

Cross-Validation Report: nnU-Net Baseline on AMOS22 Multi-Organ Segmentation

This document summarizes the results of a five-fold cross-validation experiment using nnU-Net for multi-organ abdominal segmentation on the AMOS22 dataset. The objective of this experiment was to establish a reproducible institutional baseline model and training protocol for future research and architectural comparisons.

Overall foreground mean Dice coefficient across folds was **0.8726 ± 0.0060**, indicating strong segmentation performance with low inter-fold variability.

Fold	Foreground Mean Dice
Fold 0	0.8658
Fold 1	0.8664
Fold 2	0.8741
Fold 3	0.8823
Fold 4	0.8741
Mean ± SD	0.8726 ± 0.0060

Methods Overview

Training utilized nnU-Net's automated configuration pipeline with a 3D full-resolution architecture. Five-fold cross-validation was performed on 240 labeled CT volumes, with each fold using approximately 80% of the data for training and 20% for validation. Standard nnU-Net preprocessing included resampling, intensity normalization, patch-based sampling, and extensive on-the-fly augmentation.

Scientific Interpretation

The achieved Dice performance (~0.87) is consistent with strong baseline performance for large-scale multi-organ segmentation tasks. Low variance across folds suggests dataset balance and stable optimization behavior. These findings support the suitability of this model as a reference benchmark for subsequent experimentation, including architectural modifications and MONAI-based pipelines.

Limitations

This experiment reflects baseline nnU-Net configuration without task-specific customization. Future work may evaluate ensemble strategies, region-based training, or transformer-based architectures to determine potential performance gains.