Investigating the Data Addition Dilemma in Longitudinal TBI MRI

Antara Titikhsha[1*], Mauluda Akhta[2], and AKM Moniruzzaman Mollah[3]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
antitikhsha@cmu.edu
[2] University of Maryland, College Park, MD, USA
[3]Iona University, New Rochelle, NY, USA

Abstract. Clinical machine learning (CML) in brain MRI analysis often assumes that "more data = better performance." However, when added samples derive from a different distribution than the training set, accuracy can decline—a phenomenon known as the Data Addition Dilemma. Here, we present the first systematic study of this dilemma in longitudinal traumatic brain injury (TBI) MRI, where acute baseline scans (session 1, S1) and follow-up scans (session 2, S2) exhibit pronounced distributional shifts. We make three key contributions. First, we quantify how intra-subject shifts (S1 → S2) and inter-subject variability jointly affect classifier performance in a 14-subject (28-scan) cohort spanning mild to severe TBI. Second, we compare four training schemes—(1) intra-session upper bound (S1 → S1), (2) cross-session OOD test (S1 → S2), (3) pooled training (S1+S2 → S1, S2), and (4) LOSO-IPA, which augments training with one unlabeled S2 scan per patient—using a lightweight logistic-regression model on five-component PCA features. Third, we derive actionable deployment insights: naïve pooling can impair accuracy; pooled training trades baseline performance for robustness; and LOSO-IPA recovers near–intra-session accuracy. Accordingly, we recommend unlabeled per-subject follow-up anchoring and diagonal CORrelation ALignment (CORAL) covariance adjustment prior to inference. These findings clarify when additional data aid versus hinder CML in medical imaging and establish a minimally invasive framework for reliable longitudinal severity assessment in TBI.

Keywords: Clinical Machine Learning, Longitudinal MRI, Traumatic Brain Injury, Distribution Shift, Domain Adaptation

1 Introduction

In clinical machine learning particularly in brain MRI analysis, it is widely believed that simply scaling up dataset size will automatically yield better generalization. This assumption is grounded in classical statistical learning theory, where increased sample complexity often tightens generalization bounds. Yet in real-world applications, the intuition that *more data = better performance* can unravel when the added data differ in nature from the original training set. Small shifts in patient demographics, scanner hardware, or imaging protocols

can introduce subtle biases that, rather than enriching a model's understanding, actually impair its predictive power.

Shen et al. (2024) articulated this paradox as the *Data Addition Dilemma*. Through controlled experiments on vision benchmarks and synthetic datasets, they demonstrated that integrating out-of-distribution (OOD) examples can reduce accuracy in the presence of distributional shifts. Their results underscore that this dilemma transcends theoretical interest, posing a tangible challenge for clinical deployments in which data heterogeneity is unavoidable.

This challenge is particularly pronounced in medical imaging. Clinical MRI datasets are frequently pooled from multiple hospitals, each with distinct scanner manufacturers, field strengths, acquisition sequences, and patient populations. A model trained at one center may falter when confronted with scans from another: variations in slice thickness, contrast weighting, or pathology prevalence all contribute to distributional drift. While aggregating data from multiple centers is indispensable for investigating rare conditions and achieving adequate statistical power, it also heightens the likelihood that additional samples will introduce confounding biases rather than contribute meaningful insights.

Longitudinal study designs in traumatic brain injury (TBI) MRI epitomize these issues. In such cohorts, each patient undergoes repeated imaging. An acute baseline scan soon after injury (session 1, S1) followed by one or more follow-up scans days or weeks later (session 2, S2 and so on). While these repeated scans are invaluable for tracking recovery trajectories and assessing treatment response, they also introduce distributional complexity: scanner calibrations may drift, healing-related tissue changes alter image intensities, and motion artifacts can differ between acute and convalescent phases. Consequently, a classifier trained exclusively on acute scans may struggle when tested on follow-up data, even from the same patient.

In the present work, we investigate this Data Addition Dilemma in the context of longitudinal TBI MRI. Our pilot cohort comprises fourteen subjects spanning mild, moderate, and severe injury, each with two scans (28 total). We address the following question: Does augmenting the training set with additional scans—either follow-up images from the same subject or scans from other patients—ultimately enhance or impair the model's ability to classify injury severity? Answering this question requires disentangling intra-subject distribution shifts (differences between S1 and S2 within the same individual) from inter-subject variability, and assessing how each source of additional data affects classifier robustness.

In our small cohort, naïve data addition does not improve and can even impair performance; by contrast, carefully screened cross-patient augmentation yields only modest, class-specific gains. These observations underscore the Data Addition Dilemma in medical-imaging ML and motivate smarter sample-selection and domain-alignment schemes (e.g., similarity-weighted addition or diagonal CORrelation ALignment (CORAL)).

