141	0.886 ± 0.152
Robust	400/874 (45.77%)	0.707 ± 0.206	0.732 ± 0.226
Dysfunctional	20/874 (2.29%)	0.678 ± 0.196	0.575 ± 0.151
Divergent	190/874 (21.74%)	0.334 ± 0.355	0.259 ± 0.264

Conclusion

- Efficient QC scores can be computed from trained DL models to evaluate the conformity of the input image and output segmentation.
- By combining the two scores, the prediction space can be stratified into 4 regimes of varying segmentation performance:

  Optimal > Robust > Dysfunctional > Divergent
- This enriched QC procedure can be used to alert the user if the input image is far from the training distribution and/or if the output segmentation does not meet predefined quality standards.

Methods & Materials

- In-distribution (ID) images correspond to adult subjects with glioblastomas (BraTS 2023 [3], 876 for training, 30 for validation, for 227 test).
- Four MRI sequences are provided: T1, T2, T1 with contrast-enhancement, FLAIR
- Out-of-distribution images correspond to auxiliary BraTS 2023 datasets [4-7]: Sub-Saharan Africa (SSA, 60 subjects), Pediatric (99 subjects), Meningioma (250 subjects), Metastases (238 subjects)
- The segmentation model is the Optimized U-Net [7] (16.5 million parameters) trained with a combination of the Dice and Cross-Entropy losses, using the ADAM optimizer with a learning rate of 2 x 10^{-4}.
- QC thresholds are determined on the validation dataset by computing the 95-th percentiles of the QC scores.

References

[1] Calli, Erdi et al. "FRODO: An in-depth analysis of a system to reject outlier samples from a trained neural network." IEEE Transactions on Medical Imaging 42, no. 4 (2022): pp. 971-981.
[2] Evan Hann et al. “Ensemble of deep convolutional neural networks with Monte Carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets”. Annual Conference on Medical Image Understanding and Analysis (2021), pp. 280–293.
[3] Menze, Bjoern H. et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." IEEE transactions on medical imaging 34, no. 10 (2014): pp. 1993-2024.
[4] Adewole, Maruf, et al. ”BraTS 2023: Glioma segmentation in sub-saharan africa patient population." ArXiv (2023).
[5] Moawad, Ahmed et al. ”BraTS 2023: Brain metastasis segmentation on pre-treatment MRI." ArXiv (2023).
[6] Kazerooni, Anahita Fathi et al. ”BraTS 2023: Focus on pediatrics.” ArXiv (2023).
[7] LaBella, Dominic et al. ”BraTS 2023: Intracranial meningioma." ArXiv (2023).
[8] Futrega, Michał et al. "Optimized U-Net for brain tumor segmentation." International MICCAI brainlesion workshop (2021), pp. 15-29.