Adaptive Adversarial Data Augmentation with Trajectory Constraint for Alzheimer's Disease Conversion Prediction

Hyuna Cho[1], Hayoung Ahn[1], Guorong Wu[2], and Won Hwa Kim[1]

[1] Pohang University of Science and Technology (POSTECH)
{hyunacho, wonhwa}@postech.ac.kr
[2] University of North Carolina at Chapel Hill

Abstract. Distinguishing progressive mild cognitive impairment (pMCI) from stable MCI (sMCI) is crucial for timely treatment of Alzheimer's disease (AD), yet it is challenging due to inherent class imbalance and limited data. While recent data synthesis methods have shown successful results, they often disregard distributional differences between groups and individual heterogeneity in disease progression. Also, they treat the whole-brain as a unified entity, overlooking region-specific features despite their varying associations with AD. To address this, we propose a novel end-to-end framework that augments MCI data and predicts their future conversion to AD. This is realized by using adversarial attacks that directly control data points in the feature space considering group differences. The attacks are adaptively applied with region-wise learnable attack intensities and subject-specific attack steps, which are flexibly adjusted based on each subject's observation interval. Moreover, we introduce a trajectory constraint that ensures the attacked (i.e., augmented) data follow plausible disease progressions and preserve realistic neurodegeneration patterns. Extensive validations on two AD biomarkers across three classifiers show our method's superiority over six baselines.

Keywords: Adversarial attack · Data augmentation · Data imbalance.

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder and a leading cause of dementia [15, 32]. Before the onset of AD, patients first experience mild cognitive impairment (MCI) [20], a transitional stage between normal aging and AD. While some MCI patients remain cognitively stable over time (sMCI), others progress to AD (pMCI), making early identification of high-risk patients crucial for timely intervention. However, the annual conversion rate from MCI to AD is only 5-10% and many subjects with MCI do not progress to AD even after 10 years of follow-up [18, 24]. This naturally causes a substantial class imbalance between sMCI and pMCI in neuroimaging studies [5, 17, 22], which eventually hinders the development of reliable predictive models for pMCI identification.

	cohort	mean	std
Amyloid	sMCI	1.81	0.36
	pMCI	2.08	0.38
	AD	2.09	0.39
FDG	sMCI	0.83	0.22
	pMCI	0.80	0.22
	AD	0.77	0.22

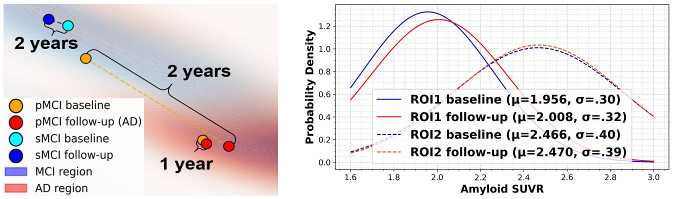

Fig. 1: Group-, subject- and ROI-level data analyses in the ADNI study [22]. (Left) Amyloid and FDG SUVR statistics for the sMCI, pMCI, and AD cohorts. (Middle) t-SNE visualization of disease trajectories based on Amyloid SUVR, showing distinct progression patterns across subjects. (Right) Amyloid SUVR distributions of two ROIs in the pMCI cohort, showing regional variations in longitudinal disease progression.

To address this issue, data augmentation and generative models have been studied. However, typical augmentation methods such as noise injection [30, 34] and interpolation [21, 23] overlook distributional differences between groups and individual variability in disease progression. As shown in the left panel of Fig. 1, AD biomarkers in the ADNI study [22] exhibit sequential distributional shifts from sMCI to pMCI to AD, indicating that pMCI inherently has more AD-related features than sMCI. Moreover, MCI-to-AD transition speeds vary across subjects, where patients who develop AD earlier tend to locate closer to AD in the disease spectrum compared to those who progress more slowly, as shown in the second panel of Fig. 1. Yet, existing augmentation methods disregard these variabilities and apply generic transformations (e.g., adding random noises or blending pMCI data), without considering group- and individual-level heterogeneity.

