Latent Space Oversampling for Addressing Adult–Pediatric Imbalance in Brain Tumor Segmentation

Saeideh Ghanbari Azar[1] SAEIDEH.GHANBARI@UMU.SE
Tufve Nyholm[2] TUFVE.NYHOLM@UMU.SE
Tommy Löfstedt[1] TOMMY.LOFSTEDT@UMU.SE
[1] *Department of Computing Science, Umeå University, Umeå, Sweden*
[2] *Department of Diagnostics and Intervention, Biomedical Engineering and Radiation Physics, Umeå University, Umeå, Sweden*

Editors: Under Review for MIDL 2026

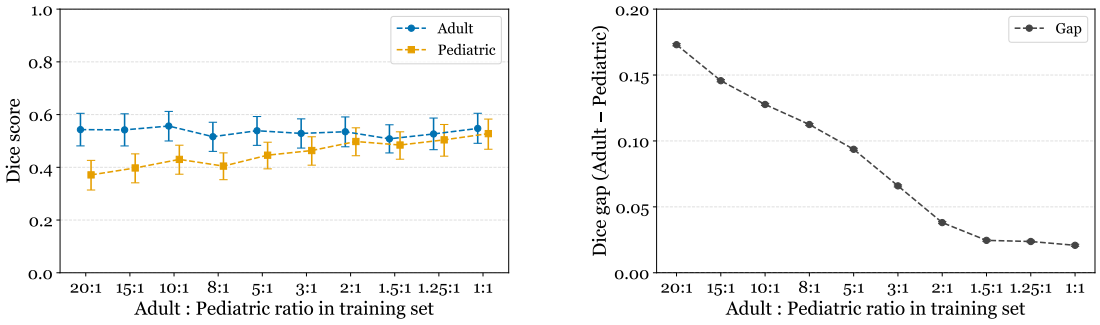
Abstract

Deep-learning models tend to perform unevenly when there are imbalances in training sample subgroups—*e.g.*, in the number of adult and pediatric training samples in brain tumor segmentation. There are often fewer pediatric scans and they differ from adult scans in anatomy, contrast, and tumor characteristics. In this work, we studied if enriching the pediatric cohort with realistic synthetic samples would resolve this imbalance. Specifically, we sought to know if latent space minority oversampling methods would resolve the imbalance and whether the quality of the latent space would make a difference in this application. We first constructed an adult–pediatric dataset by unifying the BraTS and BraTS-PEDs datasets, making them compatible through consistent preprocessing and labeling. We then developed a cohort-conditioned StyleGAN2 model to jointly model multi-modal MRI slices and tumor masks. Pediatric slices were embedded into the generator's latent space and, using the Synthetic Minority Over-sampling Technique (SMOTE), new pediatric latent vectors were produced. These new latent vectors were decoded into MRI–mask sets and added to the training set to balance the adult–pediatric counts. This proposed latent space oversampling strategy was compared to several imbalance-mitigation baselines. Evaluations on a balanced test set of 200 adult and 200 pediatric subjects showed that the proposed latent space oversampling improves pediatric Dice scores without decreasing the adult performance and obtains the smallest adult–pediatric performance gap of all evaluated methods.

Keywords: brain tumor segmentation, latent space oversampling, adult–pediatric imbalance, StyleGAN2, SMOTE

1. Introduction

Deep learning has become the main approach for medical image segmentation, but its performance is often limited by data imbalances. Imbalances can appear both at the label level (*e.g.*, some classes are rare) or at the subgroup level, where certain populations (*e.g.*, age groups, scanners, or institutions) have fewer samples and act as the *minority* group, while the larger, well-represented populations form the *majority* group (Chen et al., 2024). Standard deep neural networks tend to focus on patterns from the majority group leading



(a) (b)

Figure 1: The effect of adult:pediatric imbalance in the training set. (a) Test Dice scores with error bars showing 95% bootstrap confidence intervals. (b) Dice gap computed as Adult − Pediatric. The gap decreases rapidly until the ratio approaches 1.5:1, after which further balancing has only a small effect.

to weaker performance on minority cases, which directly impacts generalization and thus reliability in clinical applications.