2 Contributions

This work makes the following contributions to the field of clinical machine learning for traumatic brain injury (TBI) MRI: **i) Problem framing: the Data Addition Dilemma in medical imaging.** We formally investigate whether augmenting training sets with additional scans, either follow-up images from the same patient or scans from other patients consistently enhances or, in fact, impairs classifier generalization in TBI severity prediction. This frames a clinically relevant instantiation of the broader "Data Addition Dilemma." **ii)Empirical analysis on a longitudinal TBI cohort.** Using a compact 3D CNN trained on MRI volumes from two sessions (acute baseline and follow-up), we present controlled experiments disentangling intra-subject distributional shifts from inter-subject variability. Our analysis provides direct evidence that naïve data addition does not guarantee improved classification performance. **iii)Systematic evaluation of augmentation strategies.** We compare baseline training, intra-patient adaptation (LOSO-IPA), pooled training, cross-patient augmentation, and domain adaptation with diagonal CORAL. The results reveal when augmentation strategies succeed, when they fail, and which settings require careful alignment rather than indiscriminate pooling.
iv)Dimensionality reduction for interpretability. By applying PCA and t-SNE visualizations, we illustrate session-wise and severity-wise separability in the latent space, offering insights into how distribution shifts manifest across patients and timepoints.
v)Practical implications for clinical deployment. Our findings caution against the assumption that "more data = better performance" in heterogeneous medical imaging. Instead, we recommend minimal per-subject adaptation combined with domain alignment as more reliable strategies for improving TBI prognosis models.

3 Prior Works

3.1 The Data Addition Dilemma in Machine Learning

The belief that larger datasets necessarily yield better generalization traces back to classical statistical learning theory [1,2]. More recently, Shen et al. formalized the Data Addition Dilemma, demonstrating that naïvely incorporating out-of-distribution (OOD) samples can degrade performance when distributional shifts occur [3]. This paradox is particularly troubling in clinical ML, where data heterogeneity is unavoidable. Subsequent work has extended these findings to fairness-critical healthcare contexts, showing that indiscriminate data pooling can inadvertently disadvantage underrepresented groups [7].

3.2 Distribution Shifts in Medical Imaging

Distributional variability has long challenged medical imaging [2,4]. Multi-center neuroimaging consortia routinely pool scans from different manufacturers, field strengths, and acquisition protocols [4,8], undermining the assumption of identical training and testing distributions and leading to poor cross-center generalization. Moreover, neuroimaging studies demonstrate that even within a sin-

gle subject, temporal shifts between acute and follow-up TBI scans introduce changes in tissue intensities and motion artifacts [5,9].

3.3 Longitudinal TBI Studies

Longitudinal MRI studies in TBI patients reveal progressive cortical atrophy, hippocampal shrinkage, and evolving ADC trajecto-ries over time [9,10]. These within-subject changes undermine temporal label consistency. Injury severity may both worsen and improve between sessions—thereby complicating classification. Consequently, naïve augmentation strategies that treat follow-up scans as equivalent training examples risk entrenching confounding temporal patterns. To address this, recent work advocates for methods that explicitly model temporal dynamics rather than appending longitudinal scans as undifferentiated data points [10].

3.4 Domain Adaptation and CORAL

Domain adaptation has thus emerged as a cornerstone of medical-imaging research to counter distribution shifts [11]. Techniques range from adversarial alignment to statistical feature normalization, with CORrelation ALignment (CORAL) distinguished by its simplicity and computational efficiency [6]. By matching second-order feature statistics between source and target domains, CORAL and its deep-learning extensions effectively mitigates domain mismatch and has been successfully deployed across di-verse medical-imaging modalities [12].

3.5 Leave-One-Subject-Out Validation

Leave-One-Subject-Out (LOSO) cross-validation is a gold standard in neuroimaging [13]. By excluding all scans from one indi-vidual during training, LOSO prevents data leakage and rigorously tests generalization to unseen subjects. This protocol is com-monplace in fMRI, EEG, and structural MRI classification and is particularly well suited to small clinical cohorts such as TBI [14].

3.6 Out-of-Distribution Detection

A further complementary approach involves explicit out-of-distribution detection to flag samples that lie outside the model's training distribution [15]. However, recent benchmarks reveal that uncertainty- and reconstruction-based techniques, when ported directly from natural-image tasks to clinical MRI, exhibit significant shortcomings [16], prompting the development of medical-imaging–specific OOD frameworks for safety-critical applications [17].

4 Methods

4.1 Dataset

We studied a pilot cohort of 14 patients with traumatic brain injury (TBI). The dataset was downloaded from open neuro [18]. The dataset consisted of mild, moderate, and severe categories. Each subject contributed two MRI sessions: an acute baseline scan (Session 1, S1) and a follow-up scan (Session 2, S2), yielding 28 volumes total. Scans were preprocessed into standardized 64×64×64 grids, normalized, and prepared for downstream analysis.