On the other hand, generative models such as GANs [6, 31], VAEs [14, 29], and diffusion models [3, 11, 28] can learn these differences with a conditional training scheme by using disease stages and observational intervals as conditions. However, they often overlook brain regional variations in disease progression. As illustrated in the right panel of Fig. 1, different brain regions of interest (ROIs) have distinct longitudinal changes, where ROI 1 (*the left paracentral lobule and sulcus*) undergoes more pronounced shifts than ROI 2 (*the right middle frontal sulcus*) within the same observation window. Despite this regional heterogeneity, generative models treat the whole brain as a unified entity, failing to capture localized dynamics within the brain.

To address these limitations, we propose a novel data augmentation method for pMCI data synthesis that explicitly accounts for global differences between sMCI and pMCI, individual variability, and region-specific dynamics in disease progression. This is realized by using adversarial attacks [7, 16], which directly manipulate data points in feature space by iteratively applying small yet strategic perturbations. In our work, these perturbations are designed to gradually push samples along the disease trajectory by considering the relative positions between diagnostic groups. Unlike existing works [2, 9, 10, 13, 19, 27] that use fixed attack steps and magnitudes, our method dynamically adjusts the number of attack steps based on each subject's observation interval. This ensures that shorter-interval samples receive less perturbation, while longer-interval samples undergo more transformations, effectively capturing individual heterogeneity in disease progression. Also, we use ROI-wise learnable attack magnitudes, which allow the model to capture diverse localized progression patterns across different brain regions. The biological plausibility of the synthesized samples is further ensured by our proposed trajectory consistency constraint, which aligns the attack pathways with realistic disease trajectories. The augmented pMCI data are then combined with the given MCI dataset to predict future AD conversion of MCI patients.

Contributions of our work: **1)** We propose a novel end-to-end framework for synthesizing small-size pMCI data and predicting their AD conversion. **2)** By using adversarial attacks with adaptive steps and ROI-wise trainable perturbation intensities, realistic samples are augmented while preserving individual and brain regional heterogeneity in neurodegeneration. **3)** We introduce trajectory consistency regularization that ensures the augmented data follow plausible disease progression. As a result, the synthesized data preserve biological plausibility and enhance downstream predictive performance, outperforming six augmentation and generative methods across two AD biomarkers and three classifiers.

2 Method

In this section, we introduce a data augmentation method to address class imbalance in longitudinal data. Since longitudinal data inherently exhibit temporal variations in both features and labels over time, leveraging such dynamics is the key to effective augmentation. Specifically, in this work, our method is applied to synthesize underrepresented pMCI samples to improve the classification performance of sMCI and pMCI patients. Given that pMCI samples undergo temporal changes with evolving diagnostic labels and ages, our approach incorporates such variations to generate realistic data that align with plausible disease progression.

2.1 Problem Definition for pMCI and sMCI Classification

Consider a longitudinal sequence of samples $\mathbf{X} = \{X_b, X_f\}$, where $X_b = \{x_{b,n}\}_{n=1}^N$ is a set of brain measurements from N ROIs at baseline time point and $X_f = \{x_{f,n}\}_{n=1}^N$ is a set of follow-up brain measurements. For each time point, the samples come with age a_b and a_f, where the age difference is less than 3 years (i.e., $a_f - a_b < 3$). All baseline samples X_b are diagnosed as MCI, while follow-up samples X_f are either MCI or AD. If a patient remains MCI at follow-up, the label Y of $\mathbf{X}$ is defined as stable MCI (sMCI, $Y = 0$), otherwise if the patient progresses to AD, the label is defined as progressive MCI (pMCI, $Y = 1$). Our goal is to classify Y based *solely on* X_b *and* a_b, without any information from the follow-up. In other words, a classifier $f_\theta(X_b, a_b)$ predicts whether MCI patients will convert to AD within 3 years, without knowing the X_f and a_f.

