Data-Efficient Pediatric Demyelinating Lesion Segmentation with Cross-Age Lesion Augmentation and Trustworthy Predictions

Arti Taneja[1], Hasmila A. Omer[2], and Shantaram Vasikarla[3]

[1] Independent Researcher tanejaai@mail.uc.edu, arti.taneja@gmail.com
[2] Coventry University, Coventry, UK Hasmilaakmar@gmail.com
[3] California State University, Northridge, CA, USA shantaram@computer.org

Abstract. Pediatric demyelinating diseases like multiple sclerosis (MS) are rare, yielding small annotated MRI cohorts where supervised lesion segmentation fails and adult-trained models transfer poorly. We study 2D FLAIR segmentation on PediMS (9 patients; leave-one-patient-out) with transfer to PediDemi (13 non-MS cases). Brain-constrained synthesis improves PediMS Dice over no synthesis (e.g., copy-paste from 0.510 ± 0.094 to 0.602 ± 0.046, Wilcoxon $p = 0.004$); Patch-VAE (0.589 ± 0.042), Patch-DDPM (0.590 ± 0.046), and Cross-Age (0.606 ± 0.042) paste are similar in-domain and not significantly different from one another ($p = 0.65$). On PediDemi transfer, cross-age paste yields the highest point estimate (0.466 ± 0.066), whereas pasting MS lesions onto lesion-free PediDemi hosts yields a negative result. Adult pretraining on MS3SEG followed by PediMS fine-tuning with copy-paste remains competitive (0.598 ± 0.039 / 0.461 ± 0.042). We also report Lesion Evidence-And-Trust (LEAT), an exploratory per-lesion score from counterfactual erasure and TTA uncertainty (High/Medium/Low TPR 0.45/0.32/0.21; Spearman $r = 0.25$, not exceeding lesion size as a trust signal). Thus, synthesis reliably addresses pediatric label scarcity, cross-age paste gives the best out-of-cohort transfer, and per-lesion trust scores provide exploratory quantitative explainability.

Keywords: Pediatric MRI · Demyelinating lesion segmentation · Copy-paste augmentation · Cross-age transfer · Explainable AI.

1 Introduction

Deep learning has substantially advanced adult brain MRI analysis, including multiple sclerosis (MS) lesion segmentation on public adult cohorts and challenge benchmarks [4, 5, 9]. Pediatric applications are harder: the developing brain differs from adult anatomy, motion artifacts are common, and because pediatric demyelinating disease is rare, large labeled cohorts are scarce. Models trained on adults consequently tend to degrade when applied to children[1, 10].

Pediatric-onset MS is uncommon relative to adult MS, yet MRI remains central to diagnosis and follow-up [16], and until recently open lesion-segmentation

resources were almost exclusively adult [4, 5, 9]. PediMS provides the first public dataset dedicated to pediatric MS lesion segmentation, a FLAIR cohort of 9 patients and 28 scans with expert consensus masks [11]. PediDemi complements it with 13 pediatric non-MS demyelinating cases, a clinically important and underrepresented setting in which adult MS models transfer poorly [10]. Together these datasets enable data-efficient pediatric research, but by deep-learning standards the labeled samples remain very few.

When labels are scarce, synthesis and augmentation matter. Simple copy-paste is surprisingly strong in natural images [6]. In medical imaging, lesion-aware methods such as CarveMix and LesionMix mix or edit real lesions under anatomical constraints[2, 17]. Generative models (e.g., variational autoencoders and diffusion) can also synthesize lesions, but with only tens of pediatric volumes they may overfit or add little beyond geometry-preserving paste. Separately, adult MS corpora such as MS3SEG [3] offer denser lesion banks that could support adult-to-pediatric transfer at the data level, by pasting adult lesions onto pediatric hosts rather than only fine-tuning weights [1].

Clinical deployment, finally, demands more than a Dice score. Saliency maps are popular but often coarse for small lesions [14]; calibrated, per-detection trust is more actionable when false positives carry cost.

