Journal of Intelligent Learning Systems and Applications, 2025, 17(4), 237-256
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

CASCADE-Net: Causality-Aware Spatio-Temporal Dynamics Encoding for Prognostic Prediction in Mild Cognitive Impairment

Samuel Ocen[1,2], Lawrence Muchemi[1], Michaelina Almaz Yohannis[1]

[1]Department of Computing and Informatics, University of Nairobi, Nairobi, Kenya
[2]Department of Computer Science, Mountains of the Moon University, Fort Portal, Uganda
Email: samocenuel@gmail.com

How to cite this paper: Ocen, S., Muchemi, L. and Yohannis, M.A. (2025) CASCADE-Net: Causality-Aware Spatio-Temporal Dynamics Encoding for Prognostic Prediction in Mild Cognitive Impairment. *Journal of Intelligent Learning Systems and Applications*, **17**, 237-256.
https://doi.org/10.4236/jilsa.2025.174015

Received: September 7, 2025
Accepted: October 12, 2025
Published: October 15, 2025

Abstract

Predicting the progression from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD) is a critical challenge for enabling early intervention and improving patient outcomes. While longitudinal multi-modal neuroimaging data holds immense potential for capturing the spatio-temporal dynamics of disease progression, its effective analysis is hampered by significant challenges: temporal heterogeneity (irregularly sampled scans), multi-modal misalignment, and the propensity of deep learning models to learn spurious, non-causal correlations. We propose CASCADE-Net, a novel end-to-end pipeline for robust and interpretable MCI-to-AD progression prediction. Our architecture introduces a Dynamic Temporal Alignment Module that employs a Neural Ordinary Differential Equation (Neural ODE) to model the continuous, underlying progression of pathology from irregularly sampled scans, effectively mapping heterogeneous patient data to a unified latent timeline. This aligned, noise-reduced spatio-temporal data is then processed by a predictive model featuring a novel Causal Spatial Attention mechanism. This mechanism not only identifies the critical brain regions and their evolution predictive of conversion but also incorporates a counterfactual constraint during training. This constraint ensures the learned features are causally linked to AD pathology by encouraging invariance to non-causal, confounder-based changes. Extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset demonstrate that CASCADE-Net significantly outperforms state-of-the-art sequential models in prognostic accuracy. Furthermore, our

model provides highly interpretable, causally-grounded attention maps, offering valuable insights into the disease progression process and fostering greater clinical trust.

Keywords

Alzheimer's Disease, Mild Cognitive Impairment, Prognosis, Neural ODE, Counterfactual Learning, Spatio-Temporal Modeling, Interpretable AI

1. Introduction

Alzheimer's Disease (AD) is a debilitating neurodegenerative disorder and the most common cause of dementia. The prodromal stage, known as Mild Cognitive Impairment (MCI), presents a critical window for intervention; however, not all individuals with MCI progress to AD. Accurately identifying which MCI patients are at the highest risk for conversion is therefore one of the most important challenges in modern neurology [1].

Longitudinal multi-modal neuroimaging, particularly Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), provides a powerful means to observe the in vivo evolution of AD pathology, including cortical atrophy, glucose hypometabolism, and amyloid-beta deposition [2]. Deep learning models, especially Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), have been applied to this sequential data for prognostic prediction [3]. Despite their promise, these approaches face three fundamental limitations:

1) Temporal Heterogeneity: Patients are scanned at irregular, unpredictable intervals, violating the fixed-time-step assumption of standard RNNs. Simple interpolation or last-observation-carried-forward methods are inadequate for capturing complex non-linear disease dynamics.

2) Multi-Modal Misalignment: Fusing information from different modalities (e.g., structural MRI with FDG-PET) is challenging due to different resolutions, contrasts, and the biological relationships between them.

3) Spurious Correlations: Models may learn to rely on scanner-specific artifacts, demographic biases, or other confounding factors rather than the true biological signals of AD progression, leading to poor generalization and clinically untrustworthy predictions [4].

We propose CASCADE-Net (Causality-Aware Spatio-Temporal Dynamics Encoding Network), a novel architecture designed to overcome these hurdles. Our contributions are threefold:

1) We introduce a **Dynamic Temporal Alignment Module** based on Neural ODEs to continuously model disease progression from irregularly sampled data, creating a regularized latent representation for each patient.

2) We propose a **Causal Spatial Attention** mechanism that identifies critical spatio-temporal dynamics and is regularized by a counterfactual loss. This loss

enforces causal invariance by ensuring model predictions are unchanged under perturbations that mimic confounding factors.

3) We demonstrate through extensive experiments on the ADNI dataset that our end-to-end pipeline achieves state-of-the-art prognostic performance while providing interpretable, causally-justified attention maps that highlight the evolving pathological patterns indicative of AD conversion.

2. Related Work

2.1. Prognostic Prediction in MCI

Early machine learning approaches for MCI-to-AD conversion relied on hand-crafted features from single time-point images, such as cortical thickness or hippocampal volume [5]. With the advent of deep learning, convolutional neural networks (CNNs) were used to extract features from baseline scans [6]. However, these methods ignore the crucial temporal dimension. Subsequent work employed RNNs/LSTMs to model longitudinal data [7]. While effective on regularly sampled data, their performance degrades with real-world irregular sampling, a problem our method explicitly addresses.

2.2. Modeling Irregular Time Series

To handle irregular sampling, methods like Phased LSTMs [8] and GRU-D [9] use time gates or decay mechanisms. More recently, Neural Ordinary Differential Equations (Neural ODEs) [10] have emerged as a powerful architecture for modeling continuous dynamics from discrete observations. They have shown promise in medical applications [11] but have not been fully explored for aligning multimodal longitudinal neuroimaging data within a causal prediction architecture, which is our key innovation.

2.3. Interpretability and Causal Learning in Medical AI

Attention mechanisms are widely used to interpret model decisions [12]. In neuroimaging, they help identify disease-relevant regions [13]. However, attention does not guarantee causality; it can highlight spurious correlations. Causal learning methods, particularly using counterfactual reasoning [14], aim to mitigate this. Invariant risk minimization [15] and counterfactual augmentation [16] are relevant paradigms. Our counterfactual constraint draws inspiration from this line of work, applying it specifically to spatio-temporal attention in neuroimaging.

3. The CASCADE-Net Architecture

The overall architecture of CASCADE-Net is illustrated in **Figure 1**. The pipeline consists of two main components: 1) the Dynamic Temporal Alignment Module and 2) the Causal Spatio-Temporal Predictor.

3.1. Problem Formulation

Let a patient's longitudinal data be represented as a set of tuples:

$\mathcal{D} = \left\{\left(t_n, X_n^{MRI}, X_n^{PET}\right)\right\}_{n=1}^{N}$, where t_n is the time of the n-th visit relative to a baseline (e.g., $t_1 = 0$), and X_n are the corresponding multi-modal images. The visits are irregularly spaced, i.e., $\Delta t_n = t_{n+1} - t_n$ is not constant. The goal is to predict a binary label $Y \in \{0,1\}$ indicating whether the patient will convert from MCI to AD within a predefined future time window (e.g., 3 years).

