

Transformer Model for Alzheimer’s Disease Progression Prediction Using Longitudinal Visit Sequences

Mahdi Moghaddami
Clayton Schubring
Mohammad-Reza Siadat
Oakland University, USA

MOGHADDAMI@OAKLAND.EDU
C.SCHUBRING@OAKLAND.EDU
SIADAT@OAKLAND.EDU

Abstract

Alzheimer’s disease (AD) is a neurodegenerative disorder with no known cure that affects tens of millions of people worldwide. Early detection of AD is critical for timely intervention to halt or slow the progression of the disease. In this study, we propose a Transformer model for predicting the stage of AD progression at a subject’s next clinical visit using features from a sequence of visits extracted from the subject’s visit history. We also rigorously compare our model to recurrent neural networks (RNNs) such as long short-term memory (LSTM), gated recurrent unit (GRU), and minimalRNN and assess their performances based on factors such as the length of prior visits and data imbalance. We test the importance of different feature categories and visit history, as well as compare the model to a newer Transformer-based model optimized for time series. Our model demonstrates strong predictive performance despite missing visits and missing features in available visits, particularly in identifying converter subjects—individuals transitioning to more severe disease stages—an area that has posed significant challenges in longitudinal prediction. The results highlight the model’s potential in enhancing early diagnosis and patient outcomes.

Data and Code Availability We use data provided by the Alzheimer’s disease prediction of longitudinal evolution (TADPOLE) challenge (Marinescu et al., 2018) which is derived from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (Jack et al., 2008) and is available publicly at <https://ida.loni.usc.edu/>. Our code is also available and included as supplemental material.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Alzheimer’s disease (AD) is the most prevalent cause of dementia (Scheltens et al., 2016). Nearly 7 million Americans age 65 and older are living with AD, and it is projected that by 2050, this number may grow to a projected 12.7 million as the size of the U.S. population age 65 and older continues to grow. AD was the fifth-leading cause of death among individuals aged 65 and older in 2021 in the United States (202, 2024). Even before an individual with AD dies, their quality of life and their family’s deteriorate substantially. The significant growth rate of AD, combined with its fatality, makes AD a vital public safety matter.

Since no cure has been found for AD yet, it must be detected as early as possible so that medical professionals can offer effective treatment measures and lifestyle changes to reduce the speed of its progression (Livingston et al., 2020). However, most diagnoses of Alzheimer’s disease happen in the moderate to late stages, indicating that the optimal time for treatment has often already elapsed (Hoang et al., 2023).

Mild cognitive impairment (MCI) is a condition that exists between normal cognitive function and Alzheimer’s disease (AD). It involves a slight decline in cognitive abilities and memory, such that the patients or their nearest family members notice some symptoms. However, individuals at this stage do not reach the level of dementia. MCI is considered a precursor to AD since 10-15% of individuals with MCI will develop AD every year (Sarakhshi et al., 2022).

It is crucial to develop methods that can predict the onset of AD by forecasting the trajectory of disease progression and predicting when cognitively normal (CN) individuals will develop MCI, and when individuals with MCI will develop AD.

With recent advances in machine learning and deep learning, researchers have increasingly used computer-aided diagnosis (CAD) systems, primarily

various deep learning architectures, to detect AD. Most previous work has focused on classifying subjects using a variety of data modalities into clinical diagnoses (CN, MCI, and AD) (Alp et al., 2024; Ji et al., 2019; Ahmed et al., 2019; Bi et al., 2019).

Some studies categorize MCI subjects into two classes: stable MCI (sMCI) and progressive MCI (pMCI). Subjects that do not convert to AD in a period of time (usually three years) are labeled sMCI, and subjects that do are labeled pMCI (Li et al., 2012; Cui and Liu, 2019; Spasov et al., 2018). The issue with this approach is that the conversion window is arbitrary, and there is no difference between subjects who convert relatively early in the given time frame and subjects that convert much later.

Using multiple time points for each subject would be preferable, as the model would have the information to understand the subject’s progression trajectory comprehensively. Zhang and Shen (2012) aims to forecast future clinical changes in patients with MCI accurately. The authors utilize a multi-kernel support vector machine (SVM) to predict future cognitive scores at a 24-month follow-up and whether a patient will convert from MCI to AD, using data from time points at least 6 months before the actual conversion. Cui et al. (2019) presents a combination of a convolutional neural network (CNN) and an RNN to distinguish AD patients from normal controls and classify sMCI patients vs. pMCI patients. Nguyen et al. (2019) used longitudinal data from the TADPOLE challenge and a minimalRNN model (Chen, 2018) to predict clinical diagnosis and other biomarkers for each subject up to 6 years into the future.