Contributions. On PediMS (leave-one-patient-out cross-validation, LOOCV) with PediDemi transfer, we: (i) show that brain-constrained lesion copy-paste and generative patch synthesis (VAE, DDPM) all improve Dice over no synthesis on PediMS under a matched protocol, with no statistically significant differences among the synthesis methods; (ii) introduce a cross-age paste strategy that places intensity-matched adult MS3SEG lesions onto pediatric hosts, yielding the best out-of-cohort PediDemi transfer point estimate; (iii) report a negative result showing that pasting MS lesions onto lesion-free PediDemi hosts does not improve transfer; and (iv) propose Lesion Evidence-And-Trust (LEAT), an exploratory per-lesion trust score that couples counterfactual lesion erasure with test-time-augmentation uncertainty. We do not claim a new spatial lesion-likelihood prior [2]; our focus is establishing what reliably improves segmentation under extreme pediatric label scarcity.

2 Methods

2.1 Datasets and preprocessing

Figure 1 summarizes our pipeline: synthetic training-data generation (A), 2D U-Net segmentation with per-lesion LEAT trust scoring (B), and qualitative examples (C). We use three public MRI resources, all restricted to FLAIR since the lesion annotations and MS3SEG donors are defined on this contrast. From PediMS we take pediatric MS scans from 9 patients (28 volumes) with consensus lesion masks [11]; six patients have two or more timepoints. PediDemi provides 13 pediatric non-MS demyelinating cases as an out-of-cohort transfer set [10]. MS3SEG supplies adult MS scans with abnormal white-matter-hyperintensity

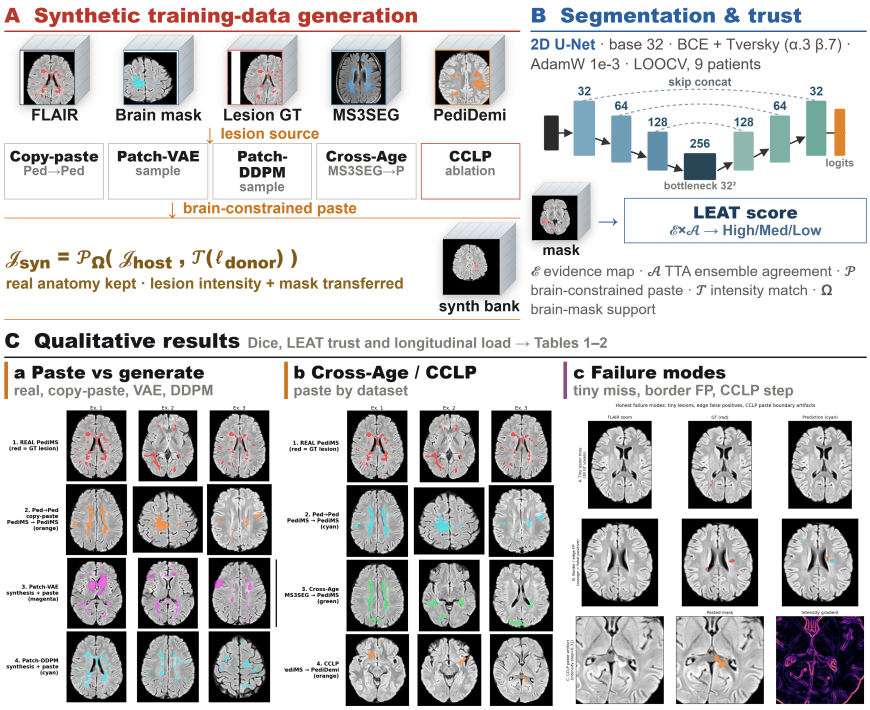

Fig. 1: Overview of data-efficient pediatric lesion segmentation. **(A) Synthesis Bank:** Brain-constrained strategies (Ped→Ped copy-paste, Patch-VAE, Patch-DDPM, Cross-Age, and CCLP ablation) generate synthetic FLAIR training pairs. **(B) Model & LEAT:** 2D U-Net (LOOCV, $N = 9$) predicts lesion masks; LEAT merges counterfactual evidence ($\mathcal{E}$) and TTA agreement ($\mathcal{A}$) into trust scores. **(C) Qualitative Results:** (a) In-cohort synthesis, (b) cross-cohort pasting, and (c) failure modes (misses, border false positives, CCLP artifacts). Quantitative results in Tables 1–2 and Fig. 2.