In this study, we focused on the practical scenario where adult and pediatric brain tumor cases coexist in the same clinical pipeline, and a single segmentation model is expected to handle both. In such settings, pediatric cases are typically far less common and differ from adult tumor images in their noise distribution, appearance, anatomy, and underlying biology. This pattern appears in public datasets: for instance, the BraTS collection includes hundreds of annotated adult glioma cases (Baid et al., 2021; de Verdier et al., 2024), whereas pediatric data are much less common. A pediatric collection was recently released in a standardized format through BraTS-PEDs (Kazerooni et al., 2024; Fu et al., 2024).

To illustrate the effect of such a cohort subgroup imbalance, Figure 1(*a*) and Figure 1(*b*) show how segmentation performance changes as the proportions of pediatric cases in the training data vary. For this illustration, we constructed a unified adult–pediatric dataset by combining the BraTS and BraTS-PEDs cohorts (details in Section 3.1), and generated several training sets with different adult to pediatric ratios while keeping the total training set size fixed and the test set balanced. As the training set becomes more balanced, pediatric Dice scores improve, the adult–pediatric gap shrinks, and the adult scores remain unaffected. This trend indicates that the performance gap arises largely from cohort imbalance, and that enriching the pediatric cohort with additional representative samples has the potential to reduce this gap.

To address this generalization gap problem, we propose a cohort-conditioned generative augmentation framework. A conditional StyleGAN2 model (Karras et al., 2020) was trained on 2D MRI slices (MRI–mask sets with four MRI sequences and a segmentation mask) using the cohort subgroup label (adult or pediatric) as a conditioning variable. Real pediatric slices were inverted into the StyleGAN2 latent space, where new latent vectors were synthesized using Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002).

These synthetic latent vectors were then decoded using the StyleGAN2 into realistic pediatric MRI–mask sets aimed to enrich the minority cohort. The augmented and now balanced dataset was then used to train a downstream U-Net segmentation model. The contributions of this work are threefold:

- A unified adult–pediatric dataset, harmonized from the BraTS and BraTS-PEDs datasets, where the label definitions have been aligned.
- A cohort-conditioned generative oversampling approach in which multi-modal MRI–mask sets are synthesized through latent space SMOTE interpolation using a StyleGAN2 model.
- An empirical comparison of standard imbalance-mitigation strategies, showing that the proposed latent space minority oversampling improves pediatric segmentation performance while maintaining accuracy on adult cases.

2. Related Work

Imbalanced learning has been studied widely in machine learning, and many methods have been proposed to handle the gap between majority and minority data (Chen et al., 2024). Sampling-based approaches operate by modifying how often different samples appear during training. This includes methods such as random oversampling of minority cases, undersampling of majority cases, or building stratified batches that have a desired minority–majority ratio (Bengio et al., 2009). Loss-reweighting methods instead modify the optimization objective so that errors on minority samples contribute more strongly to the training signal (He and Garcia, 2009). In medical image segmentation, these methods have been applied, for example, to address rare tumor subregions in the BraTS datasets (Nalepa et al., 2019).

Data augmentation is an approach that ranges from pixel-space transformations, such as rotations, flips, and intensity perturbations commonly used in tumor segmentation (Nalepa et al., 2019), to more advanced strategies based on convex combinations of existing samples, such as SMOTE (Chawla et al., 2002). Unlike classical augmentation, SMOTE produces new samples by interpolating between real instances, which works well for tabular data but is less effective for unstructured data such as images. Generative models provide another approach for augmentation and enable the synthesis of new images. GAN-based augmentation has shown improvements under limited-data regimes across MRI, CT, and X-ray modalities (Makhlouf et al., 2023). GANs have also been explored in tasks such as lesion segmentation, polyp segmentation, and brain tumor synthesis (Zhou et al., 2025). Diffusion models have recently been explored for fairness-oriented augmentation (Ktena et al., 2024).

Another important line of work focuses on latent space augmentation, where SMOTE-like mixing is applied to learned feature representations instead of raw pixels. Mondal et al. (2023) demonstrated that convex combinations of latent vectors can preserve semantic structure and outperform classical SMOTE for high-dimensional data. DeepSMOTE (Dablain et al., 2022) is a prominent example that combines an autoencoder with SMOTE-style interpolation in latent space. However, the effectiveness of DeepSMOTE is limited, partly due to issues with the penalty term used during autoencoder training. Recent analyses

have shown that this penalty can adversely affect the latent space (Azar et al., 2025), even though the underlying idea of latent space SMOTE remains promising.

