ADxPro-T: A Bootstrapped LSTM-Transformer Pipeline for Baseline-to-Long-Term Alzheimer's Prediction

Dhivyaa S P[1], Hyung-Jeong Yang*[,2], Jahae Kim[3], Myungeun Lee[4]
[1]Master's Student, Department of AI Convergence, Chonnam National University
[2,4]Professor, Department of AI Convergence, Chonnam National University
[3]Professor, Department of Nuclear Medicine, Chonnam National University Hospital
*Corresponding Author, Email: hjyang@jnu.ac.kr

ABSTRACT

Accurately modeling the progression of Alzheimer's disease (AD) is critical for timely intervention and treatment planning. However, longitudinal clinical data used for such prediction often suffer from irregular sampling and missing values, limiting the effectiveness of traditional machine learning approaches. In this study, we propose ADxPro-T, a novel end-to-end framework that combines a mask-aware LSTM for imputing and prediction of next timepoint data along with a Transformer encoder–decoder for long-range prediction from a single baseline visit. ADxPro-T jointly performs diagnostic prediction and biomarker imputation, leveraging temporal patterns even in the presence of sparse data. Evaluated on the ADNI dataset, our model outperforms state-of-the-art baselines including GRU-D, MinimalRNN, and BiPro, in both classification and imputation tasks. Further, an ablation study confirms that incorporating baseline diagnosis significantly enhances predictive performance. These results establish ADxPro-T as a robust and effective model for early-stage AD progression prediction in real-world clinical settings.

Keywords: Alzheimer's Disease Progression, Longitudinal Modelling, Missing data

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline, memory loss, and behavioral changes. It is one of the most common causes of dementia among older adults, affecting millions worldwide. As the global population ages, the prevalence of Alzheimer's disease is expected to rise significantly, posing a substantial burden on healthcare systems and society at large. Early detection and accurate prediction of disease progression are crucial for effective treatment and intervention strategies, potentially slowing the progression of symptoms and improving patients' quality of life.

However, predicting the progression of Alzheimer's disease remains a challenging task due to the complex nature of the disease and the heterogeneity of patient data. Clinical data used for prediction, such as biomarkers obtained from neuroimaging data, genetic information, and cognitive assessments, often suffer from significant missing information. Missing data can arise from various factors, including inconsistent follow-up visits, high costs of diagnostic procedures, and patient dropout, leading to incomplete records that hinder the development of robust predictive models.

Early models for predicting AD progression primarily relied on linear statistical methods, such as linear regression and the Cox proportional hazards model, which were used to analyze relationships between baseline characteristics and disease outcomes [1]. While these models provided foundational insights, they were limited in their ability to capture the non-linear and dynamic nature of disease progression.

As the complexity of AD pathology became increasingly apparent, a shift occurred towards more sophisticated modelling approaches. This shift was driven by the need to handle the complex patterns within high-dimensional data that traditional statistical models could not adequately address [2]. Primary machine learning models, such as support vector machines (SVMs) and random forests, were applied to predict AD progression. However, these models faced significant challenges when dealing with the temporal aspects of the disease, as they were not designed to handle sequential data [3].

The advent of deep learning introduced powerful tools such as recurrent neural networks (RNNs) and their variants, including long short-term memory (LSTM) and gated recurrent unit (GRU) networks, which have proven effective in modelling AD progression by capturing temporal dependencies in longitudinal data [4]. These models learn from sequences of patient data, such as cognitive test scores, neuroimaging measurements, and genetic information, offering a more deep and detailed understanding of disease progression [5].

