# TECHNICAL REPORT FOR THE DEEP-PSMA CHALLENGE: TEAM CDS

**Jakob Dexl, Katharina Jeblick, Michael Ingrisch**
Department of Radiology, University Hospital, LMU Munich
Munich Center for Machine Learning (MCML)
jakob.dexl@lmu.de

August 21, 2025

**ABSTRACT**

This report presents the methodology of our algorithm submitted to the DEEP-PSMA (Deep-learning Evaluation for Enhanced Prognostics - Prostate Specific Membrane Antigen) challenge, which focuses on automated segmentation in Positron Emission Tomography (PET) imaging. We address the joint segmentation of PSMA and FDG PET/CT scans using a compact, multi-tracer UNet architecture. Our approach operates on the native resolution of each case and resolves spatial and resolution differences between tracers through a postprocessing fusion step. The training and inference code for our method is publicly available at https://github.com/JakobDexl/DEEP-PSMA.

***K*eywords** PET, PSMA, FDG, Segmentation

## 1 Introduction

The development of prostate-specific membrane antigen (PSMA) PET tracers has transformed the detection and treatment planning of metastatic castration-resistant prostate cancer (mCRPC), particularly in the context of radioligand therapy [1, 2]. Quantitative biomarkers derived from PSMA and FDG PET imaging have shown predictive value for treatment response but require accurate and reproducible segmentation of tumor burden [3, 4]. The DEEP-PSMA challenge [5], hosted at MICCAI 2025, was established to benchmark deep learning approaches for automated segmentation across both tracers, thereby enabling scalable computation of prognostic imaging biomarkers.

## 2 Methods

### 2.1 Datasets

The training dataset comprises 100 PET/CT cases staged prior to LuPSMA therapy. Each case includes both PSMA and FDG PET/CT imaging. For each tracer, the dataset provides a CT scan, a PET scan which is rescaled to standardized uptake value (SUV) units, a TotalSegmentator mask [6], a SUV threshold, a rigid registration to the complementary tracer, and a ground-truth segmentation mask. PET volumes were annotated using a fixed threshold (FDG>=(Liver Mean + 2*SD), PSMA>=3).

The organizers provide a preliminary validation set of 10 cases and a test set of 40 cases via GrandChallenge. All cases were manually curated to ensure diversity in anatomical size, tumor distribution, and challenging segmentation regions. Only cases with malignant lesions were included, with tumor burden varying across patients. Data acquisition was performed at a single institution using GE Discovery 710 and 690 PET/CT systems, Siemens Biograph PET/CT, and Siemens Vision 600 PET/CT scanners.

### 2.2 Model and Augmentations

Given the small size of the training dataset and the associated risk of overfitting, we employed a compact MONAI [7] DynUNet architecture, closely following an nnU-Net baseline configuration. Our goal was to develop a multimodal, multi-tracer model; therefore, we trained it using both FDG and PSMA PET volumes along with their corresponding TotalSegmentator masks as input. Modality fusion was achieved by concatenation, aiming to learn a shared embedding. Each output channel employed a sigmoid activation function to allow for overlapping segmentations. Model optimization was performed using a DiceCE loss function.

For preprocessing, images from different tracers were registered bidirectionally using the provided rigid transformations. Rather than resampling all cases to a common voxel spacing, the model was trained using the original image sizes and spacings. Each epoch included each case twice as independent inputs, one registered to PSMA and one registered to FDG. The rationale behind this was to always use the best possible resolution for each tracer while simultaneously providing a form of augmentation. Volumes were cropped to the union of tracer-specific fields of view,

| | FDG | | | | | PSMA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Fold | DSC ↑ | Surface DSC ↑ | FPV ↓ | FNV ↓ | ΔMTV ↓ | DSC ↑ | Surface DSC ↑ | FPV ↓ | FNV ↓ | $\Delta\text{SUV}_{mean}$ ↓ |
| 0 | 0.8946 | 0.8874 | 4.0099 | 5.4926 | -1.6483 | 0.9147 | 0.9006 | 2.9711 | 8.4309 | 3.1364 |
| 1* | 0.9607 | 0.9566 | 3.9342 | 1.8701 | -3.4942 | 0.9685 | 0.9564 | 9.3548 | 3.9083 | 3.2231 |
| 2 | 0.8785 | 0.8756 | 3.3217 | 3.6925 | 5.4419 | 0.9208 | 0.8979 | 7.1859 | 10.8616 | 1.6609 |
| 3 | 0.8541 | 0.8490 | 3.5961 | 11.6711 | 2.8829 | 0.9279 | 0.9086 | 3.7851 | 5.7240 | -1.6718 |
| 4 | 0.8402 | 0.8346 | 2.9521 | 3.2663 | 3.5562 | 0.9077 | 0.8880 | 9.6043 | 8.2704 | 7.0732 |