Vaswani et al. (2017) introduced the Transformer architecture in 2017. The authors used a mechanism called attention to transfer context between the encoder and decoder layers in the model. Transformer-based models have been employed in CAD systems more frequently after yielding superior results in natural language processing. Chen and Hong (2023) uses a Transformer-based model to classify CN vs. AD using magnetic resonance imaging (MRI) data from multiple visits. Alp et al. (2024) uses a vision Transformer (ViT) to extract spatial features from T1-weighted brain MRIs. Sequence modeling is performed on the feature sequences to maintain interdependencies, and finally, features are classified using a Time Series Transformer (TST) (Zerveas et al., 2021).

The TST is designed for time series forecasting and representation learning. It’s an encoder-only Transformer incorporating time-series-specific inductive bi-

ases, such as causal attention and positional encoding tailored for temporal data.

Our work differs from prior work in two main ways. Most previous works either do not incorporate the longitudinal history of subjects or do not predict the clinical diagnosis in the near future. Also, to the best of our knowledge, no study focuses on assessing the significance of the length of visit history of subjects utilizing Transformers and RNN-based models.

In this work, we propose a Transformer model that utilizes a subject’s entire visit history to predict the clinical diagnosis of the subject at the next visit. We use an imputation strategy to handle missing data and avoid discarding visits with missing feature values. We focus on the conversion points of subjects from CN to MCI and MCI to AD since they are vital for early detection. Statistical testing is performed to compare the performance of the models and to show that the proposed Transformer model significantly outperforms the RNN-based models in predicting the conversion points. Finally, we analyze feature importance and the importance of visit history length.

2. Materials and Methods

2.1. Data

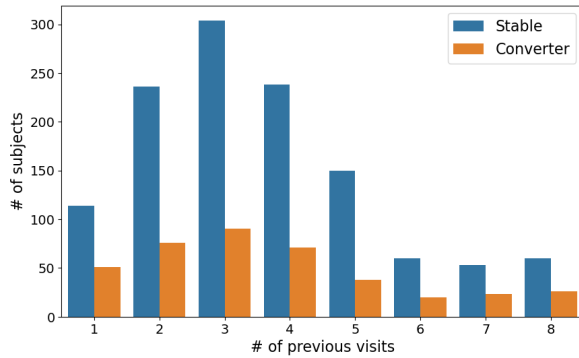
We use data from the Alzheimer’s disease prediction of longitudinal evolution (TADPOLE) challenge ¹, which contains longitudinal data from 1677 participants (Marinescu et al., 2018) from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005). Multiple data points were collected for each participant at a minimum interval of 6 months, making it a valuable dataset for longitudinal studies. The TADPOLE challenge aimed to develop models to accurately predict future clinical states and biomarker trajectories in individuals at risk for or diagnosed with AD.

We use a subset of 23 features from the TADPOLE dataset, the same ones recommended by the TADPOLE challenge. These features include information from different data modalities, such as neuropsychological test scores, clinical diagnosis (CN, MCI, or AD), MRI measures, positron emission tomography (PET) measures, and cerebrospinal fluid (CSF) markers. Table 1 shows a comprehensive list of the features used in this study.

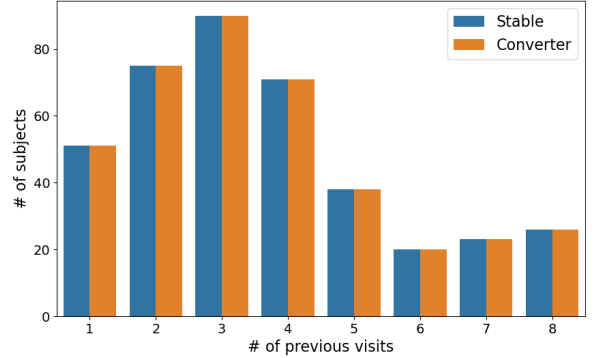
1. <https://tadpole.grand-challenge.org/>

Table 1: List of features with their means, standard deviations, and the percentage of missing values. SB: Sum of boxes, ADAS: Alzheimer’s Disease Assessment Scale, RAVLT: Rey Auditory Verbal Learning Test.