Despite these advancements, the presence of missing data in clinical records remains a significant challenge. Missing data, often due to irregular follow-up visits, patient dropout, or high diagnostic costs, can severely compromise model

performance [6]. Various models have been developed to address this issue. [7] proposed a novel approach called Gated Recurrent Unit with Decay (GRU-D), which uses two representations of missing pattern: masking and time interval, to capture long-term temporal correlations in time series data. GRU-D's decay mechanism adjusts the influence of past observations based on the elapsed time, effectively handling irregular intervals and informative missingness patterns. MinimalRNN, introduced by [8], is a lightweight recurrent neural network aimed at predicting patient diagnoses, ventricular volumes, and cognitive scores in AD. BiPro, proposed by [9], employs a bidirectional RNN model that integrates both past and future data for imputation and prediction, leveraging temporal information in both directions.

Recently, transformer architectures that rely on self-attention have begun to transform sequence modeling in natural language processing and computer vision by capturing long-range dependencies. However, to our knowledge no study has applied a purely Transformer-based model to predict Alzheimer's disease progression from a single baseline visit, since transformers generally require a substantial number of input tokens to learn effective attention patterns.

To bridge this gap, we introduce ADxPro-T, an end-to-end framework for predicting Alzheimer's disease progression from a single baseline visit. Our key contribution is a unified architecture in which an LSTM module generates intermediate visits directly from baseline data and feeds them into a Transformer encoder, with all components trained jointly. This design ensures that the LSTM's synthetic visit generation is optimized for the downstream attention-based prediction task.

2. Proposed Method

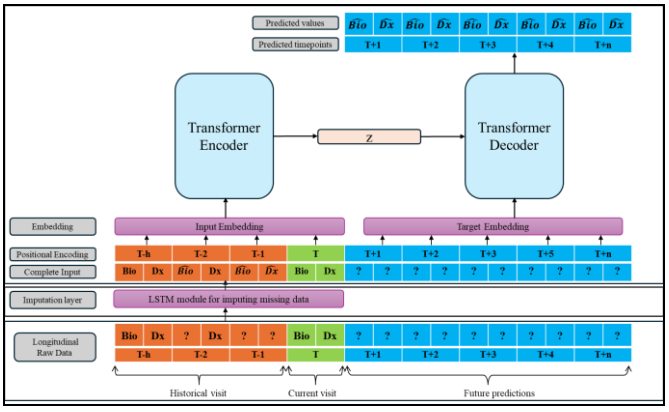

Figure 1. Overall Architecture of Proposed Model

Figure 1 illustrates the overall architecture of ADxPro-T, our end-to-end LSTM + Transformer model for Alzheimer's disease progression prediction from a single baseline visit. The model comprises three main components: (1) an LSTM-based imputation module that bootstraps missing and future visits, (2) embedding and positional encoding, and (3) a Transformer encoder–decoder that autoregressively forecasts long-horizon biomarkers and diagnostic labels.

2.1 Problem Statement

Let a multivariate time-series X with T known timepoints with *h* historical visits can be represented as

$$X = \{x_t \in R^d \mid t = T-h, T-h+1, \ldots, T\} \quad (1)$$

where some entries of x_t are missing (denoted by "?" in Figure 1). Our goal is to (a) impute missing values in the history *h* and (b) predict the next *n* time points.

$$\hat{X} = \{\widehat{x_{T+1}}, \widehat{x_{T+2}}, \ldots, \widehat{x_{T+n}}\} \quad (2)$$

2.2 LSTM Module for Initial Imputation & Prediction

At a given times step t, the input is represented as a multi-dimensional array x_t which consists of biomarker x_t^{bio} and diagnosis x_t^{dx}. Therefore, the input array at time step t can be expressed as $x_t = [x_t^{bio}, x_t^{dx}]$. To represent the observation status of elements at the t^{th} visit, we introduce a binary masking vector $m_t \in \{0,1\}$. So, the mask at time point *t* for the input x_t is represented as $m_t = [m_t^{bio} + m_t^{dx}]$. We impute the initial timepoint using the global mean and impute each subsequent timepoint with the predictions $\widehat{x_t}$ from the previous visit. Hence the imputed data x_t^{impute} can be denoted as