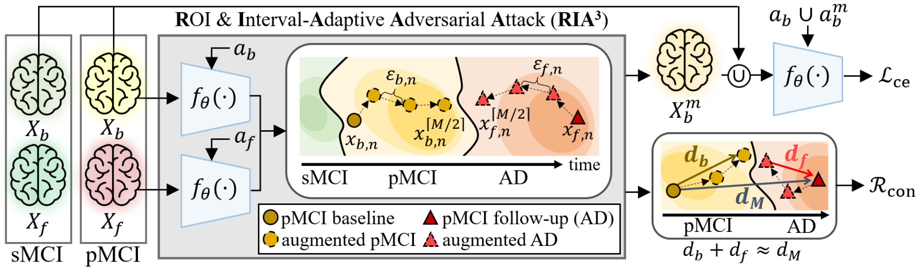

Fig. 2: Overview of the model structure. pMCI samples are augmented via adversarial attacks with ROI-wise trainable intensities $\varepsilon_{b,n}$, $\varepsilon_{f,n}$ and adaptive attack steps determined by individual observation interval M. The pMCI baseline $x_{b,n}$ is perturbed away from sMCI (i.e., relatively towards AD), while the follow-up $x_{f,n}$ shifts toward sMCI. A trajectory constraint R_{con} allows the attacked data $x_{b,n}^m$, $x_{f,n}^m$ to follow realistic neurodegeneration patterns. A downstream classifier $f_\theta(\cdot)$ is trained on the merged dataset of X_b^m and X_b, which has a balanced data size for sMCI and pMCI classification.

2.2 Adaptive Adversarial Attack for pMCI Data Augmentation

Given a population of $\mathbf{X}$ with $N_{pMCI}{=}P$ and $N_{sMCI}{=}S$ subjects, one of the main challenges in AD conversion prediction is *sample imbalance* ($P < S$). To address this, we introduce a data augmentation strategy for pMCI data synthesis using adversarial attacks. As illustrated in Fig. 2, our method performs bidirectional attacks that simultaneously perturb X_b towards AD and X_f towards MCI in the feature space. These attacks are ROI and Interval-Adaptive Adversarial Attack (RIA^3), which applies trainable ROI-wise perturbation intensities with adaptive attack steps adjusted based on individual observation interval.

To account for heterogeneous progression timelines among subjects, the number of attack steps is dynamically scaled based on the monthly interval $M = (a_f - a_b) \times 12$. Specifically, iterative attack is applied for $\lceil M/2 \rceil$ steps to each of X_b and X_f, ensuring that shorter-interval samples receive smaller perturbations, while longer-interval samples have larger adjustments to simulate gradual progression. This design is motivated by the observation that shorter-interval pMCI samples already share more characteristics with their opposing class (as illustrated in Fig 1), making them inherently closer to their target distribution. Consequently, they require fewer attack steps, whereas longer-interval samples require more perturbations.

For attack steps $m = 1, \ldots, \lceil M/2 \rceil$, we use monthly-updated ages a_b^m and a_f^m as conditions to enable a classifier $f_\theta(\cdot)$ to consider age-dependent variations. Since the attacks on X_b and X_f are applied in the opposite direction, their ages are updated accordingly as $a_b^m = a_b + \frac{m}{12}$ and $a_f^m = a_f - \frac{m}{12}$. Given initial data $X_b^0 = X_b$, $X_f^0 = X_f$ with ages $a_b^0 = a_b$, $a_f^0 = a_f$, perturbed data X_b^m and X_f^m are obtained by iteratively applying adversarial perturbations δ_b^m and δ_f^m as follows:

$$\begin{aligned} X_b^{m+1} &= X_b^m + \delta_b^m = X_b^m + |\varepsilon_b|\,\texttt{sign}(\nabla_{X_b^m} L_{\text{ce}}(f_\theta(X_b^m, a_b^m), Y=0)) \quad \text{and} \\ X_f^{m+1} &= X_f^m + \delta_f^m = X_f^m - |\varepsilon_f|\,\texttt{sign}(\nabla_{X_f^m} L_{\text{ce}}(f_\theta(X_f^m, a_f^m), Y=0)), \end{aligned} \tag{1}$$

where L_{ce} is a standard cross-entropy and $\varepsilon_b, \varepsilon_f \in \mathbb{R}^N$ are trainable perturbation magnitudes that adaptively control attack strengths across different ROIs.

Eq. (1) denotes that X_b^m is gradually pushed away from sMCI ($Y = 0$) by a perturbation $\delta_b^m = \text{argmax}_\delta(L_{\text{ce}}(f_\theta(X_b^m, a_b^m), Y = 0))$ that *maximizes* a loss. This iterative attack enhances AD-related features in X_b^m, which are discriminative characteristics that help distinguish sMCI and pMCI. Note that, since $f_\theta(\cdot)$ only classifies baseline data and does not explicitly learn AD data (i.e., pMCI follow-up), sMCI serves as a target label to perturb pMCI baselines X_b. In contrast, $\delta_f^m = \text{argmin}_\delta(L_{\text{ce}}(f_\theta(X_f^m, a_f^m), Y = 0))$ guides the classifier to classify the pMCI follow-up as sMCI by *minimizing* a loss, making X_f^m exhibit more MCI-related features. In both cases, the ROI-wise perturbation magnitudes ε_b and ε_f are shared across all pMCI subjects and thus trained to capture general monthly changes between MCI and AD.