Motivated by these observations, we investigated latent space oversampling for brain tumor segmentation. While DeepSMOTE combines an autoencoder with SMOTE-based interpolation, it is not directly applicable here since the associated penalty term has been shown to adversely affect latent diversity and downstream results (Azar et al., 2025). Further, DeepSMOTE uses an autoencoder to construct the latent space for oversampling, which we hypothesized may not be expressive enough in more complicated scenarios (Azar et al., 2025), such as when representing multi-modal MRI data together with spatially consistent masks. These issues suggest that DeepSMOTE's limitations lie not necessarily in latent space SMOTE itself, but in the model formulation and the underlying representation used for SMOTE interpolation. As a reference baseline to test this hypothesis, we therefore evaluated an autoencoder-based latent space SMOTE variant, but without the defunct DeepSMOTE penalty (denoted AE-SMOTE here).

To obtain a more expressive latent representation for large-scale MRI synthesis, we thus propose and evaluate a cohort-conditioned StyleGAN2 model and perform SMOTE-style interpolation directly in its latent space. The workflow involves the following steps: train a conditional StyleGAN2 model on adult and pediatric data, invert real pediatric slices into the StyleGAN2 latent space, apply SMOTE-style interpolation to generate new latent vectors, and decode these vectors back into synthetic pediatric MRI–mask sets. Finally, add these synthesized samples to the training set to enrich the minority cohort and encourage a more balanced downstream segmentation performance across adult and pediatric cases.

3. Methods

3.1. Data and Preprocessing

We constructed a unified adult–pediatric dataset by combining the adult BraTS Glioma 2023 dataset (Baid et al., 2021; de Verdier et al., 2024) with the pediatric BraTS-PEDs 2025 dataset (Kazerooni et al., 2024). The adult BraTS Glioma dataset comprises pre-treatment diffuse gliomas in adults, whereas BraTS-PEDs contains pediatric high-grade gliomas such as diffuse midline glioma (DMG) and diffuse intrinsic pontine glioma (DIPG), which differ in biology, anatomy, and imaging appearance. To make the two datasets compatible, the pediatric MRIs were skull-stripped using the same preprocessing pipeline applied to the adult BraTS data (Vossough et al., 2024; Gandhi et al., 2024; Isensee et al., 2019).

Although both datasets include voxelwise tumor annotations, their label definitions and prevalence differ. All labels were therefore remapped into a shared scheme with three categories: enhancing tumor (ET), non-enhancing or cystic core (NET/CC), and peritumoral edema (ED). To avoid conflating the large cohort imbalance between adults and pediatrics with an additional label imbalance, which would arise because certain tumor subregions are far less common (*e.g.*, ET, ED) or even absent in pediatric gliomas (Fu et al., 2024), we also simplified the task to be binary tumor-core segmentation. This ensured that differences in performance reflect true adult–pediatric domain effects rather than disparities in tumor

subregion frequency. Accordingly, the ET and NET/CC voxels were assigned to a tumor core class, while edema and background were merged into a single background class.

For each subject, we extracted 2D axial slices from the four MRI modalities and the remapped segmentation masks. Slices were max-normalized by their maximum intensity and zero-padded to a spatial resolution of 256×256 pixels. To study adult–pediatric cohort imbalance under controlled evaluation, we defined fixed subject-wise splits: the training set contained 400 adult and 40 pediatric subjects, while both the validation and test sets were balanced, with 20 adults and 20 pediatrics in the validation and 200 adults and 200 pediatrics in the test set. Code and instructions for reproducing this dataset is available at https://github.com/SG-Azar/brats-adult-pediatric-imbalance.

3.2. Latent Space Pediatric Data Augmentation

The proposed balancing strategy has two stages: (i) learning a cohort-conditioned StyleGAN2 model that jointly synthesizes multi-modal MRI and tumor masks, and (ii) performing SMOTE-style oversampling in the resulting latent space to generate additional pediatric samples. The synthesized pediatric slices are then added to the training set to balance it, while the validation and test sets remain unchanged and contain only real images.