masks, used for adult pretraining and as lesion donors in cross-age augmentation [3]. We use the publicly released brain-extracted volumes and masks, extract axial 2D slices (resizing to 256×256; MS3SEG is already at this size), and normalize FLAIR intensities within the brain mask using brain-tissue percentiles (1st–99th). No additional denoising or bias-field correction is applied.

2.2 Segmentation model and training

The segmenter is a standard 2D U-Net [12] with 32 base channels. We minimize $\mathcal{L} = \text{BCE} + \text{Tversky}$ [13], with Tversky parameters $\alpha = 0.3$ (false positives) and $\beta = 0.7$ (false negatives) to emphasize recall under severe foreground–background imbalance. By default, models are trained from scratch on PediMS

training folds with AdamW (learning rate 10^{-3}, weight decay 10^{-5}), batch size 8, and 25 epochs. Splits are patient-level via LOOCV on PediMS, so serial scans from the same child never appear in both training and evaluation. The primary metric is scan-level Dice: we average slice Dice per volume, average per-patient across timepoints, and report mean±std across the nine LOOCV folds. PediDemi Dice is reported as a transfer score over the same folds.

2.3 Synthetic lesion augmentation

Synthetic samples are generated from training patients only. We compare five strategies under one paste-and-train protocol (≈4 synthetic slices per real slice). Geometric jitter comprises random flips, 90° rotations, and isotropic scaling in $[0.75, 1.25]$; blending uses $\alpha \in [0.65, 0.95]$ with a hyperintensity boost of 0.05–0.18, and pastes are placed only where the local window is at least 60% brain tissue. Copy-paste, Patch-VAE, and Patch-DDPM stay within PediMS training patients, so the synthesis method is tested without a change in cohort; Cross-Age and CCLP cross cohorts by design.

Brain-constrained copy-paste (Ped→Ped). Connected lesion components are extracted from one training slice and pasted onto another training FLAIR host, restricted to brain tissue as above.

Patch-VAE and patch-DDPM paste. Lesion patches are synthesized with a variational autoencoder [8] or a denoising diffusion probabilistic model [7], then pasted onto real FLAIR hosts under the same rules.

Cross-age paste. Donor lesions come from adult MS3SEG; hosts come from PediMS training slices, with donor intensities matched to the host brain distribution.

Cross-cohort lesion paste (CCLP). PediMS training lesions are pasted only onto clean PediDemi hosts (slices without baseline lesions), again with intensity matching.

2.4 Adult pretraining and fine-tuning

As a weight-transfer baseline, we pretrain the U-Net on MS3SEG and fine-tune on PediMS training folds with brain-constrained copy-paste at a reduced learning rate (10^{-4}).

2.5 Lesion Evidence-And-Trust (LEAT)

For each connected component in the predicted lesion map, LEAT computes a trust score from two quantities. Evidence E is the drop in mean lesion probability when the component is replaced with local brain intensities (counterfactual erasure). Uncertainty U is the pixel-wise standard deviation of the lesion probability under geometric test-time augmentation (flips and small rotations) [15]. Evidence and uncertainty are mapped to $[0, 1]$ by logistic transforms centered at E=0.2 and U=0.12,

$$e = \sigma\big(8(E - 0.2)\big), \qquad c = \sigma\big(12(0.12 - U)\big),$$

and combined as $T = \mathrm{clip}(0.55\,e + 0.45\,c,\ 0,\ 1)$, where σ is the logistic sigmoid, e is an evidence term (high when E is large), and c is a certainty term (high when U is low). Components are assigned to High ($T \geq 0.65$), Medium ($0.40 \leq T < 0.65$), or Low ($T < 0.40$) trust. Within each trust level we report true-positive rate (IoU ≥ 0.1) and mean Dice against ground truth.