2.3 Trajectory-Constrained Adversarial Training

To ensure that the perturbed samples follow the natural disease progression, we introduce *trajectory consistency regularization*, which constrains the total adversarial displacement to align with the real longitudinal trajectories. Let $d_M = X_f - X_b$ denote the true difference between the observed samples X_b and X_f over M months, and d_b and d_f be the adversarial trajectory displacements at baseline and follow-up, respectively. The d_b and d_f are defined as

$$d_b = X_b^{\lceil \frac{M}{2} \rceil} - X_b = \sum_{m=1}^{\lceil \frac{M}{2} \rceil} \delta_b^m \quad \text{and} \quad d_f = X_f - X_f^{\lceil \frac{M}{2} \rceil} = -\sum_{m=1}^{\lceil \frac{M}{2} \rceil} \delta_f^m, \tag{2}$$

where $X_b^{\lceil \frac{M}{2} \rceil} = X_b + \sum_{m=1}^{\lceil \frac{M}{2} \rceil} \delta_b^m$ and $X_f^{\lceil \frac{M}{2} \rceil} = X_f + \sum_{m=1}^{\lceil \frac{M}{2} \rceil} \delta_f^m$ according to Eq. (1). To maintain a smooth transition between $X_b^{\lceil \frac{M}{2} \rceil}$ and $X_f^{\lceil \frac{M}{2} \rceil}$, the total adversarial displacement approximates the expected disease progression, i.e., $d_b + d_f \approx d_M$, using a trajectory consistency regularization, defined as

$$R_{\text{con}}(\theta, \varepsilon_b, \varepsilon_f) = \frac{1}{P} \sum^{P} ||d_M - (d_b + d_f)||_{l2} = \frac{1}{P} \sum^{P} ||X_f^{\lceil \frac{M}{2} \rceil} - X_b^{\lceil \frac{M}{2} \rceil}||_{l2}, \tag{3}$$

where P is the number of pMCI subjects. The R_{con} penalizes perturbed data deviating from the natural trajectory, preventing unrealistic adversarial shifts.

2.4 Training a Classifier for AD Conversion Prediction

Among the perturbed pMCI data X_b^m, we randomly select $P' = S - P$ instances to balance the number of sMCI and pMCI samples. The selected pMCI data X_b^m and corresponding ages a_b^m are combined with the original data X_b and a_b, forming a dataset of $S + P + P'$ samples. This merged dataset is then used to train a classifier $f_\theta(\cdot)$, which outputs a prediction $\hat{Y}$ for sMCI and pMCI classification. The classifier is trained using the following cross-entropy loss:

$$L_{\text{ce}}(\theta, \varepsilon_b) = -\frac{1}{S + P + P'} \sum_{i=1}^{S+P+P'} \Big(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log(1 - \hat{Y}_i) \Big). \tag{4}$$

Table 1: Demographics of MCI subjects based on Amyloid and FDG SUVR measurements in the ADNI study.

Category	Amyloid		FDG	
	sMCI	pMCI	sMCI	pMCI
Number of subjects	631	98	1827	459
Gender (M / F)	362 / 269	53 / 45	1230 / 597	278 / 181
Age (mean ± std)	72.2 ± 7.3	74.1 ± 7.2	75.2 ± 7.6	75.0 ± 7.2
Monthly interval (mean ± std)	24.2 ± 3.7	25.0 ± 3.2	15.8 ± 8.7	20.5 ± 8.7

To ensure realistic transitions in adversarially augmented data, we incorporate the R_{con} in the final training loss $\mathcal{L} = L_{\text{ce}}(\theta, \varepsilon_b) + \alpha R_{\text{con}}(\theta, \varepsilon_b, \varepsilon_f)$ with a hyperparameter α, such that the network parameters θ and perturbation magnitudes ε_b, ε_f are jointly optimized to improve AD conversion prediction in MCI patients.