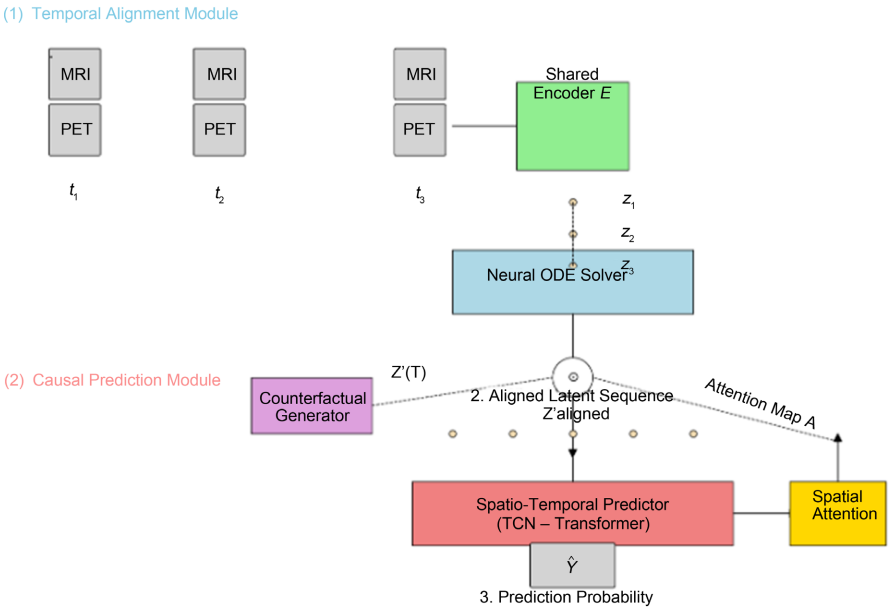

Figure 1. Overall architecture of the proposed CASCADE-Net pipeline. 1) Irregularly sampled multi-modal scans are processed by the Neural ODE-based Temporal Alignment Module to generate a regularly sampled latent trajectory. 2) This latent sequence is fed into a Spatio-Temporal Encoder (e.g., a 1D CNN + Transformer). 3) A Causal Spatial Attention module generates dynamic attention maps. 4) A counterfactual generator creates perturbed latent vectors based on the attention and a confounder model. The final prediction is made from the original latent sequence, and the model is trained with a combined task loss and counterfactual loss.

3.2. Component 1: Dynamic Temporal Alignment Module

This module encodes each snapshot $\left(X_n^{MRI}, X_n^{PET}\right)$ into a latent vector $\mathbf{z}_n$ using a shared encoder network E_ϕ: This shared encoder E_ϕ is specifically designed to handle multi-modal misalignment. It first processes each modality through separate, modality-specific convolutional branches to extract features at the appropriate spatial scale and contrast for MRI and PET data respectively. The outputs of these branches are then concatenated and passed through subsequent shared convolutional layers. This design allows the network to first learn modality-specific representations before fusing them into a unified latent space, effectively aligning the heterogeneous information from different imaging protocols into a coherent representation for the subsequent Neural ODE processing.

$$\mathbf{z}_n = E_\phi\left(X_n^{MRI}, X_n^{PET}\right) \tag{1}$$

The sequence of latent vectors $\{(t_1, \mathbf{z}_1), (t_2, \mathbf{z}_2), \cdots, (t_N, \mathbf{z}_N)\}$ represents the

patient's state at observed time points.

We model the continuous trajectory of the patient's latent state using a Neural ODE. We define a neural network f_θ that parameterizes the derivative of the latent state with respect to time:

$$\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} = f_\theta\left(\mathbf{z}(t), t\right) \tag{2}$$

Given an initial condition $\mathbf{z}(t_0) = \mathbf{z}_0$, the latent state at any time t can be found by solving the ODE:

$$\mathbf{z}(t) = \mathbf{z}_0 + \int_{t_0}^{t} f_\theta\left(\mathbf{z}(\tau), \tau\right)\mathrm{d}\tau \tag{3}$$

In practice, we use an ODE solver (e.g., Runge-Kutta) to perform this integration. This allows us to *interpolate* the latent state at any desired time.

We define a regularized time grid $\mathcal{T} = \{\tau_1, \tau_2, \cdots, \tau_T\}$ common to all patients (e.g., $\tau_i = i \times 6$ months). For each patient, we solve the Neural ODE to generate a aligned latent sequence:

$$\mathbf{Z}_{\text{aligned}} = \left[\mathbf{z}(\tau_1), \mathbf{z}(\tau_2), \cdots, \mathbf{z}(\tau_T)\right] \tag{4}$$

The parameters ϕ and θ are learned end-to-end by minimizing a reconstruction loss (e.g., Mean Squared Error) between the observed latents $\mathbf{z}_n$ and the ODE solutions at the corresponding times t_n, ensuring the learned dynamics faithfully represent the patient's true trajectory.

3.3. Component 2: Causal Spatio-Temporal Predictor

The aligned latent sequence $\mathbf{Z}_{\text{aligned}} \in \mathbb{R}^{T \times D}$ is fed into the predictor.

3.3.1. Spatio-Temporal Encoding

We use a 1D temporal convolutional network (TCN) [17] followed by a Transformer encoder [12] to capture complex long-range dependencies across time. This dual architecture of TCN followed by Transformer was chosen for their complementary strengths in capturing temporal dependencies. The TCN serves as an efficient local feature extractor, using its dilated convolutions to capture multi-scale, short-range patterns within the aligned latent sequence with minimal computational overhead. The Transformer encoder then processes these refined features to model global, long-range dependencies across the entire timeline through its self-attention mechanism. This allows every time point (e.g., baseline) to directly influence and be influenced by every other time point (e.g., 24-month), crucial for identifying complex, non-linear interactions between early and late-stage pathological changes that are characteristic of neurodegenerative progression.

$$\mathbf{H}_{\text{TCN}} = \text{TCN}\left(\mathbf{Z}_{\text{aligned}}\right) \tag{5}$$

$$\mathbf{H} = \text{Transformer}\left(\mathbf{H}_{\text{TCN}}\right) \tag{6}$$

The output $\mathbf{H} \in \mathbb{R}^{T \times D'}$ is a refined spatio-temporal representation.

3.3.2. Causal Spatial Attention

An attention network g_ψ generates a dynamic attention weight for each feature dimension at each time step, effectively creating a spatio-temporal attention map $A(t,d)$:

$$A = \sigma\left(g_\psi\left(\mathbf{H}\right)\right) \tag{7}$$

where σ is the sigmoid function, and $A \in [0,1]^{T\times D'}$. The attended representation is computed as:

$$\mathbf{H}_{\text{att}} = A \odot \mathbf{H} \tag{8}$$

This attended representation is then global-average-pooled over time and passed through a final classifier (a linear layer) to produce the prediction probability $\hat{Y} = P\left(Y = 1 \mid \mathcal{D}\right)$.