Category	Feature	Mean \pm std	Missingness %
-	Diagnosis	-	30.11%
Cognitive	Functional Activities Questionnaire (FAQ)	4.65 ± 6.96	29.40%
	Clinical Dementia Rating Scale (SB)	1.83 ± 2.29	29.64%
	Mini-Mental State Examination (MMSE)	26.92 ± 3.5	29.88%
	ADAS-Cog11	10.74 ± 7.76	30.05%
	RAVLT immediate	35.08 ± 13.24	30.67%
	RAVLT learning	4.14 ± 2.81	30.67%
	ADAS-Cog13	16.62 ± 10.68	30.72%
	RAVLT forgetting	4.26 ± 2.55	30.88%
	RAVLT forgetting percent	58.73 ± 37.57	31.43%
	Montreal Cognitive Assessment (MOCA)	23.52 ± 4.18	61.01%
MRI	Intracranial volume	1534699.07 ± 164732.93	37.57%
	Whole brain volume	1010781.21 ± 111280.94	39.65%
	Ventricles volume	42119.98 ± 23274.12	41.56%
	Hippocampus volume	6684.54 ± 1224.13	46.61%
	Entorhinal cortical volume	3455.9 ± 801.46	49.22%
	Fusiform cortical volume	17117.41 ± 2798.63	49.22%
	Middle temporal cortical volume	19206.76 ± 3098.07	49.22%
Biomarker	Fluorodeoxyglucose (FDG) - PET	1.21 ± 0.16	73.69%
	Phosphorylated tau	27.59 ± 11.7	81.38%
	Beta-amyloid (CSF)	1052.48 ± 502.57	81.40%
	Total tau	288.67 ± 105.95	81.45%
	Florbetapir (18F-AV-45) - PET	1.19 ± 0.22	83.38%



(a) Raw dataset



(b) Balanced dataset

Figure 1: Distribution of visit sequences by group number

2.2. Preprocessing

Like many long-term studies, the TADPOLE dataset suffers from some problems: follow-up appointments not occurring, irregular scheduling of visits, and some subjects dropping out before the completion of the study. Also, some measurements are not taken at every visit due to the procedure’s invasiveness. These issues, combined with unintentional mistakes during data collection or scanning cause missing (See Table 1) and incorrect values in the dataset.

To address these points, we perform the following preprocessing steps:

1. Drop all visits belonging to subjects with only one visit.
2. Drop all visits belonging to *reverter* subjects. These are subjects who convert from AD to MCI or from MCI to CN at any point in their visit history.
3. There are subjects who convert from CN to MCI or from MCI to AD multiple times. We identify these subjects and discard all their visits dated after the first conversion.
4. Drop all visits with a missing value for clinical diagnosis.

To handle missing data, we use an imputation method called “model filling” proposed by Nguyen et al. (2019). This method uses predictions from an RNN model to fill in missing values in the dataset. We use predictions from a minimalRNN model.

Since there are irregular time gaps between visits, we introduce an additional feature to encode temporal information into every visit in a subject’s visit history. This feature represents the number of months until the final visit.

2.3. Sequence Generation and Dataset Construction

To better focus on the key points of conversion in participants’ visit history, we construct datasets using stable and converter sequences. We define these terms as follows:

Definition 1 (Converter Sequence) *A sequence of visits for a subject sorted by examination date where at the last visit in the sequence, the subject converts to the next stage of the disease (CN to MCI, CN to AD, or MCI to AD).*

Definition 2 (Stable Sequence) *A sequence of visits for a subject sorted by examination date where at the last visit in the sequence, the subject’s diagnosis remains the same as the penultimate visit (CN to CN, MCI to MCI, or AD to AD).*

We generate one converter sequence of maximal length for every subject who at any point in their visit history, converted from CN to MCI or from MCI to AD. Similarly, we generate one stable sequence of maximal length from every other subject. The set of all the stable and converter sequences we obtain by this process is the *raw* dataset. The raw dataset is highly imbalanced, containing 1222 stable sequences and only 405 converter sequences, of which only 70 contain a CN to MCI conversion.

Definition 3 (Group n) *All stable or converter sequences with $n + 1$ visits.*

Note that every sequence only belongs to one group and groups cannot share any sequences.

We also construct one additional dataset: the *balanced* dataset. This dataset contains an equal number of stable and converter sequences in each group. It is created in order to ensure that the models learn equally from all classes, preventing bias towards the majority class. This dataset is built by randomly discarding stable sequences in each group until the number of stable and converter sequences in each group is equal.

Figure 1 shows the distribution of visit sequences by group number in the raw and balanced datasets. Figure 2 shows the distribution of the diagnosis (DX) in the target visit by group number in the raw dataset.