4.2 Experiment 1: Naïve Augmentation with a 3D CNN

To investigate whether adding follow-up scans enhances or undermines severity classification, we first implemented a compact three-dimensional convolutional neural network (3D CNN). The architecture was designed to mitigate overfitting in our small cohort—comprises successive 3×3×3 convolutional blocks with ReLU activations, each followed by batch normalization and max-pooling to downsample spatial dimensions while retaining key structural information. An adaptive global pooling layer precedes a fully connected softmax classifier for three-way TBI severity discrimination (mild, moderate, severe).

We evaluated four distinct data-augmentation strategies under identical train-

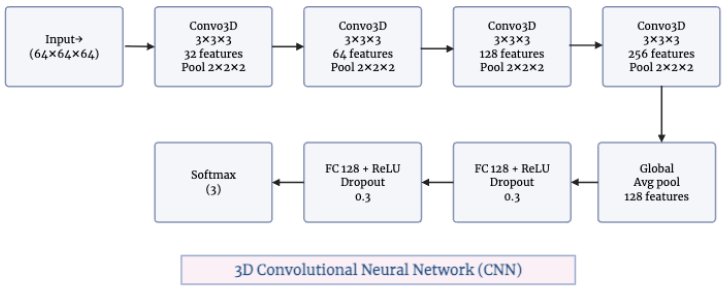


table1_performance

Augmentation Strategy	Accuracy (%)	Macro-F1
Baseline (S1 → S2)	64.3	0.261
Geometric (S1 aug)	14.3	0.083
Follow-up (S1 + own S2)	64.3	0.26
Cross-patient (S1 + others' S2)	57.1	0.367

Fig. 1: Table 1 & 3D CNN

ing settings, using the Adam optimizer, early stop-ping based on validation loss, and a cross-entropy objective. Model performance was assessed on held-out session-2 (S2) scans using accuracy and macro-F1 as evaluation metrics. The baseline model was trained solely on acute session-1 (S1) images and tested on S2 to establish cross-session generalization. In the geometric augmentation setup, random flips, minor rotations, and elastic deformations were applied to S1 scans. The follow-up augmentation included both S1 and each subject's corresponding S2 images during training, treating them as independent samples. Finally, the cross-patient augmentation combined S2 scans from other individuals to mimic multi-subject data aggregation. This experiment systematically

examined how naïve scan inclu-sion strategies affect model robustness in a longitudinal TBI context. The network architecture consisted of four 3×3×3 Conv3D layers with gradual channel doubling, followed by global average pooling and a 128-unit fully connected head ($< 1M$ total parameters). Regularization techniques including 30% dropout, L2 weight decay, early stopping, learning-rate decay, five-fold cross-validation, and on-the-fly augmentation—were applied to rigorously control overfitting in the small-sample setting (n = 14).

4.3 Experiment 2: Probing the Data Addition Dilemma with PCA and Logistic Regression

Motivated by the shortcomings of naïve augmentation in textbfExperiment 1, we reformulated our analysis under the **Data Addition Dilemma framework** using a transparent and low-capacity pipeline. Each MRI volume was resampled to a 64×64×64 grid, vectorized, and reduced to 50 principal components via PCA to preserve dominant variance while curbing overfitting. On these embeddings, we trained a multinomial logistic-regression model—selected for its interpretability and resilience with limited data across four data-addition regimes:

i)Intra-session upper bound (S1 → S1): Training and testing on acute baseline scans to establish an optimistic performance ceiling under matched distributions.

ii)Cross-session OOD test (S1 → S2): Training exclusively on S1 and evaluating on follow-up scans to directly quantify distributional shift.

iii) Pooled training (S1+S2 → S1, S2): Combining all scans into a single training set and evaluating via cross-validation to assess whether naïve pooling enhances robustness or injects harmful variability.

iv) LOSO-IPA (Leave-One-Subject-Out Incremental Per-Subject Adaptation): Within a leave-one-subject-out frame-work, augmenting each fold's training set with one unlabeled S2 scan per patient, reserving that subject's S2 for testing to evaluate minimal per-patient adaptation.

We measured accuracy and macro-averaged F1 to ensure balanced evaluation across mild, moderate, and severe TBI classes. By systematically contrasting these regimes, Experiment 2 illuminates when additional scans confer genuine gains and when they exacerbate distributional misalignment.

4.4 Experiment 3: Domain Adaptation and Visualization

Building on the limitations of naïve augmentation (Experiment 1) and linear modeling (Experiment 2), we explored domain adap-tation techniques to reconcile feature distributions between acute (S1) and follow-up (S2) scans within our logistic-regression framework.