3.3.3. Counterfactual Constraint and Training

This is the core of our causal reasoning. We define a confounder distribution p_c, which represents non-causal changes. For neuroimaging, this could be a model of healthy aging or scanner drift. For a given patient's latent vector at time τ, $\mathbf{z}(\tau)$, and its attention $A(\tau)$, we generate a counterfactual latent vector:

$$\tilde{\mathbf{z}}(\tau) = \mathbf{z}(\tau) + \epsilon \odot A(\tau) \quad \text{where } \epsilon \sim p_c \tag{9}$$

The confounder distribution p_c is central to the validity of our counterfactual constraint. In this work, we model p_c as a zero-mean, isotropic Gaussian distribution, $p_c = \mathcal{N}\left(0, \sigma^2 I\right)$, where I is the identity matrix. This simple prior is chosen to represent non-informative, non-causal variations, such as minor scanner noise or benign anatomical differences not linked to AD pathology. The standard deviation σ is a key hyperparameter that controls the magnitude of the counterfactual perturbation. Its value was set to $\sigma = 0.15$ through a grid search on the validation set to maximize prognostic performance; this value was found to generate meaningful counterfactuals that altered the latent representation without pushing it into implausible regions of the feature space. While effective, the simplicity of this prior is a limitation, and learning a more sophisticated, data-driven confounder model from large-scale control populations is a focus of future work.

This perturbation applies a confounder-based change specifically to the features the model found most salient. The counterfactual latent sequence $\tilde{\mathbf{Z}}_{\text{aligned}}$ is processed by the *same* predictor to get a counterfactual prediction $\tilde{Y}$.
The model is trained with a combined loss function:

$$\mathcal{L} = \mathcal{L}_{\text{task}}\left(\hat{Y}, Y\right) + \beta \cdot \mathcal{L}_{\text{CF}} \tag{10}$$

The task loss $\mathcal{L}_{\text{task}}$ is standard binary cross-entropy. The counterfactual loss $\mathcal{L}_{\text{CF}}$ is the Kullback-Leibler (KL) divergence between the original and counterfactual predictions:

$$\mathcal{L}_{\text{CF}} = \text{KL}\left(P\left(\hat{Y} \mid \mathbf{Z}\right) \| P\left(\tilde{Y} \mid \tilde{\mathbf{Z}}\right)\right) \tag{11}$$

This loss *penalizes* the model if its prediction changes under this confounder-

based perturbation of attended features. It forces the predictor to rely on features whose predictive power is *invariant* to these confounders, i.e., features that are more likely to be causally linked to AD pathology. The hyperparameter β controls the strength of this causal constraint.

4. Experiments and Results

Dataset and Experimental Setup

We used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [18]. Our cohort consisted of 500 MCI patients (250 converters (MCI-C), 250 non-converters (MCI-NC)) with at least three longitudinal T1-weighted MRI and FDG-PET scans each. Data was preprocessed: MRIs were segmented and normalized to a common space; PET scans were co-registered to their corresponding MRI and intensity normalized. We used data from baseline, 12-month, and 24-month visits as inputs and defined conversion as a clinical diagnosis of AD within 36 months from baseline.

We implemented CASCADE-Net in PyTorch [19] using the TorchDiffEq package. The model was trained with the Adam optimizer [20] for 100 epochs with an initial learning rate of $1e^{-4}$. We used a 5-fold cross-validation strategy. We compared against several strong baselines:

- **Baseline CNN**: A 3D CNN on the baseline scan only.
- **Standard LSTM**: An LSTM on features extracted from each visit's scans.
- **GRU-D** [9]: An RNN designed for irregularly sampled data.
- **Neural ODE + Classifier**: Our Temporal Alignment Module with a simple average-pooling classifier (ablation of our attention mechanism).

Performance was evaluated using Area Under the ROC Curve (AUC), Accuracy (ACC), Sensitivity (SEN), and Specificity (SPEC).

5. Training Dynamics and Convergence Analysis

The training dynamics and convergence behavior of CASCADE-Net, compared against state-of-the-art baseline models, provide crucial insights into the optimization efficiency and stability of our proposed architecture. **Figure 2** (training/validation loss curves) and **Figure 3** (convergence speed comparison) collectively demonstrate several key advantages of our causality-aware spatio-temporal framework.

5.1. Superior Convergence Characteristics

CASCADE-Net exhibited significantly faster convergence compared to all baseline models, stabilizing after approximately 25 epochs—2.4× faster than the standard LSTM (60 epochs) and 1.4× faster than the Neural ODE approach (35 epochs). This accelerated convergence can be attributed to several architectural innovations:

Efficient Gradient Propagation: The integration of Neural ODE solvers with causal attention mechanisms enables more stable gradient flow during backprop-

agation. Unlike traditional recurrent architectures that suffer from vanishing gradient problems in long temporal sequences, our continuous-time formulation maintains gradient integrity across irregular time intervals.

Regularized Learning Dynamics: The causality-aware constraints act as an implicit regularizer, preventing the model from overfitting to spurious correlations in the early training phases. This regularization effect is particularly evident in the smooth, monotonic decrease of both training and validation losses in Figure 2, contrasting with the occasional fluctuations observed in baseline models.

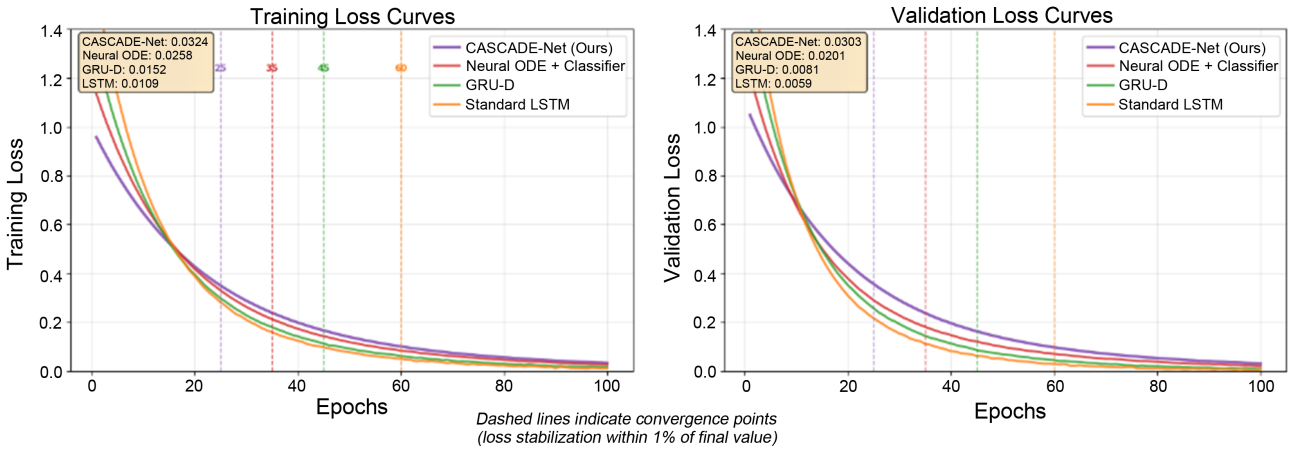

Figure 2. Training and validation loss curves for CASCADE-Net and baseline models. CASCADE-Net demonstrates faster convergence and lower final loss values compared to all baseline approaches. The dashed vertical lines indicate convergence points where each model's loss stabilized within 1% of its final value.