Cohort-conditioned StyleGAN2. We trained a conditional StyleGAN2 generator (Karras et al., 2020) on the five-channel slices described in Section 3.1, where each slice consisted of four MRI modalities and a binary tumor mask, all resampled to 256×256 pixels to match the resolution expected by the StyleGAN2 architecture. Each slice was treated as an independent training example and associated with a binary cohort label (*adult* or *pediatric*). We adapted the official StyleGAN2-ADA implementation[1] to support five-channel inputs (four MRI modalities and one mask), instead of the standard three-channel RGB configuration. The cohort label was embedded and provided to the mapping network, allowing the generator to learn a conditional distribution over $(\text{MRI}, \text{mask} \mid \text{cohort})$, where the cohort thus was adult or pediatric. Training followed the standard StyleGAN2-ADA setup, using adaptive discriminator augmentation (ADA) with the `bg` augmentation pipeline (geometric transformations such as flips, rotations and translations, but no intensity or color changes). During training, we monitored the Fréchet Inception Distance (FID) and selected the checkpoint with the lowest FID for latent inversion and sample generation. By modelling images and masks jointly, the generator produced synthetic slices in which tumor masks remained spatially consistent with the corresponding MRI appearance.

Latent inversion of pediatric slices. After training the generator, all real pediatric training slices were embedded into the StyleGAN2 latent space to obtain latent vectors that approximate the pediatric manifold. For this, we followed an optimization-based inversion scheme in the extended W^+ space, as proposed by Karras et al. (2020). Compared to the original StyleGAN2 projector, which was designed for RGB images, we adapted the inversion objective to the five-channel setting here (four MRI sequences and a tumor mask). Specifically, we applied the VGG16-based perceptual loss only to the MRI channels and

1. https://github.com/NVlabs/stylegan2-ada-pytorch

added a light Dice term (weight 0.01) on the mask channel to encourage spatial agreement. This resulted in one pediatric W^+ latent vector per real pediatric five-channel slice.

SMOTE-style oversampling in latent space. To synthesize new pediatric samples, we applied SMOTE (Chawla et al., 2002) directly in the W^+ latent space. All latent vectors obtained from real pediatric slices were flattened and used to fit a k-nearest-neighbor graph ($k = 5$). To generate a synthetic sample, $\mathbf{w}_{\text{syn}}$, we selected a real pediatric latent vector, $\mathbf{w}_i$, randomly chose one of its k nearest neighbors, $\mathbf{w}_j$, and formed the convex combination

$$\mathbf{w}_{\text{syn}} = \lambda \mathbf{w}_i + (1 - \lambda)\, \mathbf{w}_j, \quad \text{where} \quad \lambda \sim \mathcal{U}(0, 1). \tag{1}$$

The resulting latent vector, $\mathbf{w}_{\text{syn}}$, was reshaped back into the W^+ format and passed through the StyleGAN2 synthesis network, with the pediatric conditioning label, to generate a five-channel slice. We generated as many synthetic pediatric slices as needed to match the number of adult slices in the training set, such that the augmented train set was fully balanced, *i.e.*, the adult:pediatric ratio was 1:1.

3.3. Segmentation Model

For the downstream task we trained a 2D U-Net for tumor-core segmentation on axial slices, using the four MRI modalities as input channels and a single binary tumor-core label as output. The architecture was adapted from a standard U-Net (Ronneberger et al., 2015), with design choices similar to common nnU-Net 2D configurations (Isensee et al., 2021). All models were trained with a loss consisting of a cross-entropy term and a soft Dice term weighted equally. The network was trained using AdamW (with weight decay $3{\times}10^{-5}$) and a cosine annealing learning-rate schedule over a maximum of 400 epochs. Before training, we performed an automated learning-rate range test (Smith, 2017), sweeping the learning rate logarithmically from 10^{-6} to 10^{-2} over roughly 200 optimization steps, and used the rate suggested by this procedure as the initial learning rate for all final experiments. Early stopping on the validation mean Dice score was used (with a patience of 30 epochs), and the checkpoint with the best validation Dice was used for test-time evaluation.

4. Experiments

4.1. Training and Evaluation Setup

All methods were trained on the same adult–pediatric dataset using the same U-Net architecture and training protocol described in Section 3.3. The validation and test sets were the same in all experiments. The competing methods are listed below and they differ only in how they address the adult–pediatric imbalance during training:

Baseline. A U-Net is trained on the original imbalanced training set. It serves as a reference for the adult–pediatric performance gap.