2.4. Model

We consider a multi-class classification (CN, MCI, AD) problem with the input being the first n visits from a visit sequence $v = (v_1, v_2, \dots, v_n)$ and the target being the diagnosis at visit $n + 1$ (d_{n+1}).

We utilized three RNN-based models as a baseline for predicting disease progression: long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), gated recurrent unit (GRU) (Cho et al., 2014), and minimalRNN (minRNN) (Chen, 2018). RNN-based models try to preserve the context of previous time points by carrying a hidden state vector h . At time point t , predictions for the next time point (\hat{y}_t) are made using the current observation x_t and the

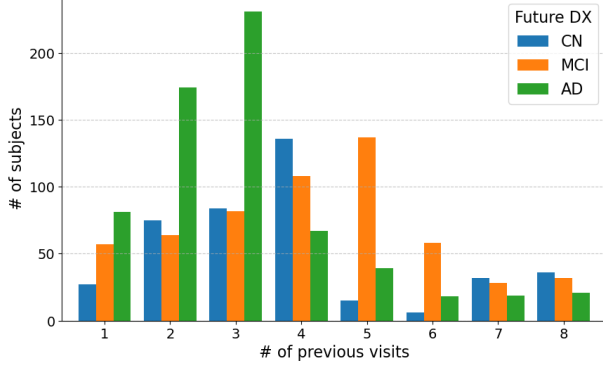


Figure 2: Distribution of Future DX across groups

hidden state from the prior time point h_{t-1} . The equations for updating the hidden state and calculating the output vary between LSTM, GRU, and min-RNN.

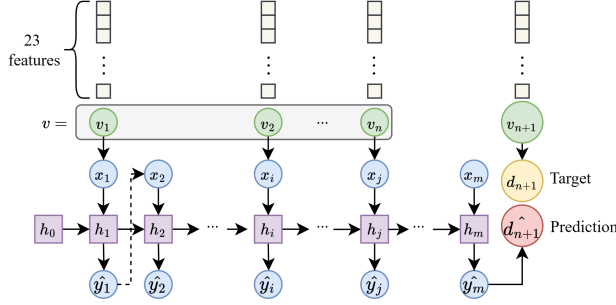


Figure 3: Prediction process for RNN-based models

Following [Nguyen et al. \(2019\)](#), we train these models by predicting clinical diagnosis for every month into the future up to six years, starting with only the first visit in the sequence v_1 . If the actual feature values for timepoint t exist in the visit sequence v , a loss value is calculated between y_{t-1} and x_t and weights are updated accordingly; otherwise, y_{t-1} will be used as the input (x_t) for the next month. Figure 3 shows an overview of the training and prediction process for RNN-based models. They continuously make predictions for the next month until the end of the visit sequence is reached. At this point, the model’s prediction for the diagnosis in the last visit is compared to the actual diagnosis.

Our proposed model is based on the Transformer architecture ([Vaswani et al., 2017](#)). Unlike an RNN-

based model, the Transformer can take in the entire input visit sequence at once, increasing training speed significantly. Each visit in an input visit sequence represents a token to the Transformer. The clinical diagnosis of the given sequence in the next visit (d_{n+1}), given their visit sequence v , is calculated by the Transformer as depicted in Figure 4.

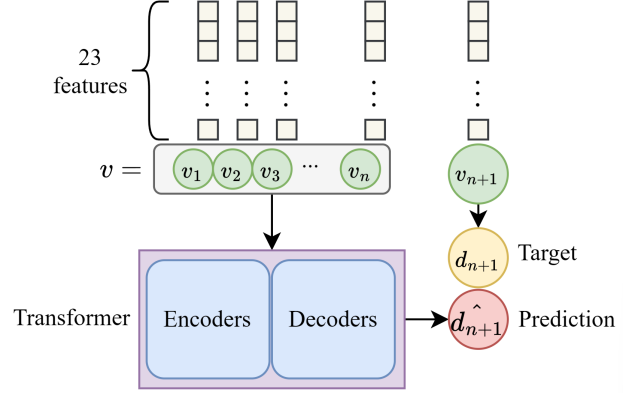


Figure 4: Prediction process for the Transformer

To compare our proposed model to a newer architecture, we also consider the Time Series Transformer (TST) ([Zerveas et al., 2021](#)).