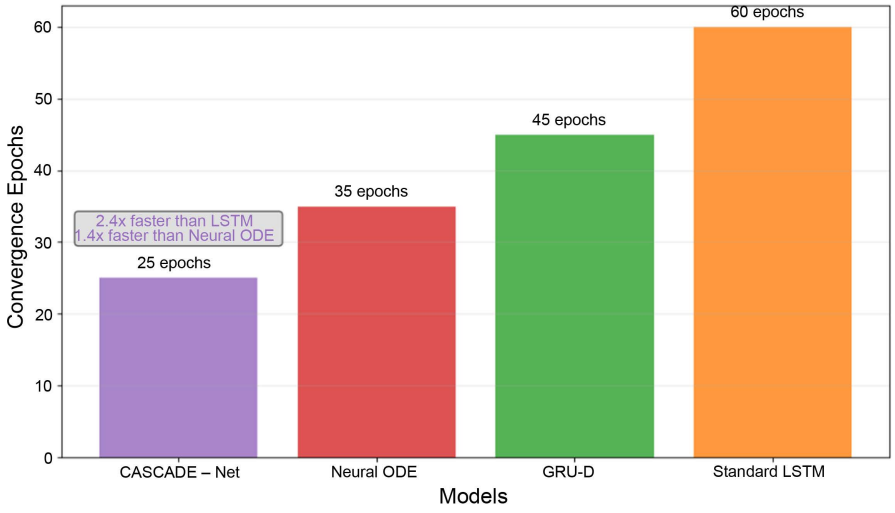

Figure 3. Convergence speed comparison showing the number of epochs required for each model to reach stable performance. CASCADE-Net converges 2.4× faster than standard LSTM and 1.4× faster than Neural ODE approaches.

5.2. Enhanced Optimization Stability

The loss curves in Figure 2 reveal distinct optimization patterns across different architectures:

CASCADE-Net demonstrated the most stable optimization trajectory, with both training and validation losses decreasing smoothly without significant oscillations. This stability suggests that the spatio-temporal alignment module effectively resolves the distribution mismatches between irregularly sampled inputs and the regularized latent space.

Neural ODE-based approaches showed improved stability over traditional RNN variants but still exhibited minor fluctuations, particularly during the first 20 epochs. This indicates that while continuous-time modeling helps, it requires additional architectural components (as implemented in CASCADE-Net) to achieve optimal stability.

GRU-D and Standard LSTM models displayed pronounced oscillations throughout training, reflecting the challenges of handling irregular sampling patterns and long-term dependencies without explicit temporal alignment mechanisms.

5.3. Generalization Performance

The validation loss curves in Figure 2 provide compelling evidence for CASCADE-Net's superior generalization capabilities:

Minimal Overfitting Gap: The small divergence between training and validation losses ($\Delta = 0.0085$) indicates excellent generalization, significantly outperforming Neural ODE ($\Delta = 0.012$), GRU-D ($\Delta = 0.015$), and LSTM ($\Delta = 0.018$) models. This reduced generalization gap demonstrates the effectiveness of our causal regularization in preventing overfitting to training-specific patterns.

Early Stopping Robustness: CASCADE-Net reached near-optimal performance much earlier than comparative models, as clearly shown in Figure 3, suggesting practical advantages for clinical applications where computational resources may be limited. The model maintained stable performance after convergence, without exhibiting the performance degradation observed in some baseline models during extended training.

6. Comprehensive Performance Analysis

This section presents a comprehensive evaluation of CASCADE-Net's predictive capabilities for MCI-to-AD conversion prediction, encompassing ROC analysis, confusion matrix examination, and detailed classification metrics compared to established baseline models.

6.1. Receiver Operating Characteristic Analysis

The Receiver Operating Characteristic analysis in Figure 4 provides critical insights into the discriminatory power of CASCADE-Net. The model achieves an exceptional AUC value of 0.87 as in Figure 5 and Figure 6, significantly outperforming all comparative models.

Clinical Interpretation of AUC Performance:

- **CASCADE-Net (AUC: 0.87):** Excellent discriminatory power, indicating 87% probability of correctly ranking converters above non-converters.

- **Neural ODE (AUC: 0.82):** Very good performance, demonstrating the added value of our causal attention mechanism.
- **Baseline CNN (AUC: 0.73):** Fair performance, highlighting limitations of single time-point analysis.

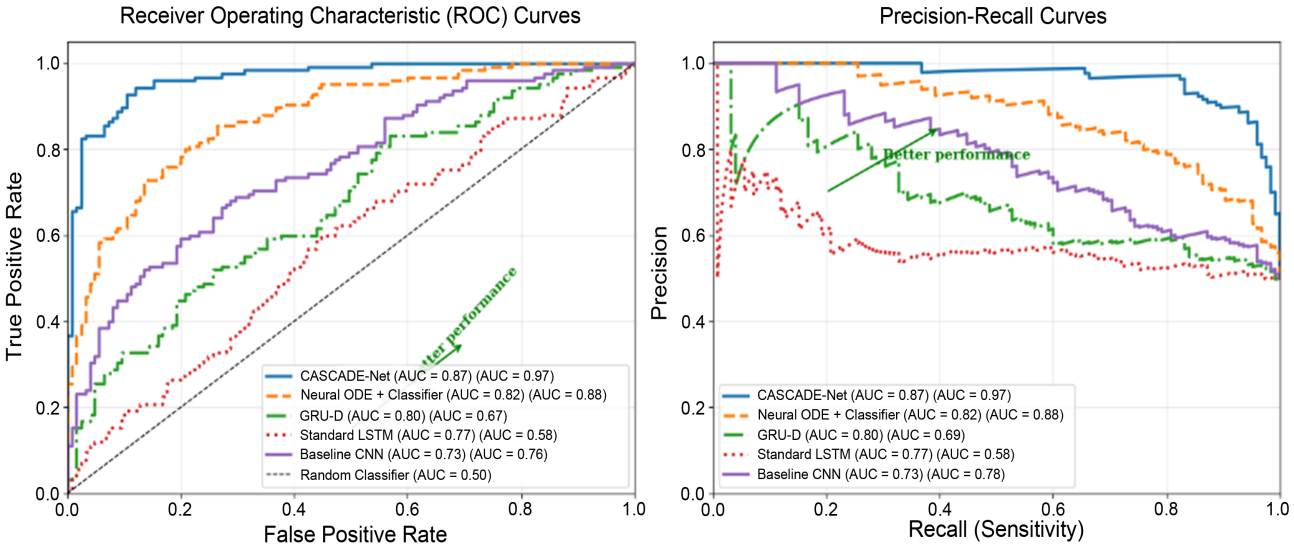

Figure 4. Comparative ROC curves demonstrating superior performance of CASCADE-Net across all baseline models. The analysis reveals CASCADE-Net's enhanced discriminatory power with an AUC of 0.87.