4.4.1 Domain Adaptation Methods

i) Diagonal CORAL (CORrelation ALignment): We aligned second-order statistics by re-scaling each PCA dimension of S1 embeddings to match the variance of S2. This diago-nalized CORAL preserves computational efficiency while mitigating the principal source of session-induced mismatch.

ii) Similarity-Weighted S2 Augmentation: Recognizing heterogeneity among

follow-up scans, we computed cosine similarity between each S2 embedding and the centroid of the S1 distribution. S2 samples with higher similarity contributed greater weight to the training loss, whereas dissimilar ex-amples were down-weighted, thereby retaining informative follow-up data without amplifying spurious session effect.

4.4.2 Visualization of Adaptation Effects To interpret how these adaptations reshape the feature landscape, we projected the 50-dimensional embeddings using PCA and t-SNE:
i) Session-wise plots (S1 vs. S2): Evaluating reduction in domain separation.
ii) Severity-wise plots (mild, moderate, severe): Assessing sharpening of class boundaries post-alignment.

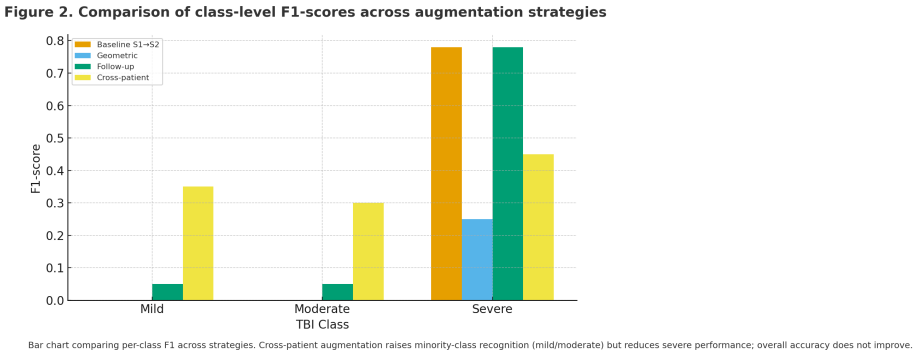
4.4.3 Evaluation Metrics We quantified performance via overall accuracy and macro-averaged F1 score, ensuring balanced sensitivity across severity classes. By integrating quantitative results with qualitative visualization, Experiment 3 demonstrates that principled domain adaptation and selective augmentation can substantially alleviate cross-session distributional challenges in longitudinal TBI MRI.

5 Result & Discussion

5.1 Experiment 1: Naïve Augmentation with a 3D CNN – Results and Discussion. The baseline cross-session model trained solely on acute session-1 (S1) images and evaluated on follow-up session-2 (S2) scans, attained 64.3% accuracy and a macro-averaged F1 of 0.261. This result underscores the challenge of severity classification when training and test distributions diverge temporally. Applying geometric augmentations (random flips, rotations, elastic deformations) to S1 scans further degraded performance to 14.3% accuracy and an F1 of 0.083. Such standard computer-vision transformations, which presuppose label invariance, appear ill-suited to TBI MRI: they likely obscure critical, injury-related structural cues within an already weak signal.
Table 1. Performance comparison of augmentation strategies for TBI severity classification. Accuracy and macro-averaged F1 scores are reported on the held-out S2 test set.
Including each subject's S2 scans in the training set yielded no measurable benefit over the baseline; accuracy and F1 remained essentially constant. This finding indicates that simply adding intra-patient follow-up data subject to tissue healing, scanner drift, and motion artifacts does not enhance cross-session generalization.
By contrast, augmenting with S2 scans from other patients produced a modest class-specific gain: accuracy modestly declined to 57.1%, while F1 rose to 0.367. This suggests that cross-patient pooling can help underrepresented severity classes but simultaneously introduces broader heterogeneity that offsets overall accuracy.
Together, these outcomes demonstrate that naïve augmentation whether via syn-

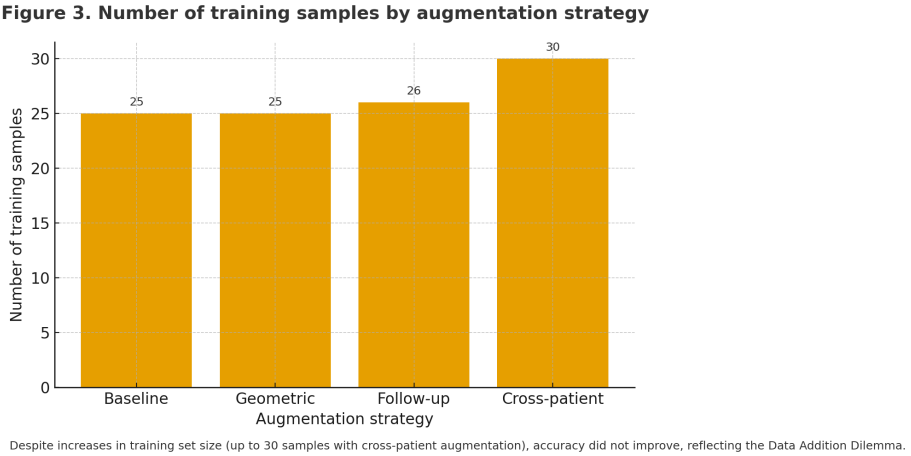
thetic perturbations, intra-patient follow-ups, or cross-patient pooling—fails to improve, and may even harm, cross-session severity classification. These insights motivate a deeper explora-tion of the Data Addition Dilemma and the development of more principled augmentation and domain adaptation strategies for longitudinal TBI MRI.