Table 1: Average metrics across five cross-validation folds. *Indicates the fold used for the final submission.

| | FDG | | | | | PSMA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | DSC ↑ | Surface DSC ↑ | FPV ↓ | FNV ↓ | ΔMTV ↓ | DSC ↑ | Surface DSC ↑ | FPV ↓ | FNV ↓ | $\Delta\text{SUV}_{mean}$ ↓ |
| Single | 0.9242 | 0.9077 | 1.9966 | 1.8809 | -11.6194 | 0.9466 | 0.9164 | 6.9941 | 4.0360 | 6.3460 |
| Single+ | 0.9661 | 0.9633 | 3.9441 | 1.8809 | -3.0470 | 0.9672 | 0.9571 | 12.3290 | 4.0334 | 3.1726 |
| Bothways* | 0.9607 | 0.9566 | 3.9342 | 1.8701 | -3.4942 | 0.9685 | 0.9564 | 9.3548 | 3.9083 | 3.2231 |
| Bothways+ | 0.9667 | 0.9630 | 5.4710 | 1.8627 | -0.2380 | 0.9672 | 0.9550 | 13.1942 | 3.9070 | 2.4109 |

Table 2: Postprocessing experiments for the model trained on fold 1. *Indicates the submitted final method.

defined by their respective thresholds, to improve inference speed. PET volumes were normalized by dividing by the provided annotation threshold and subsequently shifted by subtracting one. Input TotalSegmentator segmentation masks were divided by the maximum number of classes (117).

During training, images were cropped to $(256, 128, 128)$ with a 2/3 probability that the center contains tumor volume and a 1/3 probability of containing elevated physiological uptake. To improve generalization, a Trivial Augment data augmentation strategy was applied, including dimension-independent rotations up to $10°$, scaling in the range [0.9, 1.4], and random flipping across all axes. We also experimented with random tracer channel dropping and segmentation mask shrinking/growing based on gamma augmentation; however, no improvements were observed.

### 2.3 Training and evaluation

Models were trained for 500 epochs using the Adam [8] optimizer (learning rate = 2e-4) with a cosine annealing schedule. Training was performed on two A100 GPUs with an effective batch size of 4. The Dice Similarity Coefficient (DSC) for each tracer was the primary evaluation metric, while joint validation loss was used for model selection. False positive and false negative voxels were also monitored throughout training.

For all of our experiments, we make use of five-fold cross-validation. Final evaluation employed the challenge metrics, including DSC, Surface DSC, False Positive Volume (FPV), and False Negative Volume (FNV) per tracer, as well as the average error percentage of Metabolic Tumor Volume (ΔMTV) for FDG and the average deviation of $\Delta\text{SUV}_{mean}$ (as a percentage) for PSMA.

### 2.4 Experiments

We investigated four postprocessing strategies. The first (Single) served as a baseline, using only the predicted masks in the native space of each tracer. The second (Single+) applied the postprocessing method provided by the organizers, which expands the prediction by a certain radius, thresholds the mask based on the SUV, and optionally prevents mask growth in certain organs. High-uptake regions such as the liver, urinary tract, and kidneys were excluded in all experiments.

The third configuration (Bothways) combined the predictions for each tracer in its native resolution and space with the corresponding prediction transformed into the space of the other tracer. To achieve this, the binarized mask was resampled and registered back to its original space us-

| | FDG | | | | | PSMA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | DSC ↑ | Surface DSC ↑ | FPV ↓ | FNV ↓ | ΔMTV ↓ | DSC ↑ | Surface DSC ↑ | FPV ↓ | FNV ↓ | $\Delta\text{SUV}_{mean}$ ↓ |
| Single+ | 0.7963 | 0.7837 | 10.7518 | 4.7530 | 100.0000 | 0.8258 | 0.7972 | 25.2296 | 6.8963 | 14.7523 |
| Bothways* | 0.8069 | 0.7882 | 4.0329 | 5.7606 | 100.0000 | 0.8723 | 0.8404 | 6.3068 | 7.6993 | 14.4144 |
| Bothways+ | 0.7837 | 0.7723 | 13.3687 | 4.6844 | 100.0000 | 0.8154 | 0.7885 | 26.6467 | 6.8963 | 15.2873 |
| Ensemble++ | 0.8212 | 0.8121 | 7.8891 | 10.9953 | 100.0000 | 0.8200 | 0.7948 | 23.3158 | 6.5786 | 15.0503 |

Table 3: Results of different configurations on the preliminary validation set (N = 10). Note that the Metabolic Tumor Volume (MTV) values are incorrect due to an error in the validation container. *Denotes the final submitted method.