All models are implemented using the PyTorch Python Library ([Paszke et al., 2019](#)). Some hyperparameter values worth mentioning are the number of encoder layers (4), the number of decoder layers (8), the number of attention heads (4), and input dimensionality (256). The total number of parameters for the Transformer model is 6,603,267.

2.5. Training and Evaluation

We train and evaluate separate models on raw and balanced datasets. We use 10-fold cross-validation to train and evaluate the models. Each fold is divided such that 10% of each group will be in the test split and 90% of each group will be in the train split. We split the data temporally, ensuring that each fold has an equal proportion of sequences from each group.

We use grid search to tune appropriate hyperparameter values (number of encoder/decoder layers, number of attention heads, dimensionality of the input layer/classification layer).

To evaluate model performance, we compute the multiclass area under the curve (mAUC) following [Hand and Till \(2001\)](#), along with the F1 score. We

also report raw accuracy, sensitivity, and specificity for each of the three classes. For each class C_i (CN, MCI, AD), sensitivity and specificity are defined as:

$$\text{Sensitivity}(C_i) = \frac{TP_i}{TP_i + FN_i},$$

$$\text{Specificity}(C_i) = \frac{TN_i}{TN_i + FP_i},$$

where TP_i , FN_i , TN_i , and FP_i represent the number of true positives, false negatives, true negatives, and false positives, respectively, for class C_i . The BCA is then calculated by taking the mean of the sensitivity (Sens.) and specificity (Spec.) for each class, averaged over all 3 classes:

$$\text{BCA} = \frac{1}{3} \sum_{i=1}^3 \left(\frac{1}{2} \left(\text{Sens.}(C_i) + \text{Spec.}(C_i) \right) \right).$$

In addition to evaluating these metrics on the full dataset, we compute them separately for two subsets of the data: one containing only stable sequences and the other containing only converter sequences.

3. Results

All statistical comparisons were performed using a two-sample Welch’s t-test with $n_1 = 10$ and $n_2 = 10$ for each model. The results are reported as mean \pm standard deviation across the ten folds.

3.1. Raw Dataset Results

On the raw dataset, the TST recorded the highest mAUC, while the Transformer had the highest F1 score (Table 2). The Transformer model demonstrated the highest overall accuracy compared to the baseline RNN models. BCA followed the same trend, with the Transformer outperforming GRU, LSTM, and minRNN. Additionally, the Transformer exhibited the highest sensitivity and specificity.

Statistical testing revealed that the Transformer significantly outperformed all RNN models in BCA, even the best RNN (GRU) with $p = 4.12\text{e-}02$. This came from a statistically significant difference in sensitivity ($p = 3.85\text{e-}02$), while the difference in specificity was not significant. With regards to raw accuracy, the Transformer outperformed minRNN ($p = 1.53\text{e-}02$) but otherwise, the Transformer-based models were not significantly different than the RNNs.

Table 2: mAUC and F1 score (raw dataset)

Model	mAUC	F1 score
TST	0.938 \pm 0.01	0.812 \pm 0.03
Transformer	0.920 \pm 0.02	0.824 \pm 0.04
GRU	0.846 \pm 0.11	0.786 \pm 0.04
LSTM	0.857 \pm 0.11	0.786 \pm 0.05
minRNN	0.873 \pm 0.11	0.766 \pm 0.05

The significant difference between the two architectures is most apparent when considering the performance on stable and converter sequences. For stable sequences, the Transformer model underperformed relative to all RNN baselines, exhibiting lower accuracy, BCA, sensitivity, and specificity. Statistical tests confirmed that the Transformer was significantly worse in accuracy than GRU ($p = 7.85\text{e-}06$), LSTM ($p = 1.71\text{e-}05$), and minRNN ($p = 9.38\text{e-}06$). Furthermore, the Transformer had significantly lower specificity than GRU ($p = 3.81\text{e-}03$), and minRNN ($p = 4.37\text{e-}03$). No significant differences were observed in BCA or sensitivity for this subset of the data.

The TST achieved the highest accuracy on converter sequences, significantly outperforming all RNN models. This trend was consistent across BCA, sensitivity, and specificity as well, with both Transformer-based models outperforming the RNNs. Statistical analysis confirmed these results, with TST demonstrating significantly higher accuracy ($p = 1.48\text{e-}10$), BCA ($p = 1.01\text{e-}10$), sensitivity ($p = 1.13\text{e-}06$), and specificity ($p = 1.46\text{e-}10$) when compared to LSTM, the highest-performing RNN for this subset.

Table 3 includes a comprehensive list of the values for each metric pertaining to the overall, stable, and converter performance.