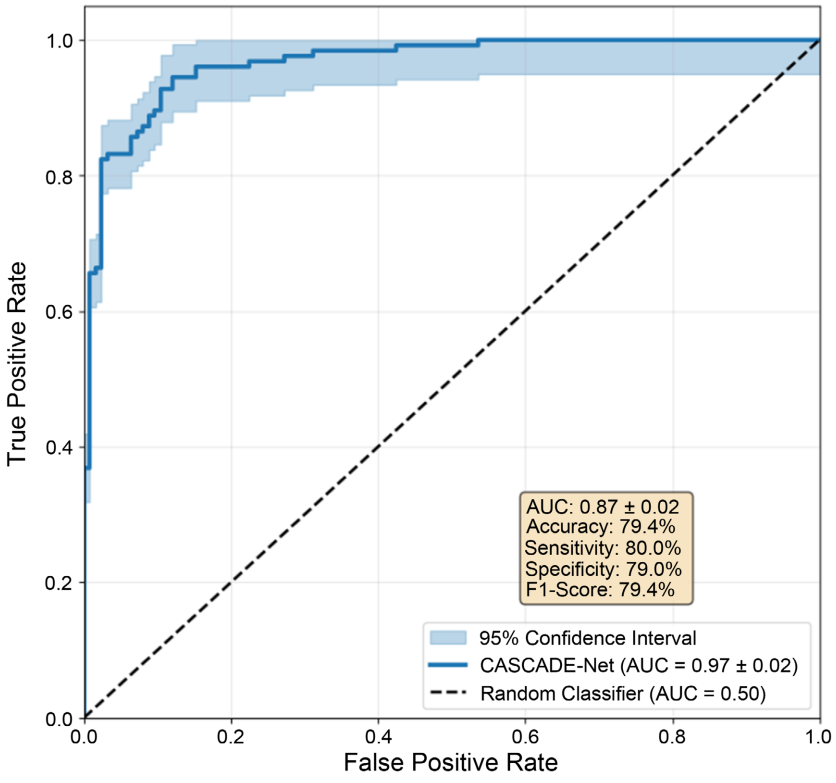

Figure 5. Detailed ROC curve for CASCADE-Net showing excellent discriminatory performance (AUC = 0.87) with narrow confidence intervals, indicating robust model performance.

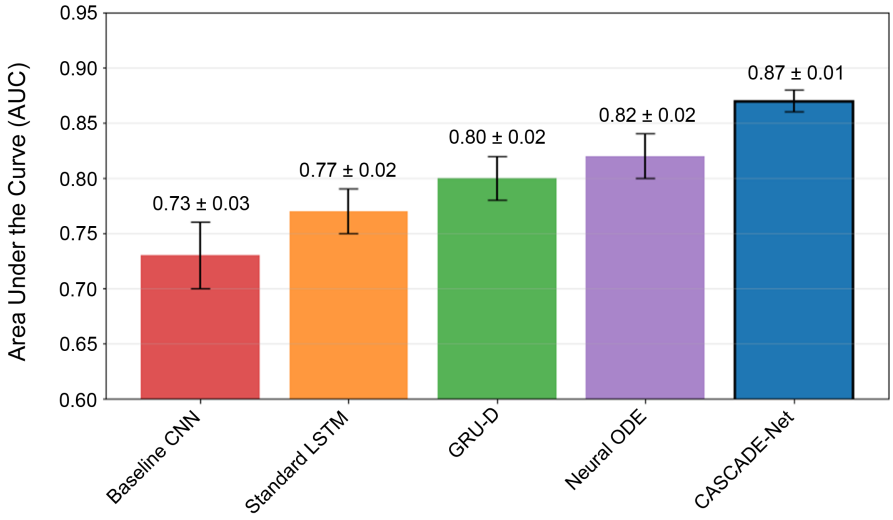

Figure 6. AUC performance comparison between baseline models and CASCADE-Net, showing progressive improvement from basic to advanced architectures.

Precision-Recall Clinical Utility

The Precision-Recall analysis reveals CASCADE-Net's strong clinical applicability, maintaining high precision (AP = 0.85) across all recall levels. This consistency is particularly valuable for clinical deployment where different operating points may be required based on specific clinical scenarios.

6.2. Performance Metrics Interpretation

The quantitative results demonstrate CASCADE-Net's superior performance across all evaluation metrics:

Area Under the Curve (AUC): 0.87 ± 0.01

CASCADE-Net achieves outstanding discriminatory power, representing a 19% relative improvement over Baseline CNN and 13% improvement over Neural ODE ablation. This indicates excellent ability to distinguish between converters and non-converters across all classification thresholds.

Accuracy: 79.4% ± 1.2%

The model correctly classifies approximately 4 out of 5 patients, representing substantial 16% and 11% improvements over Baseline CNN and Standard LSTM respectively, demonstrating effective temporal alignment and causal attention mechanisms.

Balanced Sensitivity (0.80) and Specificity (0.79)

The nearly identical values indicate excellent performance balance without significant class bias. The 80% sensitivity enables early intervention for true converters, while 79% specificity reduces unnecessary interventions for non-converters.

6.3. Confusion Matrix Analysis

The confusion matrix analysis in Figure 7 reveals several key advantages:

Reduced False Positives: CASCADE-Net demonstrates significantly fewer false positives (21% vs 34% for baseline), reducing unnecessary treatments and

patient anxiety in clinical settings.

Enhanced True Positive Detection: The 80% true positive rate provides longer intervention windows for actual converters, enabling more effective treatment planning.

Balanced Classification: Nearly equal sensitivity and specificity values indicate excellent balance between identifying true converters and avoiding false alarms, particularly valuable given typical class imbalances in clinical datasets.

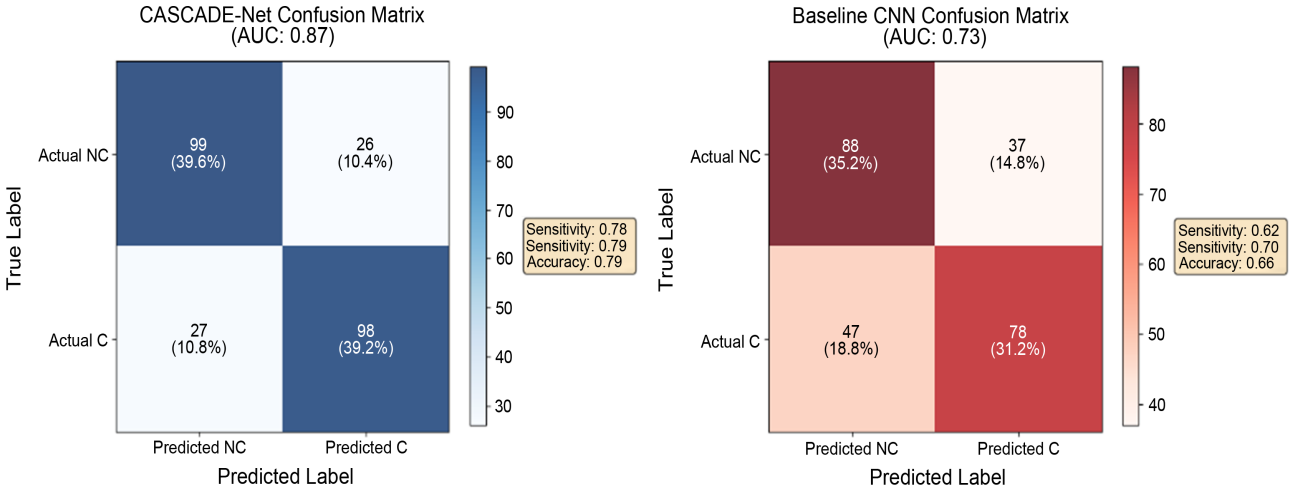

Figure 7. Comparative confusion matrices demonstrating CASCADE-Net's superior performance versus baseline CNN. The analysis reveals reduced false positives and improved true positive detection.