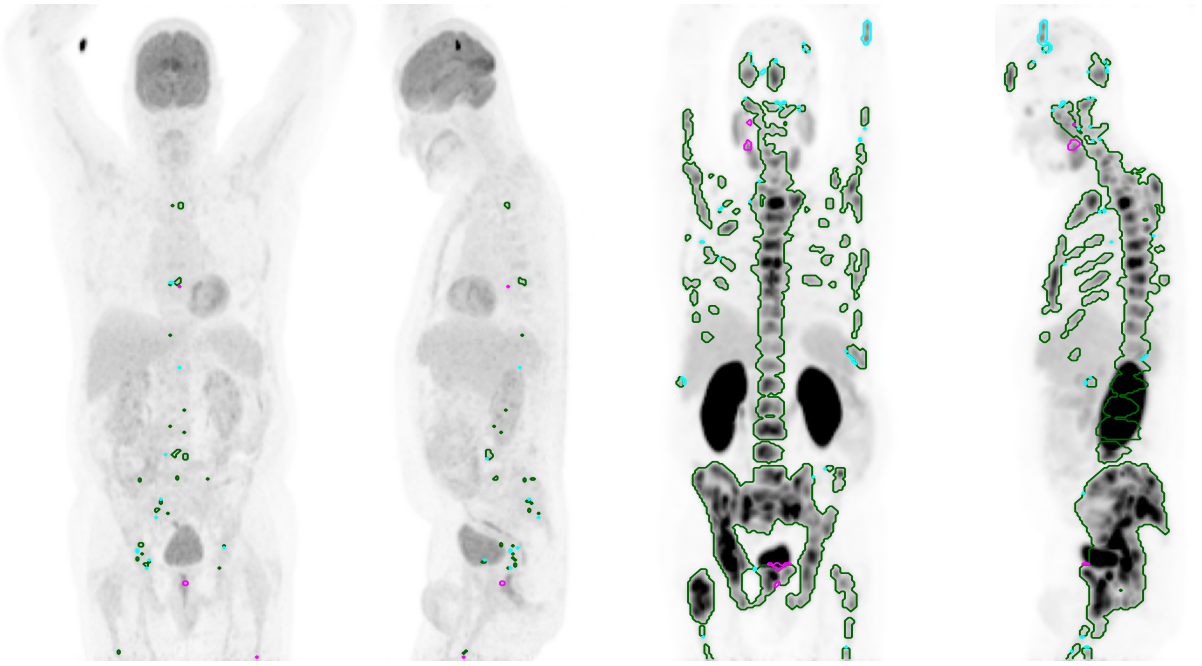

Figure 1: Maximum intensity projections for FDG (left) and PSMA (right) with predictions from our submitted model (case ID 0006). Green contours indicate True Positives, cyan False Negatives, and magenta False Positives. The sample achieved a DSC of 0.88/0.98, FPV of 0.4/1.9 ml, and FNV of 0.3/5.1 ml for FDG/PSMA, respectively.

ing linear interpolation, which unexpectedly outperformed nearest-neighbor interpolation. The binarized union of both volumes was then used. The fourth configuration (Bothways+) combined the Bothways method with the region-growing postprocessing.

We submitted Single+, Bothways, and Bothways+ configurations to the preliminary evaluation. Additionally, an ensemble of all folds in the Bothways+ configuration was submitted.

### 2.5 Results

Table 1 shows results for all cross-validation folds using the Bothways configuration. Fold 1 has a reasonably better performance than the others, and average DSC scores vary by up to 10 percent. However, it is worth mentioning that all fold models had a median DSC above 0.94 for both tracers. Bothways averaging improved DSC, Surface DSC, and error percentages at the cost of a slight increase in FPV, as shown in Table 2. Adding region-growing postprocessing to the baseline produced a similar effect. Its combination with Bothways further increased FPV. Table 3 summarizes our preliminary test submissions. Bothways was selected as the final submission.