3 Experiment

3.1 Experimental Setup

Dataset. We conducted experiments on two AD biomarkers: Standardized Uptake Value Ratio (SUVR) of Amyloid and fluorodeoxyglucose (FDG) provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) [22], whose demographics are reported in Table 1. Both biomarkers were obtained from positron emission tomography (PET) and measured across 148 ROIs on the Destrieux atlas [4]. All subjects have two time points, and we split 80% of the subjects for training and the rest 20% for testing, ensuring an equal proportion of sMCI and pMCI.
Setup. As baselines, we used both data augmentation and generative models, including Logit Uncertainty (LU) [12], SMOTE [1], Mixup [33], CTGAN [31], TVAE [14], and DDPM [11]. All generative models were trained in a conditional scheme, using labels, ages, and monthly intervals as conditions. As in ours, all methods generated P' number of pMCI data, which were combined with the given train set for downstream AD conversion prediction. To evaluate the effectiveness and generalizability of the synthesized data, we used three classifiers: Multi-Layer Perceptron (MLP) [26], FT-Transformer [8], and NODE [25]. In all settings, we reported averaged accuracy and F1-score across three replicates using different parameter initialization. All methods were fine-tuned via a grid search of learning rates, and their best results were reported.

3.2 Quantitative Results

As shown in Table 2, RIA^3 surpassed all baselines on both datasets. Specifically, on the Amyloid dataset with MLP, our method achieved 87.67% in accuracy and 61.27% in F1-score, surpassing the second-best results by $\sim$3.7%p and $\sim$4.5%p, respectively. Notably, in both datasets, training with synthesized data from generative models often results in lower F1-scores compared to training without any

Table 2: Performance comparison of RIA^3 and baseline methods.

Methods	MLP [26]		FT-Transformer [8]		NODE [25]	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Amyloid						
No Aug.	$80.59^{\pm 1.7}$	$54.11^{\pm 1.6}$	$80.36^{\pm 0.8}$	$54.27^{\pm 1.0}$	$82.19^{\pm 1.2}$	$54.14^{\pm 0.7}$
LU [12]	$81.05^{\pm 6.7}$	$53.99^{\pm 6.7}$	$83.33^{\pm 1.4}$	$55.24^{\pm 1.4}$	$83.10^{\pm 0.3}$	$54.90^{\pm 1.1}$
SMOTE [1]	$83.56^{\pm 2.4}$	$\underline{56.78}^{\pm 6.4}$	$87.67^{\pm 3.8}$	$57.07^{\pm 3.0}$	$83.56^{\pm 4.5}$	$52.52^{\pm 7.0}$
Mixup [33]	$82.65^{\pm 2.4}$	$55.43^{\pm 2.7}$	$84.25^{\pm 1.8}$	$57.73^{\pm 1.1}$	$83.33^{\pm 3.9}$	$\underline{55.04}^{\pm 3.6}$
CTGAN [31]	$\underline{84.02}^{\pm 5.5}$	$53.33^{\pm 0.0}$	$83.11^{\pm 2.4}$	$56.04^{\pm 1.2}$	$83.11^{\pm 6.9}$	$50.38^{\pm 3.8}$
TVAE [14]	$80.59^{\pm 7.9}$	$50.53^{\pm 3.8}$	$86.30^{\pm 4.5}$	$60.17^{\pm 8.2}$	$79.22^{\pm 10.0}$	$44.53^{\pm 5.1}$
DDPM [11]	$83.79^{\pm 4.2}$	$43.83^{\pm 15.5}$	$\underline{89.27}^{\pm 5.0}$	$\underline{61.57}^{\pm 5.6}$	$\underline{85.84}^{\pm 4.6}$	$40.00^{\pm 26.5}$
RIA^3 (Ours)	$\mathbf{87.67}^{\pm 1.4}$	$\mathbf{61.27}^{\pm 4.9}$	$\mathbf{90.19}^{\pm 0.8}$	$\mathbf{61.98}^{\pm 0.8}$	$\mathbf{88.36}^{\pm 1.6}$	$\mathbf{57.91}^{\pm 1.3}$
FDG						
No Aug.	$81.22^{\pm 0.9}$	$65.97^{\pm 0.7}$	$83.40^{\pm 0.8}$	$67.03^{\pm 1.9}$	$81.22^{\pm 0.2}$	$64.06^{\pm 0.8}$
LU [12]	$81.23^{\pm 0.7}$	$66.50^{\pm 0.8}$	$82.31^{\pm 0.8}$	$67.65^{\pm 0.9}$	$81.00^{\pm 0.2}$	$\underline{66.24}^{\pm 0.2}$
SMOTE [1]	$\underline{83.77}^{\pm 2.4}$	$67.10^{\pm 1.6}$	$\underline{84.13}^{\pm 0.9}$	$66.45^{\pm 1.2}$	$82.02^{\pm 1.7}$	$63.87^{\pm 0.8}$
Mixup [33]	$83.12^{\pm 1.5}$	$\underline{67.35}^{\pm 0.7}$	$84.06^{\pm 0.6}$	$\underline{67.94}^{\pm 0.2}$	$\underline{82.17}^{\pm 1.4}$	$66.13^{\pm 0.4}$
CTGAN [31]	$82.31^{\pm 1.7}$	$62.79^{\pm 3.0}$	$83.26^{\pm 1.3}$	$63.62^{\pm 2.4}$	$75.47^{\pm 10.8}$	$35.40^{\pm 31.3}$
TVAE [14]	$80.28^{\pm 4.0}$	$62.08^{\pm 3.1}$	$83.55^{\pm 0.3}$	$62.39^{\pm 3.1}$	$80.93^{\pm 3.1}$	$59.11^{\pm 3.0}$
DDPM [11]	$81.15^{\pm 2.3}$	$63.02^{\pm 2.8}$	$83.12^{\pm 2.0}$	$64.27^{\pm 2.4}$	$81.59^{\pm 2.3}$	$60.82^{\pm 1.8}$
RIA^3 (Ours)	$\mathbf{85.01}^{\pm 0.9}$	$\mathbf{68.21}^{\pm 0.5}$	$\mathbf{85.88}^{\pm 0.5}$	$\mathbf{68.72}^{\pm 0.5}$	$\mathbf{84.06}^{\pm 0.8}$	$\mathbf{66.37}^{\pm 0.1}$