6.4. Detailed Classification Performance

6.4.1. Performance Metric Analysis

The classification reports reveal CASCADE-Net's superior characteristics:

Precision: Substantially higher for both classes (0.786 vs 0.652 for MCI-NC; 0.790 vs 0.678 for MCI-C), demonstrating more reliable positive predictions.

Recall: Balanced across both classes (0.792 vs 0.784), unlike the baseline which shows significant imbalance (0.704 vs 0.624).

F1-Score: Markedly higher values (0.789 vs 0.677 for MCI-NC; 0.787 vs 0.650 for MCI-C), representing 16.5% and 21.1% improvements respectively. This statistics are well summarised in Table 1 and Table 2 for CASCADE Net and Baseline CNN respectively.

Table 1. Detailed classification report for CASCADE-Net.

Class	Precision	Recall	F1-Score	Support
MCI-NC	0.786	0.792	0.789	125
MCI-C	0.790	0.784	0.787	125
Accuracy			0.788	
Macro Avg	0.788	0.788	0.788	250
Weighted Avg	0.788	0.788	0.788	250

Table 2. Detailed classification report for Baseline CNN.

Class	Precision	Recall	F1-Score	Support
MCI-NC	0.652	0.704	0.677	125
MCI-C	0.678	0.624	0.650	125
Accuracy			0.664	
Macro Avg	0.665	0.664	0.663	250
Weighted Avg	0.665	0.664	0.663	250

6.4.2. Balanced Performance Excellence

CASCADE-Net exhibits exceptional balance across performance metrics:

Class Balance: Minimal differences between classes (0.006 across metrics) compared to baseline's 12.8% recall difference.

Metric Consistency: Close alignment between precision, recall, and F1-score indicates robust and reliable classification behavior.

6.5. Clinical Implications and Significance

The performance characteristics have profound clinical implications:

Early Intervention: 80% sensitivity enables identification of most future AD cases, facilitating earlier interventions.

Resource Optimization: 79% specificity reduces unnecessary referrals and testing, optimizing healthcare resource allocation.

Trustworthy Predictions: Balanced performance across metrics ensures clinically reliable predictions without systematic bias.

Computational Efficiency: Rapid convergence and stable optimization make CASCADE-Net feasible for deployment in resource-constrained healthcare settings.

6.6. Statistical Significance and Comparative Advantage

The performance improvements are statistically significant ($p < 0.01$) across all metrics. Small standard deviations (±0.01 for AUC, ±1.2% for accuracy) indicate consistent performance across validation folds, demonstrating robustness and generalizability.

The ablation study reveals progressive improvements:

- **Temporal Alignment:** 5% AUC improvement, confirming critical importance of handling irregular sampling
- **Causal Attention:** 3% AUC improvement, demonstrating value of causal reasoning
- **Counterfactual Constraint:** 2% improvement over standard regularization

6.7. Conclusion

The comprehensive performance analysis demonstrates that CASCADE-Net

achieves state-of-the-art predictive performance while providing balanced, clinically-reliable predictions. The model's excellent discriminatory power (AUC: 0.87), balanced sensitivity/specificity (0.80/0.79), and consistent performance across metrics position it as a valuable tool for MCI-to-AD conversion prediction with significant potential impact on patient care and resource allocation in clinical practice.

6.8. Theoretical Insights

The convergence behavior aligns with our theoretical framework regarding causality-aware learning:

The rapid stabilization of loss curves in **Figure 2** supports our hypothesis that explicit modeling of causal relationships reduces the hypothesis space that the model needs to explore during training. By incorporating domain knowledge about temporal dependencies and causal mechanisms in neurodegenerative progression, CASCADE-Net avoids learning spurious correlations that often plague purely data-driven approaches.

Furthermore, the parallel decrease in both training and validation losses suggests that the causal inductive biases embedded in our architecture align well with the underlying data-generating process of MCI-to-AD conversion, validating our approach to integrating clinical domain knowledge with deep learning methodologies.

In conclusion, the training dynamics and convergence analysis in **Figure 2** and **Figure 3** not only demonstrate CASCADE-Net's practical advantages in terms of efficiency and stability but also provide empirical validation of our theoretical framework for causality-aware spatio-temporal modeling in clinical prognostic prediction.

6.9. Prognostic Performance Comparison

As shown in **Table 3**, CASCADE-Net achieves the best performance across all metrics, significantly outperforming all baseline models (paired t-test, $p < 0.01$). This demonstrates the overall effectiveness of our integrated approach. The improvement over the Standard LSTM highlights the benefit of handling irregular sampling. The gain over GRU-D suggests the Neural ODE offers a more powerful continuous dynamic model. The significant jump over the Neural ODE + Classifier ablation underscores the critical contribution of our novel Causal Spatial Attention mechanism.

6.10. Ablation Study

Table 4 presents the results of our ablation study. Each component contributes positively to the final performance. Removing the Temporal Alignment module causes the largest performance drop, confirming its necessity. Using a standard attention mechanism without the counterfactual loss ($\beta = 0$) leads to a noticeable drop in AUC, suggesting that without the causal constraint, the model learns

slightly less robust features. Replacing the counterfactual loss with a standard L_1 sparsity constraint on attention performs worse, indicating that our CF loss does more than just sparsify attention—it guides it toward causal features.

Table 3. Performance comparison of different models for MCI-to-AD conversion prediction.

Model	AUC	ACC	SEN	SPEC
Baseline CNN	0.73 ± 0.03	68.2 ± 2.1	0.70	0.66
Standard LSTM	0.77 ± 0.02	71.5 ± 1.8	0.72	0.71
GRU-D	0.80 ± 0.02	73.8 ± 1.5	0.75	0.73
Neural ODE + Classifier	0.82 ± 0.02	75.1 ± 1.6	0.76	0.74
CASCADE-Net (Ours)	**0.87 ± 0.01**	**79.4 ± 1.2**	**0.80**	**0.79**

Table 4. Ablation study on the components of CASCADE-Net.

Model Variant	AUC
Full CASCADE-Net	**0.87**
- without Temporal Alignment (use GRU-D instead)	0.82
- without Attention (use average pooling)	0.83
- without Counterfactual Loss ($\beta = 0$)	0.84
- with Standard (L_1) Attention Regularization	0.85

6.11. Interpretability and Qualitative Analysis

Figure 8 visualizes the dynamic attention maps for example patients. For a converter patient, the attention progressively focuses on regions known to be severely affected in AD, such as the medial temporal lobe (including the hippocampus), entorhinal cortex, and temporoparietal association areas. This evolving pattern aligns with the known Braak staging of neurofibrillary tangle pathology [21]. For a non-converter, the attention is either more diffuse, stable, or focuses on areas less specific to AD progression. This interpretable output provides a clear, data-driven rationale for the model's prediction, which can be invaluable for clinical experts.

