Probabilistic nnU-Net for Multi-Class Brain Hemorrhage Segmentation: Addressing Inter-Observer Variability in Clinical Practice

Anonymized Authors[1]

[1] Anonymized Affiliations
email@anonymized.com

1 Methods

Our proposed Conditional Probabilistic nnU-Net extends the nnU-Net architecture to model segmentation variability in multi-annotator datasets by integrating a conditional variational autoencoder (CVAE) into the U-Net's latent space and decoding pathway.

1.1 Baseline Architecture: nnU-Net

We build upon nnU-Net, which automatically configures preprocessing, network architecture, and post-processing based on dataset properties. Our probabilistic extension retains these automated capabilities while extending the deterministic baseline to generate diverse segmentations.

1.2 Probabilistic Latent Space Modeling

To capture inter-observer variability, we introduce a low-dimensional latent vector $\mathbf{z}$ that represents the unique stylistic characteristics of each annotator. The model learns a conditional distribution $p(z|\mathbf{X}, s)$, where $\mathbf{X}$ is the input image and s is the annotator ID. This is achieved through a CVAE framework composed of prior and posterior networks.

Prior Network: The prior network q_φ $(z|\mathbf{X}, s)$ predicts the distribution of styles for a given annotator s. It takes the U-Net's bottleneck feature map concatenated with a one-hot encoded annotator ID vector s as input, then outputs the parameters (mean μ_{prior} and log-variance $log(\sigma^2)_{prior}$) of a diagonal Gaussian distribution. During inference, we sample $\mathbf{z}$ from this learned prior $N(\mu_{prior}, \sigma^2_{prior})$ to generate segmentations in the style of annotator s.

Posterior Network: The posterior network p_θ $(z|\mathbf{X}, \mathbf{Y}_s)$ is used only during training to guide latent space learning. It takes the bottleneck feature map and the corresponding ground truth segmentation $\mathbf{Y}_s$ from annotator s as input. Its purpose is to estimate the "ideal" latent distribution for reconstructing that specific ground truth, outputting parameters ($\mu_{posterior}$, $log(\sigma^2)_{posterior}$) of the posterior distribution.

Both networks use lightweight 1×1 convolutional layers to map features to latent distribution parameters.

1.3 Hierarchical Latent Vector Injection

A key contribution is our method for modulating segmentation output through the sampled latent vector **z**. Instead of single injection at the bottleneck, we employ a hierarchical conditioning scheme across multiple decoder scales. First, **z** is passed through a 1×1 convolutional layer and added to the bottleneck feature map, controlling global segmentation structure. Subsequently, **z** is injected into skip-connection pathways at each decoder stage through separate 1×1 convolutions that match the corresponding feature dimensions. This multi-scale injection ensures annotator style is reflected at both global structural levels and local boundary details.

1.4 Annotator Augmentation Strategy

While this challenge provided five annotators, we created two additional virtual annotators to enhance stylistic diversity. The sixth annotator represents a larger segmentation tendency using the union of all five existing annotations, while the seventh represents a conservative tendency using their intersection.

1.5 Loss Function

The network is trained end-to-end by optimizing a composite loss function L_{total}:

$$L_{total} = L_{recon} + \beta L_{KL} \tag{1}$$

Reconstruction Loss (L_{recon}): This term measures the fidelity of generated segmentation to ground truth. It is a weighted sum of batch-wise Dice loss and Cross-Entropy loss, identical to the standard nnU-Net loss function. To emphasize precise boundary delineation, we increased the Cross-Entropy component weight (*weight_ce* = 2).

KL Divergence Loss (L_{KL}): This regularizes the latent space by minimizing the Kullback-Leibler divergence between the prior distribution $q_\varphi(z|\mathbf{X}, s)$ and posterior distribution $p_\theta(z|\mathbf{X}, \mathbf{Y}_s)$. This forces the prior network to learn meaningful latent distributions conditioned on annotator ID, without requiring ground truth segmentation as input. The hyperparameter β balances both terms with annealing.

References

1. Kohl, Simon, et al. "A probabilistic u-net for segmentation of ambiguous images." Advances in neural information processing systems 31 (2018).
2. Wu, Yicheng, et al. "Diversified and personalized multi-rater medical image segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
3. Pu, Yunchen, et al. "Variational autoencoder for deep learning of images, labels and captions." Advances in neural information processing systems 29 (2016).