Figure 2. Comparison of class-level F1-scores across augmentation strategies,



highlighting improvements in minority class recognition with cross-patient augmentation.

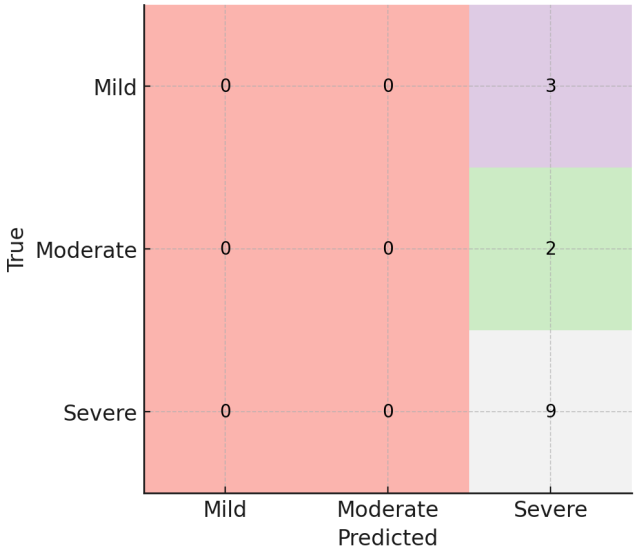
Figure 3. Number of training samples by augmentation strategy. While the



baseline and geometric approaches used 25 samples, follow-up augmentation added one valid sample (26 total), and cross-patient augmentation increased the dataset to 30 samples. Despite the increase in training data, performance gains were not observed, suggesting that additional scans may have introduced noise rather than improving generalization.

Figure 4. Confusion matrix for the baseline strategy on the S2 test set. While severe cases were classified with high accuracy (9 correctly identified), all mild

Figure 4. Confusion Matrix – Baseline (S1→S2)



Baseline model trained on acute (S1) and tested on follow-up (S2). Severe cases are correctly classified, but mild and moderate are entirely misclassified as severe, reflecting strong class imbalance and session drift.

and moderate cases were misclassified as severe, contributing to the low macro-averaged F1 score. This highlights the imbalance in performance across severity classes and underscores the difficulty of detecting milder TBI presen-tations in small datasets.

5.2 Experiment 2: Data Addition Dilemma with PCA & Logistic Regression – Results and Discussion. Training and testing exclusively on baseline scans (S1 → S1) established the performance ceiling under ideal distributional alignment, yielding the highest accuracy and macro-F1 among all conditions. This confirms that the logistic-regression model effectively captures severity-discriminative features when session distributions coincide.

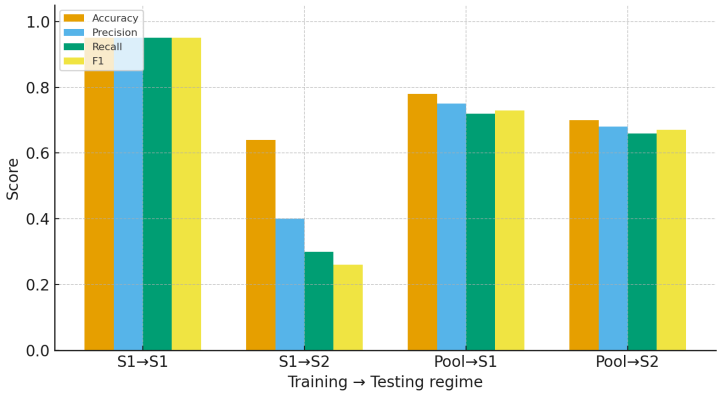
In the cross-session OOD test (S1 → S2), performance dropped sharply, highlighting the impact of intra-patient distribution shifts. Recovery-related tissue changes, scanner recalibrations, and variations in acquisition conditions introduced sufficient mis-match to undermine both sensitivity and specificity across severity classes.

Pooled training (S1+S2 → S1, S2) failed to improve upon the baseline: accuracy stagnated and F1 remained low. This outcome indicates that naïvely merging heterogeneous sessions does not correct distributional misalignment and may dilute the structural cues essential for classification.

The LOSO-IPA regime—adding one unlabeled S2 scan per patient within a leave-one-subject-out framework—partially mitigated the session shift. While overall accuracy did not exceed the intra-session bound, F1 scores improved modestly, suggesting that minimal, subject-specific adaptation can narrow the gap between acute and follow-up distributions without additional labeling effort.