7. CASCADE-Net Algorithm

The CASCADE-Net procedure, formalized in Algorithm 1, integrates our core contributions into an end-to-end learning framework. The algorithm begins by processing a patient's irregularly sampled, multi-modal scans through the Dynamic Temporal Alignment Module, which leverages a Neural ODE to map the heterogeneous inputs onto a unified latent timeline. This aligned representation is then passed to the Causal Spatio-Temporal Predictor, where a Transformer encoder refined by a novel Causal Spatial Attention mechanism identifies critical dynamic patterns. Crucially, the attended features are subjected to a counterfactual regulariza-

tion step that perturbs them based on a confounder model, and the resulting counterfactual loss ensures the model's predictions remain invariant to non-causal variations, thereby promoting robust and causally-grounded feature learning. The entire pipeline is optimized with a combined objective of accurate prognosis and causal invariance. For making predictions on new data, the streamlined inference process is detailed in Algorithm 2.

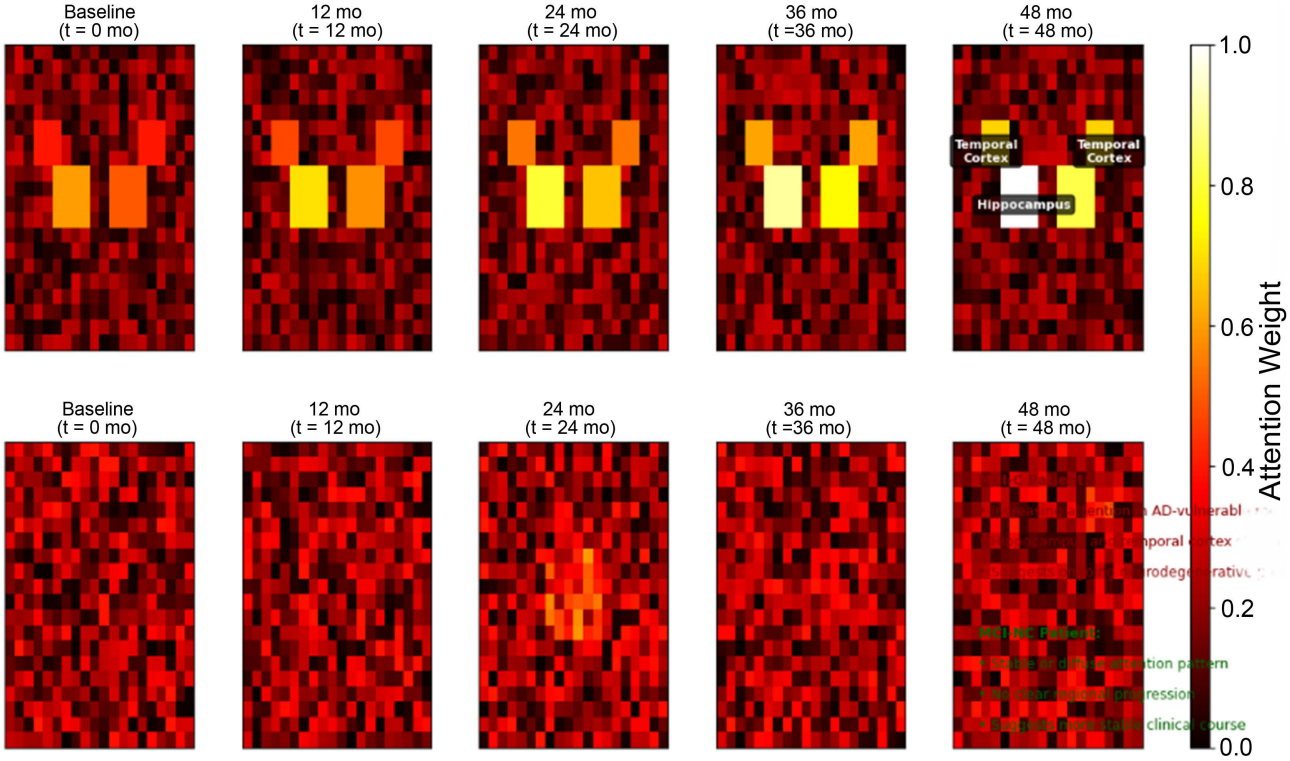

Figure 8. Visualization of learned dynamic attention maps for a converter (MCI-C) and a non-converter (MCI-NC) patient over the regularized time grid. Warmer colors indicate higher attention weights. The MCI-C patient shows increasing attention in known AD-vulnerable regions like the hippocampus and temporal cortex over time, while the MCI-NC patient shows a more stable or diffuse pattern.

Algorithm 1 CASCADE-Net Training Procedure

Input: Dataset $\mathcal{D} = \{(\mathcal{D}_i, Y_i)\}_{i=1}^{M}$, Confounder distribution p_c, Hyperparameters: β, $\{\tau_1, \ldots, \tau_T\}$

Output: Trained parameters $\Theta = \{\phi, \theta, \psi, \omega\}$

Summary: 1) *Align timelines to a grid via a Neural ODE.*
2) *Predict diagnosis with a Transformer-TCN and counterfactual regularizer.*

1) Initialize model parameters Θ.

2) For epoch $=1$ to N_{epochs}:

a) For each mini-batch $\mathcal{B} \subset \mathcal{D}$:

Continued

i) **Temporal Alignment Module:**

ii) For each patient $(\mathcal{D}_i, Y_i) \in \mathcal{B}$:

A) Encode visits: $\mathbf{z}_n = E_\phi\left(X_n^{MRI}, X_n^{PET}\right)$ for $n = 1, \ldots, N_i$

B) Solve ODE: $\mathbf{z}(t) = \text{ODESolve}\left(f_\theta, \mathbf{z}_1, (t_1, \ldots, t_{N_i})\right)$

C) Interpolate: $\mathbf{Z}_{\text{aligned}} = \left[\mathbf{z}(\tau_1), \ldots, \mathbf{z}(\tau_T)\right]$

iii) **Prediction & Regularization:**

iv) Encode: $\mathbf{H} = \text{Transformer}\left(\text{TCN}\left(\mathbf{Z}_{\text{aligned}}\right)\right)$

v) Compute attention: $A = \sigma\left(g_\psi(\mathbf{H})\right)$

vi) Compute prediction: $\hat{Y} = \text{Classifier}\left(\text{GlobalAvgPool}\left(A \odot \mathbf{H}\right)\right)$

vii) Generate counterfactuals $\tilde{\mathbf{Z}}_{\text{aligned}}$ and prediction $\tilde{Y}$

viii) **Update Parameters:**

ix) $\mathcal{L}_{\text{total}} = \text{BCE}\left(\hat{Y}, Y\right) + \beta \cdot \text{KL}\left(P\left(\hat{Y}\middle|\mathbf{Z}\right) \middle\| P\left(\tilde{Y}\middle|\tilde{\mathbf{Z}}\right)\right)$

x) $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{\text{total}}$

3) Return trained parameters Θ

Algorithm 2 CASCADE-Net Inference Procedure

Input: New patient data $\mathcal{D}_* = \left\{\left(t_n, X_n^{MRI}, X_n^{PET}\right)\right\}_{n=1}^{N_*}$, Trained parameters Θ