augmentation (i.e., ‘No Aug.’). This suggests that the generative models struggle to learn a generalized pMCI data distribution, likely due to the small sample size. As a result, the generated data may lack realism and diversity, causing classification bias toward either pMCI or sMCI. In contrast, RIA^3 effectively synthesizes generalized and diverse pMCI samples using strategic adversarial attacks that directly push data points away from sMCI in feature space, leading to improved classification performance.

3.3 Model Behavior Analysis and Ablation Study

Effect of $R_{\mathbf{con}}$. Fig. 3 demonstrates that the perturbed samples X_b^m and X_f^m exhibit realistic disease trajectories, seamlessly changing from the baseline X_b to follow-up X_f. By using R_{con}, the most perturbed samples $X_f^{\lceil \frac{M}{2} \rceil}$ and $X_b^{\lceil \frac{M}{2} \rceil}$ shift smoothly, indicating that the learned perturbations are not arbitrary but rather preserve underlying disease progression characteristics between MCI and AD. The R_{con} not only enhances the reliability of augmented data but also improves classifier performance, as shown in Table 3. Notably, FT-Transformer and NODE achieve peak performance at $\alpha = 0.1$, surpassing the setting without R_{con} (i.e., $\alpha = 0$) by 4.35%p and 5.48%p in accuracy. MLP performs best at $\alpha = 1$, achieving 5.93%p improvement in accuracy over $\alpha = 0$, highlighting the importance of R_{con} in capturing generalized and discriminative pMCI features.

Discussion on trained $\boldsymbol{\varepsilon}_b, \boldsymbol{\varepsilon}_f$. As shown in Fig. 4, trained perturbation magnitudes $|\varepsilon_b|$ and $|\varepsilon_f| \in \mathbb{R}^N$ converged adaptively across different brain regions. Moreover, the regional changes differ from the baseline and follow-up, indicating