2.6 Load-tracking protocol

For patients scanned more than once, we compare the model's predicted lesion load with the consensus load at each timepoint, using the matching LOOCV Cross-Age U-Net [12] (threshold 0.5). Load is the lesion voxel count summed over slices. We assess whether predicted load follows consensus load in magnitude and direction of change.

3 Results

3.1 Main ablation

Table 1 reports LOOCV PediMS and PediDemi transfer Dice, both mean±std over the nine folds. On PediMS, every synthesis method improves significantly over no synthesis but the methods do not differ significantly from one another (Table 1); only CCLP is clearly lower. On PediDemi, Cross-Age gives the highest point estimate, with adult fine-tuning and Patch-VAE close behind, while CCLP stays low. All methods except CCLP exceed the published zero-shot reference, which was measured under a different protocol[1, 10, 11].

3.2 Tracking lesion load over time

Six of nine PediMS patients have at least two timepoints (25 scans). Predicted load from the matching LOOCV Cross-Age U-Net [12] tracks consensus load closely (Pearson $r = 0.993$ across all scans, $r = 0.965$ excluding P2; Fig. 2), and the direction of change from the first scan agrees for P2, P3, P7, and P8 but not P1 or P4. Agreement is weakest for the highest-load patient, P2, where predicted load increasingly under-estimates consensus over time.

3.3 LEAT trust levels

Table 2 pools LEAT over all nine PediMS LOOCV folds (true positive IoU ≥ 0.1). Higher-trust detections yield higher true-positive rates (0.45, 0.32, 0.21 for high/medium/low trust), though the trust–Dice association is modest (Spearman $r = 0.25$). Counterfactual erasure and TTA uncertainty [15] did not exceed lesion size as a trust signal (AUROC 0.61 vs. 0.73 for size alone); multi-fill averaging and nested calibration offered no improvement because size dominates detection confidence in this regime. Because LEAT evaluates only predicted lesions, it cannot flag false negatives. Given its fixed heuristic parameters, LEAT is presented as an exploratory framework.

Table 1: PediMS leave-one-patient-out Dice and PediDemi transfer Dice.

Method	PediMS Dice	PediDemi Dice
No synthetic	0.510 ± 0.094	0.329 ± 0.050
Copy-paste (Ped→Ped)	0.602 ± 0.046	0.425 ± 0.077
Patch-VAE (synth+paste)	0.589 ± 0.042	0.454 ± 0.058
Patch-DDPM (synth+paste)	0.590 ± 0.046	0.426 ± 0.033
Cross-Age (MS3SEG→PediMS)	0.606 ± 0.042	0.466 ± 0.066
CCLP (PediMS→PediDemi)	0.426 ± 0.056	0.349 ± 0.103
MS3SEG→PediMS FT + CP	0.598 ± 0.039	0.461 ± 0.042
Published ref.[a]	≈ 0.54	≈ 0.34

[a] Zero-shot, different protocol from our LOOCV: PediMS UNet++/MSSEG-2016[1, 11]; PediDemi [10].

Paired Wilcoxon tests over the nine PediMS folds confirm that each synthesis method improves over no synthesis (copy-paste Δ=+0.092, p=0.004, 95% CI $[0.046, 0.157]$; Cross-Age Δ=+0.096, p=0.004, CI $[0.055, 0.152]$; VAE Δ=+0.079, p=0.008, CI $[0.043, 0.124]$), whereas the synthesis methods do not differ significantly from one another (copy-paste vs Cross-Age p=0.65; copy-paste vs VAE p=0.20).