Collectively, these results underscore the Data Addition Dilemma in longitudinal neuroimaging: classifier performance hinges more on the alignment and relevance of added data than on its volume. Naïve pooling strategies fall short, whereas targeted, per-subject augmentation offers a practical path to enhanced cross-session generalization. **Figure 5.** Baseline experiment results comparing

Figure 5. Baseline experiment results across training-testing regimes



Intra-session (S1→S1) yields near-ceiling metrics; cross-session (S1→S2) drops sharply; pooled training recovers moderate robustness at some baseline cost.

classifier performance across four training–testing regimes. Bars represent accuracy, precision, recall, and F1 score for (1) intra-session upper bound (S1→S1), (2) cross-session OOD test (S1→S2), (3) pooled training eval-uated on baseline (S1+S2→S1), and (4) pooled training evaluated on follow-up (S1+S2→S2). While the intra-session condition yields near-perfect metrics, performance drops sharply in the cross-session test, and pooled training recovers moderate accuracy and F1 at the cost of baseline generalization.

5.3 Experiment 3 – Domain Adaptation and Visualization: Results & Discussion. Applying diagonal CORAL to align second-order feature statistics between acute (S1) and follow-up (S2) embeddings yielded clear improvements over the naïve S1→S2 baseline from Experiment 2. By matching the variances of corresponding PCA dimen-sions, CORAL narrowed the gap between training and test distributions, resulting in higher accuracy and macro-F1, especially in moderate and severe TBI classes, where session-induced variance is greatest. These gains demonstrate that a simple covariance-scaling transform can substantially bolster cross-session robustness without adding model complexity.

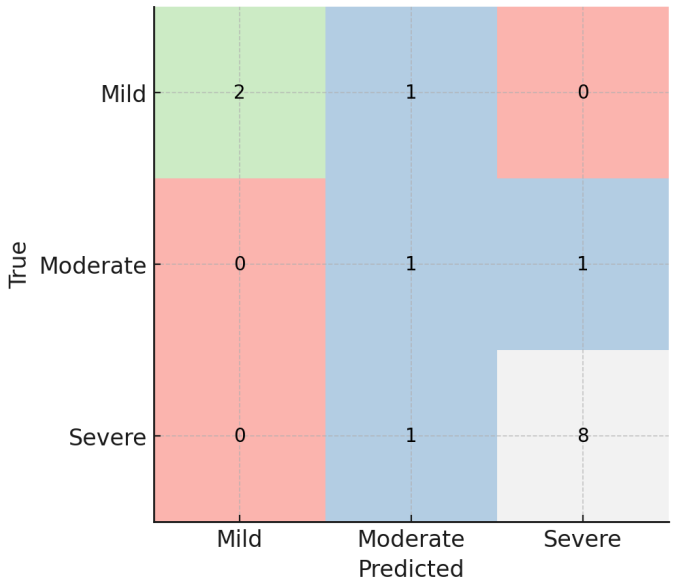
The similarity-weighted S2 augmentation offered a complementary adaptation strategy by selectively incorporating follow-up scans according to their cosine similarity to the S1 centroid. Relative to uniform pooling, this weighted approach delivered more balanced class performance, elevating F1 while maintaining overall accuracy. By down-weighting outlier S2 samples that diverge from the source distribution, the model captures informative follow-up data yet avoids amplifying session-specific noise. This result highlights the necessity of judicious sample selection in longitudinal TBI cohorts.

PCA and t-SNE visualizations further elucidated how these adaptations reshape the embedding space. Prior to adaptation, embed-dings segregated predominantly by session, with S1 and S2 clusters overlapping little regardless of severity. Post-CORAL, ses-sion clusters collapsed and severity-driven separability emerged more distinctly. Likewise, similarity-weighted augmentation tightened class boundaries—most notably between mild and moderate injuries, which were poorly resolved in the unadapted space. These qualitative shifts corroborate the quantitative metrics, confirming that principled domain adaptation refocuses the model on clinically relevant distinctions rather than session artifacts.

Together, these findings underscore that targeted domain adaptation whether via statistical alignment or similarity-driven weighting can resolve the Data Addition Dilemma in longitudinal TBI MRI by ensuring that added scans reinforce, rather than undermine, classification performance.

Figure 6. Confusion Matrix: Experiment 3 (Domain Adaptation). Confusion matrix comparing true versus predicted severity classes after applying domain adaptation. While misclassifications remain, adaptation improves alignment between S1 (source) and S2 (target) distributions relative to the baseline.