## 3 Discussion

Our goal was to train and investigate a single model capable of handling both tracers simultaneously. Cross-validation suggests strong average performance, but results on the preliminary validation set were lower. The final approach was selected based on preliminary validation, although we are well aware that the small sample size (10 cases) limits statistical power for model selection.

We ultimately chose the Bothways model due to its lowest combined FPV and FNV, even though this configuration may not be optimal for challenge ranking. Since most evaluation metrics are highly correlated, prioritizing a model with a higher DSC could have provided a better trade-off.

An important design choice was to operate primarily in the native space of each case. Combined with the lower-bound SUV threshold, this strategy emphasizes mask accuracy and mitigates errors introduced by resampling and spatial transformations. This may partly explain why single-tracer models remain competitive, as avoiding multi-tracer alignment can reduce compounding sources of error.

It is important to note that the proposed algorithms may have limited generalizability beyond the challenge setting. Cross-validation relied on only 20 samples, meaning that a small number of cases can disproportionately influence reported performance metrics. Additionally, the curated nature of the dataset introduces systematic bias. While using a lower-bound SUV threshold simplifies annotation consistency and problem formulation, it has drawbacks compared to approaches in other challenges (e.g., HECKTOR [9] or AutoPET[10]), which primarily leverage morphological CT-based annotations. In addition, low-intensity lesions commonly observed post-therapy may require further methodological considerations.

## References


[1] Hossein Jadvar et al. "Appropriate Use Criteria for Prostate-Specific Membrane Antigen PET Imaging". en. In: *Journal of Nuclear Medicine* 63.1 (Jan. 2022). Publisher: Society of Nuclear Medicine Section: Appropriate Use Criteria, pp. 59–68. ISSN: 0161-5505, 2159-662X. DOI: 10.2967/jnumed.121.263262.

[2] Michael S. Hofman et al. "Prostate-specific membrane antigen PET-CT in patients with high-risk prostate cancer before curative-intent surgery or radiotherapy (proPSMA): a prospective, randomised, multicentre study". English. In: *The Lancet* 395.10231 (Apr. 2020). Publisher: Elsevier, pp. 1208–1216. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(20)30314-7.

[3] James P. Buteau et al. "PSMA and FDG-PET as predictive and prognostic biomarkers in patients given [177Lu]Lu-PSMA-617 versus cabazitaxel for metastatic castration-resistant prostate cancer (TheraP): a biomarker analysis from a randomised, open-label, phase 2 trial". English. In: *The Lancet Oncology* 23.11 (Nov. 2022). Publisher: Elsevier, pp. 1389–1397. ISSN: 1470-2045, 1474-5488. DOI: 10.1016/S1470-2045(22)00605-2.

[4] Justin Ferdinandus et al. "Prognostic biomarkers in men with metastatic castration-resistant prostate cancer receiving [177Lu]-PSMA-617". en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 47.10 (Sept. 2020), pp. 2322–2327. ISSN: 1619-7089. DOI: 10.1007/s00259-020-04723-z.

[5] Price Jackson et al. "Deep-learning Evaluation for Enhanced Prognostics - Prostate Specific Membrane Antigen". In: (Mar. 2025). Publisher: Zenodo.

[6] Jakob Wasserthal et al. "TotalSegmentator: robust segmentation of 104 anatomical structures in CT images". In: *Radiology: Artificial Intelligence* 5.5 (Sept. 2023). 49 citations (Crossref) [2024-02-08] arXiv:2208.05868 [cs, eess], e230024. ISSN: 2638-6100. DOI: 10.1148/ryai.230024.

[7] M. Jorge Cardoso et al. *MONAI: An open-source framework for deep learning in healthcare*. MONAI. Nov. 2022. DOI: 10.48550/arXiv.2211.02701.

[8] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980 [cs]. Jan. 2017. DOI: 10.48550/arXiv.1412.6980.

[9] Vincent Andrearczyk et al. "Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images". en. In: *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by Vincent Andrearczyk et al. HECKTOR? Cham: Springer International Publishing, 2022, pp. 1–37. ISBN: 978-3-030-98253-9. DOI: 10.1007/978-3-030-98253-9_1.

[10] Sergios Gatidis et al. *The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging*. en. preprint. In Review, June 2023. DOI: 10.21203/rs.3.rs-2572595/v1.