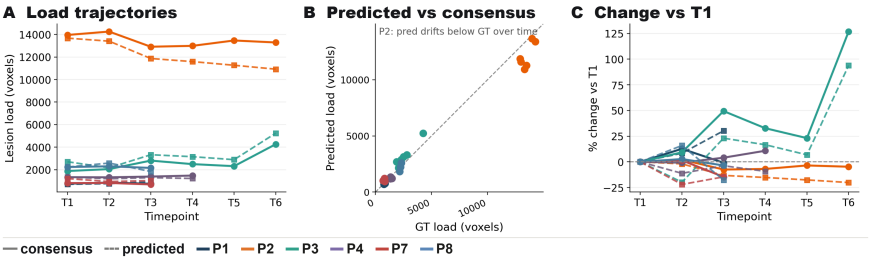

Fig. 2: Longitudinal lesion-load tracking with the held-out Cross-Age U-Net (LOOCV) against consensus masks in six multi-timepoint patients. **(A)** Per-patient load trajectories on a $\log_{10}$ axis (solid: consensus; dashed: predicted); the shaded band marks P2, which is progressively under-segmented (-17.8% at T6). **(B)** Predicted versus consensus load; Pearson $r = 0.993$ across all scans ($r = 0.965$ excluding P2). **(C)** Percentage change relative to T1; Pearson correlation between ground-truth and predicted % change, pooled over all post-T1 scans, $r = 0.882$.

Table 2: LEAT trust groups pooled over all nine PediMS LOOCV folds (exploratory).

Trust	n	TPR	Mean Dice
High	1182	0.45	0.24
Medium	497	0.32	0.20
Low	174	0.21	0.13

4 Discussion

Brain-constrained copy-paste and Cross-Age lesion pasting improve 2D FLAIR lesion segmentation on PediMS under a LOOCV protocol. In contrast, CCLP does not yield similar improvements, while Cross-Age gives the best PediDemi point estimate among the evaluated methods, though it is not significantly ahead of the other synthesis strategies (Table 1). On the PediMS cohort, standard copy-paste augmentation performs comparably to generative methods such as Patch-VAE and Patch-DDPM under identical validation conditions[2, 6–8, 17]. Given the limited sample of nine patients, generative patch architectures introduce texture variation but do not clearly outperform simpler spatial pasting within the training domain.

For the cross-cohort PediDemi evaluation, pasting intensity-matched adult MS3SEG lesions onto pediatric PediMS hosts (Cross-Age) gives the best transfer point estimate [3, 10]. Patch-VAE also improves generalization relative to baseline copy-paste, while pretraining on adult MS3SEG followed by PediMS fine-tuning with copy-paste remains competitive [1]. Cross-Age paste is therefore a practical way to use adult lesion annotations when pediatric training data are limited.

On the other hand, CCLP, which pastes pediatric PediMS lesions onto unaffected PediDemi host images, fails to improve segmentation scores and often leaves visible intensity steps along the paste boundaries. Cross-cohort pasting is therefore not inherently beneficial: the donor and host domains matter far more than the pasting mechanism.

Evaluating longitudinal performance across multi-timepoint scans shows that the matching LOOCV Cross-Age U-Net [12] tracks consensus lesion load across the majority of cases. The predicted direction of change from the initial scan aligns with consensus for P2, P3, P7, and P8, and disagrees only for P1 and P4 (Fig. 2). The primary weakness occurs in patient P2, where the model consistently underestimates lesion load, causing the gap to widen over subsequent visits.

To support clinical trust, LEAT computes an instance-level confidence score from counterfactual erasure and test-time-augmentation uncertainty [15]. Higher-trust detections tend toward higher true-positive rates across folds, though the pooled association is modest and we present LEAT as exploratory. Multi-fill averaging and nested calibration of the LEAT signals offered no improvement, as lesion size dominates detection confidence in this low-data regime. Whereas standard saliency tools such as Grad-CAM provide only coarse reference maps [14], LEAT offers a quantitative trust metric, though it remains bounded by the model's recall and cannot recover lesions omitted entirely at inference.