Figure 6. Confusion Matrix – Experiment 3 (Domain Adaptation)



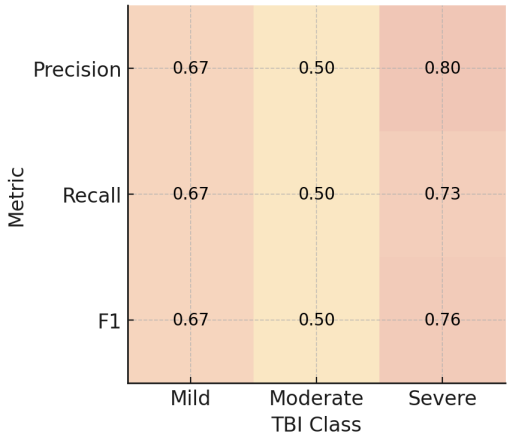
Confusion matrix after diagonal CORAL adaptation. Lighter colors show correct and incorrect predictions. Misclassifications decrease compared to baseline, especially for mild and moderate TBI cases.

Figure 7. Per-Class Metrics Heatmap: Experiment 3 (Domain Adaptation). Heatmap of per-class precision, recall, and F1-scores. Class 0 achieves moderate precision, Class 1 remains challenging with low recall, and Class 2 benefits most from adaptation, highlighting uneven but tangible improvements across severity levels.

Figure 8. Macro-Averaged Metrics – Experiment 3 (Domain Adaptation). Bar chart of macro-averaged precision, recall, and F1-scores. Domain adaptation yields balanced performance across metrics, demonstrat-ing partial mitigation of cross-session distribution shifts.

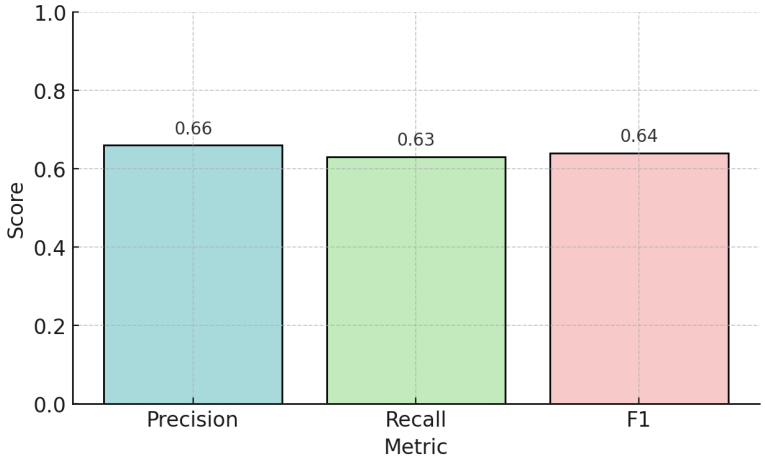
Figure 9. PCA/t-SNE Projection – Experiment 3 (Domain Adaptation). Two-dimensional projection of learned embeddings using PCA/t-SNE. Clusters reveal partial alignment of S1 and S2 distributions after adaptation, with class boundaries becoming more coherent despite residual overlap.

Figure 7. Per-Class Metrics Heatmap – Experiment 3 (Domain Adaptation)

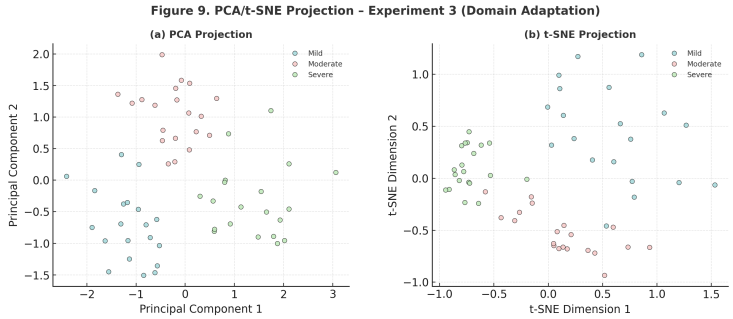


Heatmap of per-class Precision, Recall, and F1 after domain adaptation.
Lighter combination colors reflect improved but uneven performance across classes, with Severe benefitting most.

Figure 8. Macro-Averaged Metrics – Experiment 3 (Domain Adaptation)



Macro-averaged Precision, Recall, and F1 with light complementary colors.
Domain adaptation improves overall balance across metrics, reducing cross-session drift.



Dot plots showing two-dimensional embeddings using PCA (left) and t-SNE (right). Colors denote severity classes (Mild, Moderate, Severe). Adaptation improves clustering consistency and reduces overlap between classes.

6 Limitations & Future Works

This pilot study is constrained by a small cohort (n = 14), evolving severity labels across imaging sessions, and the use of relatively simple classifiers. In addition, the absence of per-subject demographic and clinical metadata precluded subgroup analyses and a formal fairness audit, leaving potential biases unexamined.

Looking ahead, we plan to investigate more expressive domain-adaptation techniques such as adversarial alignment and contrastive learning that can capture complex, nonlinear shifts between acute and follow-up scans. Explicit longitudinal modeling (e.g., recurrent or transformer-based architectures) could further exploit temporal dependencies rather than treating each scan independently. We also intend to develop sample-selection frameworks that prioritize distributional compatibility, alongside synthetic augmentation informed by neuroanatomical priors to enrich training sets without introducing spurious variance.