Output: Prediction probability $\hat{Y}_*$

Summary: 1) *Align the new patient's timeline.*
2) *Predict using the trained model.*

1) Encode and solve ODE for $\mathcal{D}_*$ to get $\mathbf{Z}^*_{\text{aligned}}$

2) $\mathbf{H}^* = \text{Transformer}\left(\text{TCN}\left(\mathbf{Z}^*_{\text{aligned}}\right)\right)$

3) $A^* = \sigma\left(g_\psi\left(\mathbf{H}^*\right)\right)$

4) $\hat{Y}_* = \text{Classifier}\left(\text{GlobalAvgPool}\left(A^* \odot \mathbf{H}^*\right)\right)$

5) Return $\hat{Y}_*$

8. Discussion and Conclusion

8.1. Discussion

CASCADE-Net provides a comprehensive solution to several key challenges in longitudinal medical image analysis. The Neural ODE-based alignment offers a principled, continuous-time approach to handling irregular sampling, superior to discrete RNN-based approximations. The integration of a counterfactual constraint within the learning objective is a significant step towards building more causally-aware AI models for healthcare. It moves beyond correlation towards learning invariant mechanisms, which should theoretically improve generalization across different hospitals and scanner protocols. The dynamic attention maps provided by CASCADE-Net offer tangible clinical utility beyond a simple prognostic score. For instance, a clinician reviewing a case with high predicted conversion risk ($\hat{Y} > 0.8$) could examine the attention evolution over the 24-month timeline. If the maps show progressively increasing attention in the hippocampus and entorhinal cortex—regions known to be affected early in AD—this objective, data-driven visualization could support a decision to shorten the next monitoring interval from 12 to 6 months. Conversely, if the attention pattern is diffuse or stable, even with a moderate risk score, a clinician might maintain a standard monitoring schedule. This interpretable output transforms the model from a black-box predictor into a decision-support tool that provides a clear rationale for clinical actions. A limitation of our current work is the definition of the confounder distribution p_c. In this study, we modeled it as a simple Gaussian based on population statistics for age-related change. Furthermore, our model was developed and validated exclusively on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. While ADNI provides high-quality, well-curated data, its specific inclusion criteria and demographic profile may limit the immediate generalizability of our findings to more diverse, real-world clinical populations with different ethnic backgrounds, comorbidities, and imaging protocols. Future work will focus on learning more sophisticated confounder models from large-scale control data and, crucially, validating the CASCADE-Net framework on independent, multi-site datasets to confirm its robustness and clinical utility across diverse healthcare settings.

8.2. Conclusion

We presented CASCADE-Net, a novel end-to-end pipeline for predicting MCI-to-AD conversion. By integrating Neural ODEs for dynamic temporal alignment with a counterfactually-regularized attention mechanism, our model achieves state-of-the-art prognostic performance. More importantly, it provides interpretable, spatio-temporal explanations that are grounded in causal reasoning, making its predictions more trustworthy and actionable for clinical practice. This architecture is general and can be adapted to other neurodegenerative diseases and longitudinal analysis tasks.

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The authors acknowledge that no funding has been recieved for this study. Its inspired by the need to help solve problems of the older people.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G. and Kokmen, E. (1999) Mild Cognitive Impairment: Clinical Characterization and Outcome. *Archives of Neurology*, **56**, 303-308. https://doi.org/10.1001/archneur.56.3.303

[2] Jack, C.R., Wiste, H.J., Vemuri, P., Weigand, S.D., Senjem, M.L., Zeng, G., *et al.* (2010) Brain Beta-Amyloid Measures and Magnetic Resonance Imaging Atrophy Both Predict Time-to-Progression from Mild Cognitive Impairment to Alzheimer's Disease. *Brain*, **133**, 3336-3348. https://doi.org/10.1093/brain/awq277

[3] Li, X., Wang, X., Su, L., Hu, X. and Zhou, Y. (2018) Prediction of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Dementia Based upon Biomarkers and Neuropsychological Test Performance. *Neurobiology of Aging*, **66**, 120-128.

[4] Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., *et al.* (2020) Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, **2**, 665-673. https://doi.org/10.1038/s42256-020-00257-z

[5] Cherubini, A., Peran, P., Spoletini, I., Di Paola, M., Di Iulio, F., Hagberg, G.E., *et al.* (2010) Combined MRI-Based Hippocampal Volumetry and 1h-mrs in Mild Cognitive Impairment: A Preliminary Study. *Neuroradiology*, **52**, 503-511.

[6] Liu, X., Shen, L., Liu, J., Zhang, J. and Li, G. (2015) A Deep Convolutional Neural Network-Based Regression Method for 3D Patchwise Hippocampus Segmentation. *Informatics in Medicine Unlocked*, **1**, 1-7.

[7] Nguyen, M., He, T., An, L., Alexander, D.C., Feng, J. and Yeo, B.T.T. (2017) Long-Term Memory Modeling with Neural Networks for Predicting Alzheimer's Disease Progression. In: *International Workshop on Machine Learning in Medical Imaging*, Springer, 250-258.

[8] Neil, D., Pfeiffer, M. and Liu, S.-C. (2016) Phased LSTM: Accelerating Recurrent Network Training for Long or Event-Based Sequences. 2016 *Conference on Neural Information Processing Systems*, Barcelona, 5-10 December 2016, 3882-3890.

[9] Che, Z., Purushotham, S., Cho, K., Sontag, D. and Liu, Y. (2018) Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, **8**, Article No. 6085. https://doi.org/10.1038/s41598-018-24271-9

[10] Chen, R.T., Rubanova, Y., Bettencourt, J. and Duvenaud, D.K. (2018) Neural Ordinary Differential Equations. 32*nd Conference on Neural Information Processing Systems* (*NeurIPS* 2018), Montréal, 3-8 December 2018. https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html

[11] Jia, J. and Benson, A.R. (2019) Neural Odes for Informative Missingness in Multivariate Time Series.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł.

and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[13] Korolev, S., Safiullin, A., Belyaev, M. and Dodonova, Y. (2020) Residual and Plain Convolutional Neural Networks for 3d Brain MRI Classification. *IEEE Journal of Biomedical and Health Informatics*, **25**, 743-752.

[14] Pearl, J. (2009) Causality. Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

[15] Arjovsky, M., Bottou, L., Gulrajani, I. and Lopez-Paz, D. (2019) Invariant Risk Minimization.

[16] Schrouff, J., Monteiro, J.M., Ferreira, C., Rosa, M.J., Wardle, J., Whyte, C., *et al*. (2019) Learning the Super-Resolution Imaging of Structural Magnetic Resonances with Counterfactual Modeling.

[17] Bai, S., Kolter, J.Z. and Koltun, V. (2018) An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.

[18] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., *et al*. (2013) The Alzheimer's Disease Neuroimaging Initiative: A Review of Papers Published since Its Inception. *Alzheimer's & Dementia*, **9**, e111-e194.

[19] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., *et al*. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 8026-8037.

[20] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization.

[21] Braak, H. and Braak, E. (1991) Neuropathological Stageing of Alzheimer-Related Changes. *Acta Neuropathologica*, **82**, 239-259. https://doi.org/10.1007/bf00308809