$$x_t^{impute} = m_t \odot x_t + (1 - m_t) \odot \widehat{x_{t-1}} \quad (3)$$

We employ a mask-aware LSTM to capture temporal dependencies in the time-series, then use its hidden state h_t to predict both continuous biomarker values and discrete diagnoses at *t+1*. Biomarker predictions are obtained by linearly combining h_t with the input and diagnosis probabilities are produced by applying a SoftMax classifier directly to h_t. They are defined as

$$\widehat{x_{t+1}^{bio}} = W_{bio}\, h_t + x_t^{impute} \quad (4)$$

$$\widehat{x_{t+1}^{dx}} = Softmax\,(W_{dx}\, h_t) \quad (5)$$

Thus, the imputed input sequence is represented as below

$$\widehat{X_{t=T-h}^{T}} = \{\widehat{x_{T-h}}, \ldots, \widehat{x_{T-1}}, \widehat{x_T}\} \quad (6)$$

which serves as an input the transform for long range prediction.

2.3 Transform Module for Long-range Prediction

To capture long-range temporal dependencies across both real and LSTM-generated visits, we employ a standard Transformer encoder-decoder configuration as in [10]. To enable the Transformer to distinguish temporal order in irregularly sampled Alzheimer's visits, we augment each imputed feature vector with a learned positional embedding indexed by its relative year. Let s_t be the integer position index for visit *t*, defined by

$$s_t = \tilde{t} - (T - h) + 1 \quad (7)$$

where $s_t \in \{1, 2, \ldots, h+1\}$ for $\tilde{t}$ = T–h,…,T–1,T

timepoints. We introduce an embedding matrix P,

$$P = [p_i, p_2, \dots, p_{\{h+1\}}]^T \quad (8)$$

where $P \in \mathbb{R}^{\{(h+1)\times d_{model}\}}$ with each row $p_i \in \mathbb{R}^{d_{model}}$ is a trainable vector. This enables the model to learn how to optimally adjust its feature representations for each individual timepoint. The input embedding for each timepoint is given by

$$e_t = W_e \cdot \hat{X}_T + p_{\{s_t\}} \; ; \; W_e \in \mathbb{R}^{\{d_{model}\times d\}} \quad (9)$$

so that,

$$E = [e_{\{T-h\}}, e_{\{T-h+1\}}, \dots, e_T] \in \mathbb{R}^{\{(h+1)\times d_{model}\}} \quad (10)$$

becomes input to the transformer encoder.

In the decoder, we continue the same scheme for prediction steps *k = 1, 2, …, n* by defining

$$s_{\{T+k\}} = (h+1) + k \quad (11)$$

$$u_{\{T+k\}} = W_e \cdot \hat{x}_{\{T+k-1\}} + p_{\{s_{\{T+k\}}\}} \quad (12)$$

where $\hat{x}_{\{T+k-1\}}$ is previously predicted values. Hence the transformer decoder embedding U is formulated as

$$U = [u_{\{T+1\}}, u_{\{T+2\}}, \dots, u_{\{T+n\}}] \in \mathbb{R}^{\{n \times d_{model}\}} \quad (13)$$

Table 1. Transformer Hyperparameters

Parameter	Symbol	Values
Hidden dimension	d_{model}	512
Attention heads	n_heads	8
Encoder layers	E	4
Decoder layers	D	4