Cohort-weighted loss (WLoss). We reweighted the contribution of adult and pediatric slices in the loss. For each mini-batch we computed separate losses L_{adult} and $L_{\text{pediatric}}$ and

combined them as

$$L = \alpha L_{\text{pediatric}} + (1 - \alpha) L_{\text{adult}},$$

where α is a pediatric weight derived from the cohort slice counts as

$$\alpha = \frac{\sqrt{n_{\text{adult}}}}{\sqrt{n_{\text{adult}}} + \sqrt{n_{\text{pediatric}}}},$$

giving pediatric slices a relatively larger contribution in the loss (He and Garcia, 2009).

Stratified batch sampling (SBS). Here, the cohorts were balanced at the batch level (Buda et al., 2018). For that the training slices were divided into adult and pediatric pools and sampled for the batches such that each mini-batch had a 50/50 adult–pediatric mix by oversampling pediatric slices when their pool was exhausted.

Pediatric-only data augmentation (Aug). This method augments only the pediatric cohort using standard augmentations (Isensee et al., 2021). For pediatric slices, we applied on-the-fly perturbations (each applied with probability 0.5): random horizontal flips, random gamma intensity transform on MRI channels, and small Gaussian noise.

Image-space SMOTE (ImgSMOTE). We applied SMOTE (Chawla et al., 2002) directly in the five-channel image space. Synthetic slices were added to the train set until adult and pediatric counts were equal.

StyleGAN2-based pediatric oversampling (SG2). We used the trained cohort-conditioned StyleGAN2 model described in Section 3.2 to generate synthetic pediatric slices (Makhlouf et al., 2023). Those MRI–mask sets were added to the real training set until the adult and pediatric slice counts were equal.

Autoencoder latent SMOTE (AE-SMOTE). We also evaluated the latent oversampling strategy using an autoencoder inspired by DeepSMOTE (Dablain et al., 2022). However, in this method we trained the autoencoder without the DeepSMOTE penalty, which has been shown to be detrimental (Azar et al., 2025). The autoencoder takes five-channel slices as input and has five encoder blocks (Conv2d-BatchNorm-LeakyReLU) followed by a fully connected layer producing a latent vector. The decoder mirrors this structure with five blocks, each consisting of an upsampling layer followed by a convolution. The model was trained using an ℓ_2 reconstruction loss on the four MRI channels and a combination of ℓ_2 and soft Dice loss on the mask channel. After training, the autoencoder was frozen and all pediatric slices encoded into latent vectors. SMOTE was then applied in this latent space to generate synthetic pediatric codes. Each synthetic code was passed through the fixed decoder to reconstruct a five-channel slice. These reconstructed synthetic slices were added to the train set until the adult:pediatric ratio was 1:1.

StyleGAN2 latent SMOTE (SG-SMOTE). This is the proposed method described in Section 3.2.

Evaluation metrics. All methods were evaluated on the fixed balanced test set of 200 adult and 200 pediatric subjects, assessing performance on the downstream segmentation