5 Limitations

This study has several constraints. First, dataset sizes are small (9 PediMS patients, 13 PediDemi patients), and our longitudinal analysis is limited to six

cases where overall volume is dominated by a single patient (P2). Second, we evaluate a single 2D U-Net architecture without benchmarking against 3D models, modern backbones (e.g., transformers, Mamba, nnU-Net), or existing paste methods such as CarveMix and LesionMix. Third, models are trained strictly on 2D FLAIR without multi-sequence (T1, T2) or 3D fusion, and pasting relies on basic brain masking rather than atlas-guided alignment. Fourth, evaluations use single-run Dice scores without secondary metrics (Hausdorff distance, lesion-load errors) or repeated random seeds. Finally, differences in validation protocols prevent direct comparisons with published adult-to-pediatric benchmarks.

6 Conclusion

On small pediatric FLAIR cohorts, any lesion synthesis reliably improves segmentation over no synthesis, while individual synthesis methods (copy-paste, patch-VAE, patch-DDPM, and Cross-Age) do not differ significantly from one another. Cross-Age gives the best point estimate for transfer to the non-MS PediDemi domain. Separating in-cohort synthesis from cohort-crossing paste shows that the benefit depends on the donor and host rather than the mechanism alone. We additionally explore LEAT, an exploratory per-lesion trust score, and find that erasure- and TTA-based evidence does not exceed lesion size as a trust signal at this scale. Future work will investigate multi-sequence and 3D models, atlas- or uncertainty-guided pasting, active learning for missed lesions, stronger trust calibration, and multi-site clinical evaluation.

References

1. Badea, L., Popa, M.: Toward generalizable multiple sclerosis lesion segmentation models. IEEE Access **13**, 97859–97869 (2025). https://doi.org/10.1109/ACCESS.2025.3576145
2. Basaran, B.D., et al.: LesionMix: a lesion-level data augmentation method for medical image segmentation. In: MICCAI DALI Workshop 2023. LNCS, vol. 14379, pp. 73–83. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-58171-7_8
3. Bawil, M.B., Shamsi, M., Ghalehasadi, A., Jafargholkhanloo, A.F., Bavil, A.S.: A multiple sclerosis MRI dataset with tri-mask annotations for lesion segmentation. Sci. Data **13**, 867 (2026). https://doi.org/10.1038/s41597-026-07184-5
4. Carass, A., et al.: Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. NeuroImage **148**, 77–102 (2017). https://doi.org/10.1016/j.neuroimage.2016.12.064
5. Commowick, O., et al.: Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci. Rep. **8**, 13650 (2018). https://doi.org/10.1038/s41598-018-31911-7
6. Ghiasi, G., et al.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR 2021, pp. 2918–2928. IEEE, Piscataway (2021). https://doi.org/10.1109/CVPR46437.2021.00294
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851 (2020)

8. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: 2nd International Conference on Learning Representations (ICLR) (2014). https://doi.org/10.48550/arXiv.1312.6114
9. Lesjak, Ž., et al.: A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics **16**, 51–63 (2018). https://doi.org/10.1007/s12021-017-9348-7
10. Popa, M., Vişa, G.A.: PediDemi – a pediatric demyelinating lesion segmentation dataset. Mach. Learn. Biomed. Imaging **3**, 866–874 (2025). https://doi.org/10.59275/j.melba.2025-123f
11. Popa, M., Vişa, G.A., Şofariu, C.R.: PediMS: a pediatric multiple sclerosis lesion segmentation dataset. Sci. Data **12**, 1184 (2025). https://doi.org/10.1038/s41597-025-05346-5
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: MLMI 2017. LNCS, vol. 10541, pp. 379–387. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_44
14. Selvaraju, R.R., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV 2017, pp. 618–626. IEEE, Piscataway (2017). https://doi.org/10.1109/ICCV.2017.74
15. Wang, G., et al.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019). https://doi.org/10.1016/j.neucom.2019.01.103
16. Waubant, E., et al.: Difference in disease burden and activity in pediatric patients on brain magnetic resonance imaging at time of multiple sclerosis onset vs adults. Arch. Neurol. **66**(8), 967–971 (2009). https://doi.org/10.1001/archneurol.2009.142
17. Zhang, X., et al.: CarveMix: a simple data augmentation method for brain lesion segmentation. NeuroImage **271**, 120041 (2023). https://doi.org/10.1016/j.neuroimage.2023.120041