Our model jointly performs AD diagnosis progression and biomarker imputation as well as prediction. We optimize a total loss that combines masked cross-entropy for diagnostic predictions with masked mean-squared error between imputed and observed biomarker values. Hence, the loss functions are defined as follows:

$$L_{dx} = -\sum_{t=2}^{T}(x_t^{dx} \log \hat{x}_t^{dx}) \odot m_t^{dx} \quad (14)$$

$$L_{bio} = \sum_{t=2}^{T} | \hat{x}_t^{bio} - x_t^{bio} | \odot m_t^{bio} \quad (15)$$

$$L_{total} = L_{dx} + L_{bio} \quad (16)$$

3. Experimental Setup & Result Analysis

3.1. Dataset and Pre-processing

In this study, we utilised the ADNI ADNIMERGE dataset [11], which comprises six regional volumetric biomarkers, demographic variables (age, education, gender, APOE4 genotype) and diagnostic labels collected over a ten-year period. Participants with reverse diagnoses were excluded, and only those with a minimum of three visits were retained to ensure adequate longitudinal coverage. The dataset comprises 1,369 patients, with 739 males and 630 females. The average patient age is 73.76 ± 6.96 years, and the mean education level is 16.01 ± 2.82 years. Among them, 535 are cognitively normal (CN) and 650 have mild cognitive impairment (MCI), with no patients labelled as Alzheimer's disease (AD) at baseline. Continuous measures (biomarkers, age, education) were standardised via z-score normalisation, categorical variables (gender, APOE4 status) were encoded using one-hot vectors, and subject age was incrementally updated at each follow-up to reflect elapsed time. Figure 2 shows the decline of data as the time increases.

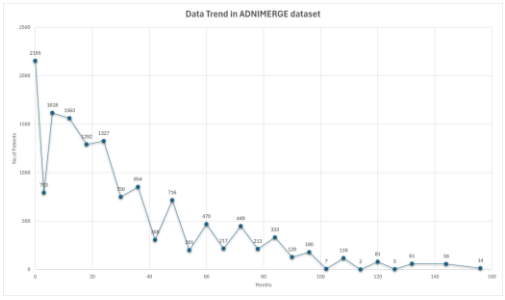

Figure 2. Longitudinal data trend in ADNI MERGE Dataset

3.2. Result Analysis

In this study, the ADxPro-T model was developed and evaluated on a computational platform featuring an NVIDIA A100 Tensor Core GPU with Multi-Instance GPU support and 7 GB of dedicated memory. The implementation relied on the PyTorch framework, with network parameters trained using the Adam optimizer at a learning rate of 0.0001. To promote stable convergence, all model weights were initialised via the Xavier uniform method, and a learning-rate 0.0001 was applied over the course of training. Model generalisability was assessed through five-fold cross-validation, ensuring robust performance estimates. For the Alzheimer's diagnosis task, standard evaluation metrics such as accuracy, precision, recall and mean area under the curve were employed, whereas imputation and prediction quality were quantified using mean absolute error. All competing models were subjected to identical hardware, software and training configurations to guarantee a fair and rigorous comparison.

Table 2. Performance Comparison among Competing Methods

Model	ACC (↑)	PRE (↑)	REC (↑)	mAUC (↑)
GRU-D	0.5750 ± 0.0122	0.5722 ± 0.0069	0.5934 ± 0.0128	0.7632 ± 0.0095
MinimalRNN	0.5680 ± 0.0203	0.5661 ± 0.0211	0.5885 ± 0.0185	0.7549 ± 0.0091
BiPro	0.5847 ± 0.0166	0.5867 ± 0.0219	0.5876 ± 0.0113	0.7650 ± 0.0087
ADxPro-T	**0.6134 ± 0.0235**	**0.6033 ± 0.0219**	**0.6068 ± 0.0194**	**0.7734 ± 0.0099**

Table 2 presents a performance comparison among four competing models—GRU-D, MinimalRNN, BiPro, and ADxPro-T—across four classification metrics: Accuracy (ACC), Precision (PRE), Recall (REC), and macro-AUC (mAUC). The results are reported as mean ± standard deviation, averaged across cross-validation folds. Among all models, ADxPro-T consistently achieves the highest performance across all metrics. BiPro ranks second, performing better than GRU-D and MinimalRNN, particularly in accuracy and mAUC. Figure 3 illustrates the biomarker imputation performance of four models - GRU-D, MinimalRNN, BiPro, and ADxPro-T, in terms of Mean Squared Error (MSE). The horizontal bar chart clearly shows that ADxPro-T achieves the lowest MSE, indicating the most accurate reconstruction of missing biomarker values. This suggests that ADxPro-T is more effective at capturing

complex temporal patterns and structural dependencies in longitudinal medical data. In contrast, GRU-D exhibits the highest MSE, reflecting relatively lower imputation accuracy. The results emphasize the effectiveness of transformer-based over traditional RNN-based approaches for medical time series imputation tasks.