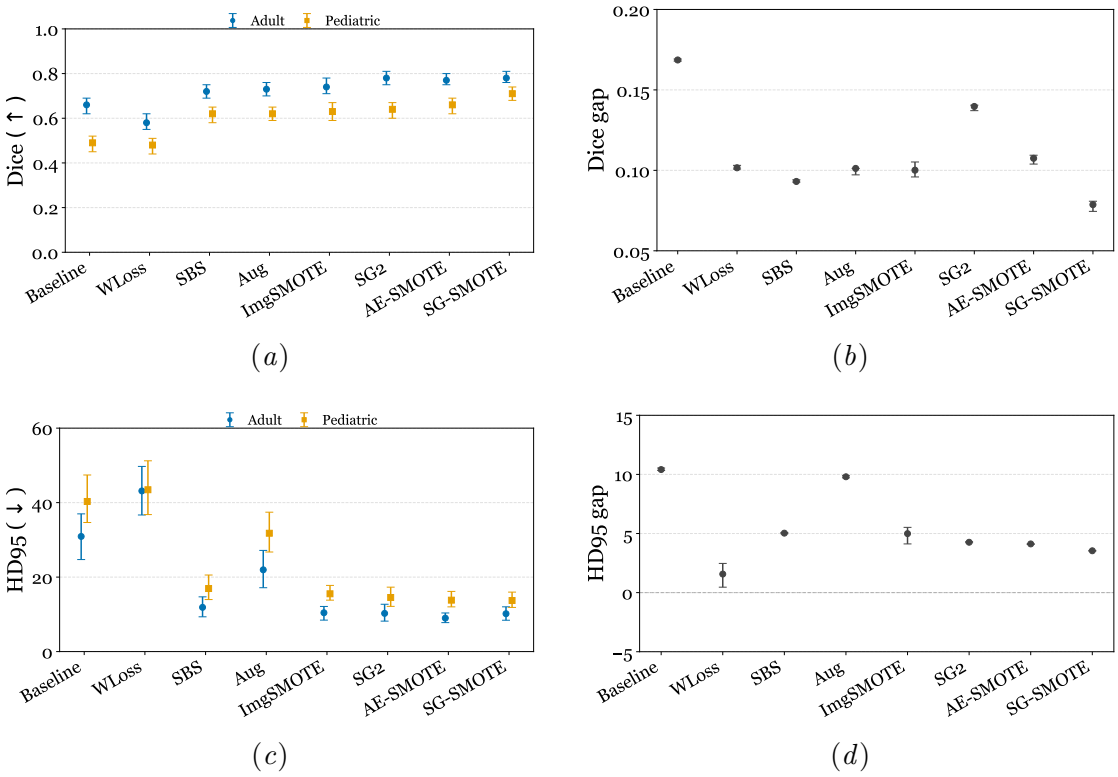


Figure 2: Comparison of adult and pediatric segmentation performance across methods. (a) and (c) show mean Dice and HD95 with 95% bootstrap confidence intervals. (b) and (d) show the corresponding cohort gaps computed as Adult − Pediatric for Dice and Pediatric − Adult for HD95.

task. Slice-wise predictions were reassembled into 3D volumes. We report the mean Dice similarity coefficient and the 95th percentile Hausdorff distance (HD95), computed separately for adult and pediatric cohorts. Subject-level metrics are summarized by their mean and 95% confidence intervals obtained via bootstrap resampling of subjects within each cohort. To quantify cohort balance, we also report performance gaps for both Dice and HD95. For Dice, the gap is computed as Adult − Pediatric, and for HD95 as Pediatric − Adult, so a larger gap always means a larger cross-cohort disparity.

4.2. Results and Discussion

The segmentation performance in terms of Dice and the corresponding Dice gap between adult and pediatric cohorts are reported in Table 1, while the HD95 results are summarized in Table 2. Figure 2 visualizes the results. The results show that the Baseline model obtains a Dice score of 0.66 on adult subjects but gets a lower Dice score of 0.49 on pediatric cases, resulting in a Dice gap of 0.17. The HD95 scores have a similar pattern obtaining an HD95

Table 1: Dice similarity coefficient for adult and pediatric cohorts with 95% bootstrap confidence intervals. Dice gap was calculated as Adult − Pediatric.

Method	**Adult Dice**	**Pediatric Dice**	**Dice Gap**
Baseline	0.66 (0.62, 0.69)	0.49 (0.45, 0.52)	0.17 (0.17, 0.17)
WLoss	0.58 (0.55, 0.62)	0.48 (0.44, 0.51)	0.10 (0.10, 0.10)
SBS	0.72 (0.69, 0.75)	0.62 (0.58, 0.65)	0.09 (0.09, 0.09)
Aug	0.73 (0.70, 0.76)	0.62 (0.59, 0.65)	0.10 (0.10, 0.10)
ImgSMOTE	0.74 (0.71, 0.78)	0.63 (0.59, 0.67)	0.10 (0.10, 0.11)
SG2	0.77 (0.75, 0.80)	0.63 (0.60, 0.67)	0.14 (0.14, 0.14)
AE-SMOTE	0.77 (0.75, 0.81)	0.66 (0.62, 0.69)	0.11 (0.10, 0.11)
SG-SMOTE	0.78 (0.76, 0.81)	0.71 (0.68, 0.74)	0.08 (0.08, 0.08)

Table 2: HD95 for adult and pediatric cohorts with 95% bootstrap confidence intervals. HD95 gap was calculated as Pediatric − Adult.