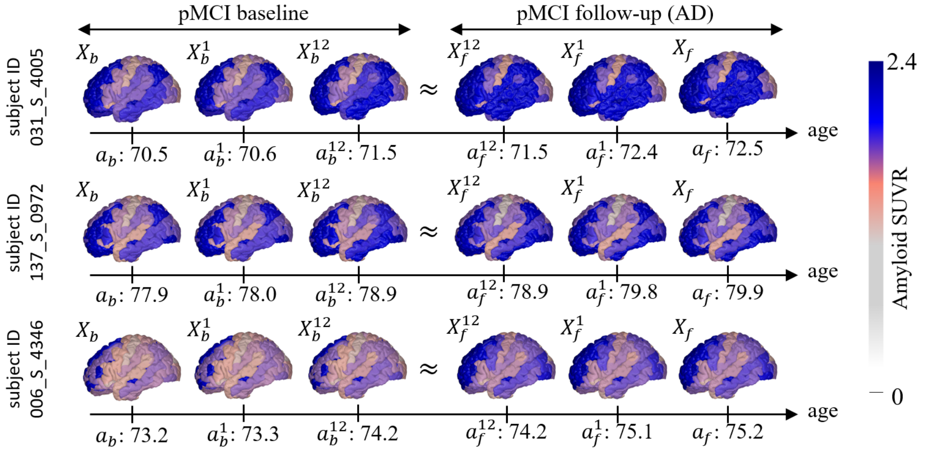

Fig. 3: Visualization of given data $\{X_b, X_f\}$ and perturbed data X_b^m and X_f^m along the disease trajectories of three pMCI subjects. Note that the real data at the far ends are highly dissimilar, whereas augmented samples in the middle share similar traits.

Table 3: Ablation study on the weight α of R_{con} on the Amyloid dataset.

α	MLP [26]		FT-Transformer [8]		NODE [25]	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
0	$81.74^{\pm 0.4}$	$56.05^{\pm 0.5}$	$85.84^{\pm 1.6}$	$56.38^{\pm 0.7}$	$82.88^{\pm 0.7}$	$56.55^{\pm 0.5}$
0.01	$\underline{87.44}^{\pm 1.4}$	$58.08^{\pm 3.3}$	$85.62^{\pm 6.0}$	$\underline{60.48}^{\pm 5.4}$	$\underline{88.13}^{\pm 1.4}$	$57.24^{\pm 2.3}$
0.05	$87.22^{\pm 1.4}$	$\underline{59.95}^{\pm 2.2}$	$87.45^{\pm 3.2}$	$59.61^{\pm 2.9}$	$85.04^{\pm 1.2}$	$57.39^{\pm 1.1}$
0.1	$85.85^{\pm 1.4}$	$58.72^{\pm 2.3}$	$\mathbf{90.19}^{\pm 0.8}$	$\mathbf{61.98}^{\pm 0.8}$	$\mathbf{88.36}^{\pm 1.2}$	$\underline{57.91}^{\pm 1.3}$
0.5	$87.21^{\pm 1.0}$	$59.44^{\pm 2.4}$	$\underline{88.59}^{\pm 1.0}$	$58.25^{\pm 1.9}$	$85.39^{\pm 1.0}$	$57.88^{\pm 0.6}$
1	$\mathbf{87.67}^{\pm 1.4}$	$\mathbf{61.27}^{\pm 4.9}$	$86.99^{\pm 1.2}$	$58.96^{\pm 0.9}$	$85.84^{\pm 1.6}$	$\mathbf{59.27}^{\pm 1.6}$

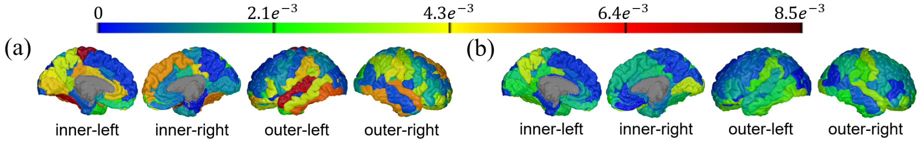

Fig. 4: Visualization of the trained perturbation magnitudes (a) $|\varepsilon_b|$ and (b) $|\varepsilon_f|$.

that the adversarial perturbations adapt to distinct data distributions observed at different disease stages. Notably, the trained $|\varepsilon_b|$ is generally larger than $|\varepsilon_f|$, i.e., the mean of $|\varepsilon_b|$ and $|\varepsilon_f|$ are $2.3e^{-3}$ (std: $2.1e^{-3}$) and $1.1e^{-3}$ (std: $8.3e^{-4}$), respectively. These results suggest that the model applies stronger shifts on X_b to emphasize AD-specific features. Using X_b^m with such pronounced disease-associated characteristics enhances the difference between sMCI and pMCI, and thus improves downstream AD conversion prediction.