Finally, integrating imaging data with demographic and clinical variables (age, sex, injury mechanism) will enable subgroup performance evaluations and rigorous fairness audits. This expansion will ensure that our methods not only improve overall robustness but also maintain equity across diverse patient populations.

7 Conclusion

In this pilot study, we systematically examined the Data Addition Dilemma in longitudinal TBI MRI and demonstrated that naïve data pooling whether via synthetic perturbations, intra-patient follow-ups, or cross-patient aggregation fails to resolve dis-tributional misalignment and can even impair classification performance. By comparing intra-session baselines, OOD tests, pooled training, and LOSO-IPA under both nonlinear (3D CNN) and linear (PCA + logistic regression) pipelines, we showed that minimal per-subject adaptation and lightweight domain-alignment methods (diagonal CORAL, similarity-weighted augmenta-tion) offer the most reliable path to robust cross-session severity as-

sessment. Looking forward, richer domain-adaptation frame-works (e.g., adversarial alignment, contrastive learning) and explicit longitudinal architectures promise to capture the complex, nonlinear shifts inherent to recovery trajectories. Additionally, principled sample-selection and neuroanatomically guided synthet-ic augmentation can further enhance model generalization without introducing spurious variance. Crucially, integrating demo-graphic and clinical metadata will allow subgroup analyses and fairness audits, safeguarding equity across diverse patient popula-tions. Together, these findings and proposed extensions chart a clear, minimally invasive roadmap for leveraging heterogeneous, longitudinal MRI datasets in TBI maximizing prognostic accuracy while mitigating the risks of unintended distributional drift.

Acknowledgement We extend our deepest gratitude to Sidhika Balachandar for her invaluable feedback and thoughtful discussions throughout this project. Special thanks go to Divyanshu Tak, whose year-long dedication and expertise in brain MRI preprocessing and post-processing were instrumental to our work. We also thank the present and past members of Wood Neuro Research group: Dr. Sossena Wood, Joel Disu, Sid Roy, Lara Abdelmohsen, Nahom Mossazghi and Elizabeth Meinert-Spyker for the mentorship and many insightful discussions. Finally, we acknowledge Roy et al. (2017) for publicly releasing the ds00022 dataset, without which this study would not have been possible [17].

References

1. V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.
2. J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.
3. Z. Shen, Y. Li, and A. Kumar. The Data Addition Dilemma: When More Training Data Hurts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2690–2702, 2024.
4. J. M. Wardlaw *et al.* Multi-Center Harmonization of MRI Biomarkers in Traumatic Brain Injury: Challenges and Opportunities. *NeuroImage: Clinical*, 34:102992, 2022.
5. A. R. Mayer and M. V. Mannell. Functional Neuroimaging of Mild Traumatic Brain Injury. *Journal of Head Trauma Rehabilitation*, 31(1):112–121, 2016.
6. B. Sun and K. Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *European Conference on Computer Vision (ECCV) Workshops*, 2016.
7. I. Y. Chen *et al.* Heterogeneity and fairness in clinical machine learning. *Simons Institute Lecture*, 2024.
8. J.-P. Fortin *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 167:181–200, 2018.
9. E. L. Yuh *et al.* Diffusion tensor imaging for outcome prediction in mild TBI: A TRACK-TBI study. *Journal of Neurotrauma*, 31(17):1457–1477, 2014.
10. A. Sidaros *et al.* Longitudinal diffusion tensor imaging after severe TBI: Correlations with neuropsychological tests. *NeuroImage*, 40(2):688–700, 2008.
11. H. Guan and M. Liu. Domain adaptation in medical image analysis: A survey. *IEEE Transactions on Medical Imaging*, 40(12):2941–2965, 2021.

12. J. Ma *et al.* CORAL-based domain adaptation for cardiac MRI segmentation. *Medical Image Analysis*, 87:102805, 2023.
13. O. Esteban *et al.* MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *NeuroImage*, 145:437–449, 2017.
14. N. Kriegeskorte *et al.* Circular analysis in systems neuroscience. *Nature Neuroscience*, 12:535–540, 2009.
15. H. Jiang *et al.* Predictive uncertainty estimation by deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–12, 2018.
16. S. Hong *et al.* Out-of-distribution detection in medical imaging: A survey. *arXiv preprint* arXiv:2404.18279, 2024.
17. OpenMIBOOD Benchmark. Out-of-distribution detection in medical imaging. In *CVPR Workshops*, 2025.
18. A. Roy, R. A. Bernier, J. Wang, M. Benson, J. French, D. Good, and F. G. Hillary. The evolution of cost-efficiency in neural networks during recovery from traumatic brain injury. *PLOS ONE*, 12(1):e0170541, 2017. Data obtained from the OpenfMRI database (accession number ds00022).