Method	**Adult HD95**	**Pediatric HD95**	**HD95 Gap**
Baseline	30.92 (24.71, 36.98)	40.31 (34.66, 47.42)	10.41 (10.33, 10.54)
WLoss	43.14 (36.68, 49.72)	43.46 (36.79, 51.23)	1.57 (0.46, 2.47)
SBS	11.88 (9.34, 14.70)	16.93 (13.97, 20.56)	5.03 (4.98, 5.09)
Aug	21.96 (17.15, 27.17)	31.78 (26.74, 37.43)	9.80 (9.71, 9.89)
ImgSMOTE	10.43 (8.45, 12.12)	15.52 (13.81, 17.77)	4.99 (4.12, 5.52)
SG2	10.26 (8.18, 12.70)	14.52 (12.28, 17.18)	4.26 (4.22, 4.31)
AE-SMOTE	8.99 (7.78, 10.37)	13.81 (12.01, 16.16)	4.12 (4.07, 4.14)
SG-SMOTE	10.17 (8.42, 12.02)	13.72 (11.86, 15.97)	3.54 (3.51, 3.57)

gap of 10.41. With the test set being balanced, these discrepancies show a generalization gap between adult and pediatric cases that matches the observation in Figure 1.

The first group of methods (WLoss, SBS, and Aug) do not synthesize new images, but address the imbalance through reweighting, resampling, and augmentation, respectively. The cohort-weighted loss (WLoss) reduces the Dice and HD95 gaps compared to the Baseline, but this is achieved by lowering the adult performance. The adult Dice drops from 0.66 to 0.58, and adult HD95 increases from 30.92 to 43.14. Stratified batch sampling (SBS), which forces adult–pediatric balance at the batch level, achieves improvements for both cohorts relative to the Baseline. Adult and pediatric Dice increase to 0.72 and 0.62, respectively, and both HD95 values also decrease. The Dice gap is reduced to 0.09, and the HD95 gap to 5.03. The pediatric-only augmentation (Aug) achieves similar results. Pediatric Dice improves compared to the Baseline while adult Dice remains high, and both cohorts get lower HD95 values.

The second group of methods use explicit oversampling of minority slices. ImgSMOTE applies SMOTE directly in the image space. It improves pediatric Dice to 0.63 and reduces

HD95 compared to the Baseline, while having good adult performance. The resulting Dice and HD95 gaps (0.10 and 4.99) are better than for the Baseline but similar to SBS and Aug. SG2, which uses a StyleGAN2 generator to synthesize pediatric slices without latent space interpolation further improves adult Dice up to 0.77 and reduces adult HD95 to around 10.26, but pediatric scores are only moderately improved (Dice 0.63, HD95 14.52). Therefore, the Dice and HD95 gaps for SG2 are larger than for some of the simpler baselines.

Based on the results, latent space oversampling methods (AE-SMOTE and SG-SMOTE) provide a stronger way to enrich the pediatric cohort. The autoencoder-based variant (AE-SMOTE) improves both adult and pediatric Dice compared to most other methods, reaching 0.77 and 0.66, respectively, and gets lower HD95 values for both cohorts. The Dice gap (0.11) and HD95 gap (4.12) are smaller than the Baseline. The proposed StyleGAN2 latent space oversampling method (SG-SMOTE) achieves the best overall performance. It obtains the highest pediatric Dice of 0.71 while also improving adult Dice to 0.78. For HD95 it slightly improves the pediatric performance compared to AE-SMOTE but not the adult performance. Importantly, it produces the smallest Dice gap among all methods (0.08) and also the lowest HD95 gap (3.54).

5. Conclusions

This work studies latent space oversampling strategies for addressing the adult–pediatric imbalance problem in brain tumor segmentation. We first constructed an adult–pediatric dataset by combining the BraTS and BraTS-PEDs datasets. For this, we aligned label definitions and preprocessing so that the adult and pediatric cases could be compared under a similar evaluation setup. Using this dataset, we studied how cohort imbalance affects segmentation performance and showed that models trained on imbalanced data (with pediatric cases underrepresented relative to adults) achieve lower Dice and higher HD95 on pediatric subjects than on adults.