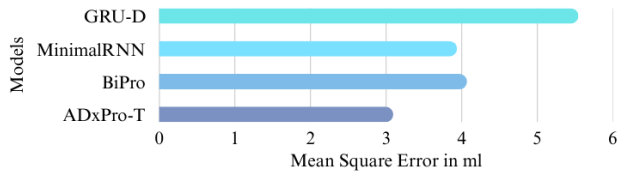

Figure 3. Comparison of mean square error (MSE) in biomarker imputation across different models.

An ablation study as shown in Figure 4, assess the effect of including baseline diagnosis (Dx) on model performance. Incorporating Dx leads to consistent improvements across all metrics highlighting its value in enhancing predictive accuracy.

Figure 4. Ablation Study based on incorporating Dx in baseline visit

4. Conclusion

In this work, we proposed ADxPro-T, a novel hybrid framework that integrates LSTM-based synthetic visit generation with Transformer-based long-range modelling to predict Alzheimer's disease progression from a single baseline visit. Our model effectively handles missing data and captures complex temporal dependencies, resulting in improved diagnostic classification and biomarker imputation. Extensive experiments on the ADNI dataset demonstrate that ADxPro-T outperforms state-of-the-art baselines such as GRU-D, MinimalRNN, and BiPro in both classification accuracy and imputation precision. Additionally, an ablation study confirms the value of incorporating baseline diagnosis in enhancing predictive performance. These findings highlight the potential of combining sequential generation with attention-based models for robust and early Alzheimer's progression modelling. In future, the study will be extended to neuro-imaging modalities.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00208397), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629), and the Information Technology Research Center (ITRC) support program (IITP-2025-RS-2024-00437718) supervised by the IITP.

Reference

[1] Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., ... & Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology, 9(1), 119-128. doi:10.1016/S1474-4422(09)70299-6

[2] Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage, 104, 398-412. doi:10.1016/j.neuroimage.2014.10.002

[3] Beckett, L. A., Harvey, D., Farias, S. T., & Mungas, D. (2012). Multidimensional characterization of trajectories in cognitive and brain aging: A harmonized approach. The Journals of Gerontology: Series B, 67(3), 271-281. doi:10.1093/geronb/gbr161

[4] Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2016). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.

[5] Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Paulsen, R. R., & Nielsen, M. (2016). Early detection of Alzheimer's disease using MRI hippocampal texture. Human Brain Mapping, 37(3), 1148-1161. doi:10.1002/hbm.23091

[6] Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., ... & Feng, D. (2014). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Transactions on Biomedical Engineering, 62(4), 1132-1140. doi:10.1109/TBME.2014.2372011

[7] Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Scientific Reports 8(1), 6085 (2018)

[8] Nguyen, M., He, T., An, L., Alexander, D.C., Feng, J., Yeo, B.T., Alzheimer's Disease Neuroimaging Initiative: Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage 222, 117203 (2020)

[9] Ho, N.H., Yang, H.J., Kim, J., Dao, D.P., Park, H.R., Pant, S.: Predicting progression of Alzheimer's disease using forward-to-backward bi-directional network with integrative imputation. Neural Networks 150, 422-439 (2022).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[11] Alzheimer's Disease Neuroimaging Initiative (ADNI). Available at: http://adni.loni.usc.edu. Last accessed 2025/04/24