4 Conclusion

In this work, we proposed a novel data augmentation method to address the data imbalance issue in sMCI and pMCI classification. Leveraging adversarial attacks with subject-wise adaptive step sizes and ROI-wise learnable attack magnitudes, our method captures heterogeneous neurodegeneration patterns across subjects and brain regions. Moreover, our proposed trajectory constraint ensures that the synthesized samples follow natural disease progression, enhancing their biological plausibility and downstream predictive performance. Extensive experiments across multiple datasets and classifiers validate the effectiveness of our method.

Acknowledgments. This research was supported by the AI Graduate Program at POSTECH (No. RS-2019-II191906, 10%), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02216257, 20% and No. NRF-2022R1A2C2092336, 70%).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chawla, N.V., et al.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
2. Cho, H., et al.: Anti-adversarial consistency regularization for data augmentation: Applications to robust medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 555–566. Springer (2023)
3. Cho, H., et al.: Conditional diffusion with ordinal regression: Longitudinal data generation for neurodegenerative disease studies. In: International Conference on Learning Representations (ICLR) (2025)
4. Destrieux, C., et al.: Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage **53**(1), 1–15 (2010)
5. Ellis, K.A., et al.: The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
6. Goodfellow, I., et al.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
7. Goodfellow, I.J., et al.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015)
8. Gorishniy, Y., et al.: Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems (NeurIPS) **34**, 18932–18943 (2021)
9. Guo, C., et al.: Simple black-box adversarial attacks. In: International Conference on Machine Learning (ICML). pp. 2484–2493. PMLR (2019)
10. Hirano, H., et al.: Universal adversarial attacks on deep neural networks for medical image classification. BMC medical imaging **21**, 1–13 (2021)
11. Ho, J., et al.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) **33**, 6840–6851 (2020)

12. Hu, Y., et al.: Data augmentation in logit space for medical image classification with limited training data. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 469–479. Springer (2021)
13. Jeong, M., et al.: Uncertainty-aware diffusion-based adversarial attack for realistic colonoscopy image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 647–658. Springer (2024)
14. Kingma, D.P.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Larson, E.B., et al.: Cognitive impairment: dementia and Alzheimer's disease. Annual review of public health **13**, 431–449 (1992)
16. Madry, A., et al.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
17. Marcus, D.S., et al.: Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
18. Mitchell, A.J., Shiri-Feshki, M.: Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. Acta psychiatrica scandinavica **119**(4), 252–265 (2009)
19. Moosavi-Dezfooli, S.M., et al.: Universal adversarial perturbations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1765–1773 (2017)
20. Morris, J.C., et al.: Mild cognitive impairment represents early-stage Alzheimer disease. Archives of neurology **58**(3), 397–405 (2001)
21. Mroueh, Y., et al.: Fair mixup: Fairness via interpolation. In: International Conference on Learning Representations (ICLR) (2021)
22. Mueller, S.G., et al.: The Alzheimer's disease neuroimaging initiative. Neuroimaging Clinics **15**(4), 869–877 (2005)
23. Oh, C., et al.: Time-series data augmentation based on interpolation. Procedia Computer Science **175**, 64–71 (2020)
24. Petersen, R.C., et al.: Mild cognitive impairment: ten years later. Archives of neurology **66**(12), 1447–1455 (2009)
25. Popov, S., et al.: Neural oblivious decision ensembles for deep learning on tabular data. International Conference on Learning Representations (ICLR) (2020)
26. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6), 386 (1958)
27. Shafahi, A., et al.: Universal adversarial training. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 34, pp. 5636–5643 (2020)
28. Song, J., et al.: Denoising diffusion implicit models. International Conference on Learning Representations (ICLR) (2021)
29. Vahdat, A., Kautz, J.: Nvae: A deep hierarchical variational autoencoder. Advances in Neural Information Processing Systems (NeurIPS) **33**, 19667–19679 (2020)
30. Xie, Q., et al.: Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems (NeurIPS) **33**, 6256–6268 (2020)
31. Xu, L., et al.: Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)
32. Yaari, R., Corey-Bloom, J.: Alzheimer's disease. In: Seminars in neurology. vol. 27, pp. 032–041 (2007)
33. Zhang, H., et al.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (ICLR) (2018)

34. Zhong, Z., et al.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 34, pp. 13001–13008 (2020)