To address the imbalance problem, we proposed a cohort-conditioned StyleGAN2 model that jointly synthesizes multi-modal MRI and tumor masks. We inverted the real pediatric slices into the StyleGAN2 latent space, applied SMOTE to generate new pediatric codes, and decoded them into synthetic MRI–mask sets to balance the training data. The proposed approach was compared to several standard imbalance-mitigation strategies. The proposed StyleGAN2 latent space oversampling (SG-SMOTE) achieved the best overall performance, outperforming the autoencoder latent space variant. It improved pediatric Dice, while also improving adult Dice, and achieved the smallest Dice and HD95 gaps across all methods.

Together with these findings, a few limitations should be noted. We have simplified the task to binary tumor–core segmentation instead of the full multi-class BraTS labels, and our pipeline operated entirely in 2D. In addition, the evaluation is performed only on a single adult–pediatric dataset. Finally, StyleGAN2 is a comparatively heavy model, and training it with limited data is challenging. Future studies can therefore include evaluating the approach on more datasets and extending it to multi-class or 3D segmentation settings. Another possible direction could be studying the resulting latent spaces to better understand how cohort-specific characteristics are represented.

Acknowledgments

We are grateful for the financial support obtained from The Swedish Childhood Cancer Fund (*sv.* Barncancerfonden; MT2021-0012), Lion's Cancer Research Foundation in Northern Sweden (LP 22-2319 and LP 24-2367), and the Cancer Research Foundation in Northern Sweden (AMP 25-1227). The computations were supported by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at C3SE partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Saeideh Ghanbari Azar, Tufve Nyholm, and Tommy Löfstedt. Rethinking the deepsmote penalty term and its role in imbalanced learning. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2025. To appear.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, M Bilello, E Calabrese, E Colak, K Farahani, J Kalpathy-Cramer, FC Kitamura, S Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint: arXiv:2107.02314*, 2021.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual International Conference on Machine Learning (ICML)*, 2009.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 2002.

Wuxing Chen, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and CL Philip Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6), 2024.

Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9), 2022.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI. *arXiv preprint: arXiv:2405.18368*, 2024.

Jingru Fu, Simone Bendazzoli, Örjan Smedby, and Rodrigo Moreno. Unsupervised domain adaptation for pediatric brain tumor segmentation. *arXiv preprint: arXiv:2406.16848*, 2024.

Deep B Gandhi, Nastaran Khalili, Ariana M Familiar, Anurag Gottipati, Neda Khalili, Wenxin Tu, Shuvanjan Haldar, Hannah Anderson, Karthik Viswanathan, Phillip B Storm, et al. Automated pediatric brain tumor imaging assessment tool from CBTN: Enhancing suprasellar region inclusion and managing limited data with deep learning. *Neuro-Oncology Advances*, 6(1), 2024.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 2009.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human brain mapping*, 40(17), 2019.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 2021.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation (BraTS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv preprint: arXiv:2305.17033 [eess.IV], 2024.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4), 2024.

Ahmed Makhlouf, Marina Maayah, Nada Abughanam, and Cagatay Catal. The use of generative adversarial networks in medical image augmentation. *Neural Computing and Applications*, 35(34), 2023.

Arnab Kumar Mondal, Lakshya Singhal, Piyush Tiwary, Parag Singla, and Prathosh AP. Minority oversampling for imbalanced data via class-preserving regularized auto-encoders. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Jakub Nalepa, Michal Marcinkiewicz, and Michal Kawulok. Data augmentation for brain-tumor segmentation: a review. *Frontiers in Computational Neuroscience*, 13, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Leslie N Smith. Cyclical learning rates for training neural networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

Arastoo Vossough, Nastaran Khalili, Ariana M Familiar, Deep Gandhi, Karthik Viswanathan, Wenxin Tu, Debanjan Haldar, Sina Bagheri, Hannah Anderson, Shuvanjan Haldar, et al. Training and comparison of nnU-Net and DeepMedic methods for autosegmentation of pediatric brain tumors. *American Journal of Neuroradiology*, 45(8), 2024.

Meng Zhou, Matthias W Wagner, Uri Tabori, Cynthia Hawkins, Birgit B Ertl-Wagner, and Farzad Khalvati. Generating 3D brain tumor regions in MRI using vector-quantization generative adversarial networks. *Computers in Biology and Medicine*, 185, 2025.