3.2. Balanced Dataset Results

On the balanced dataset, the TST recorded the highest mAUC, while the RNNs had higher F1 scores. The Transformer model outperformed all of the baseline RNNs across accuracy, BCA, sensitivity, and specificity. Statistical testing confirmed that the Transformer was significantly better than all RNN models for all metrics; for the best RNN (LSTM), we observed a significant difference in accuracy ($p = 6.59\text{e-}04$), BCA ($p = 2.44\text{e-}03$), sensitivity ($p = 1.97\text{e-}03$), and specificity ($p = 5.19\text{e-}03$).

Table 3: Performance metrics for the raw dataset across all subsets

Subset	Metric	TST	Transformer	GRU	LSTM	minRNN
Overall	Acc.	0.814 ± 0.04	0.824 ± 0.04	0.792 ± 0.04	0.789 ± 0.04	0.777 ± 0.04
	BCA	0.866 ± 0.02	0.874 ± 0.03	0.830 ± 0.05	0.829 ± 0.06	0.819 ± 0.06
	Sens.	0.828 ± 0.03	0.837 ± 0.04	0.777 ± 0.08	0.773 ± 0.08	0.761 ± 0.09
	Spec.	0.904 ± 0.02	0.910 ± 0.02	0.884 ± 0.04	0.885 ± 0.04	0.877 ± 0.04
Stable	Acc.	0.856 ± 0.04	0.883 ± 0.04	0.984 ± 0.01	0.973 ± 0.03	0.986 ± 0.01
	BCA	0.907 ± 0.02	0.922 ± 0.02	0.951 ± 0.05	0.939 ± 0.07	0.950 ± 0.05
	Sens.	0.880 ± 0.03	0.898 ± 0.03	0.921 ± 0.08	0.905 ± 0.10	0.920 ± 0.08
	Spec.	0.934 ± 0.02	0.945 ± 0.02	0.981 ± 0.03	0.972 ± 0.04	0.981 ± 0.03
Converter	Acc.	0.694 ± 0.08	0.648 ± 0.14	0.184 ± 0.07	0.217 ± 0.08	0.129 ± 0.07
	BCA	0.662 ± 0.04	0.648 ± 0.05	0.464 ± 0.03	0.476 ± 0.04	0.442 ± 0.03
	Sens.	0.446 ± 0.03	0.430 ± 0.05	0.320 ± 0.04	0.334 ± 0.05	0.298 ± 0.04
	Spec.	0.879 ± 0.05	0.866 ± 0.06	0.609 ± 0.02	0.618 ± 0.03	0.586 ± 0.03

Table 4: Performance metrics for the balanced dataset across all subsets

Subset	Metric	TST	Transformer	GRU	LSTM	minRNN
Overall	Acc.	0.779 ± 0.04	0.802 ± 0.05	0.648 ± 0.04	0.684 ± 0.07	0.623 ± 0.04
	BCA	0.822 ± 0.03	0.839 ± 0.04	0.758 ± 0.03	0.770 ± 0.05	0.742 ± 0.03
	Sens.	0.766 ± 0.04	0.788 ± 0.05	0.684 ± 0.03	0.699 ± 0.06	0.666 ± 0.04
	Spec.	0.878 ± 0.03	0.890 ± 0.03	0.832 ± 0.02	0.842 ± 0.04	0.818 ± 0.03
Stable	Acc.	0.787 ± 0.07	0.814 ± 0.07	0.974 ± 0.02	0.926 ± 0.06	0.956 ± 0.03
	BCA	0.866 ± 0.03	0.883 ± 0.05	0.910 ± 0.07	0.885 ± 0.08	0.900 ± 0.06
	Sens.	0.829 ± 0.04	0.849 ± 0.06	0.859 ± 0.09	0.832 ± 0.11	0.849 ± 0.08
	Spec.	0.904 ± 0.03	0.917 ± 0.03	0.960 ± 0.05	0.939 ± 0.06	0.951 ± 0.04
Converter	Acc.	0.775 ± 0.06	0.792 ± 0.05	0.323 ± 0.08	0.439 ± 0.14	0.291 ± 0.05
	BCA	0.703 ± 0.03	0.713 ± 0.03	0.512 ± 0.03	0.566 ± 0.06	0.499 ± 0.02
	Sens.	0.495 ± 0.05	0.504 ± 0.05	0.358 ± 0.05	0.419 ± 0.08	0.346 ± 0.04
	Spec.	0.910 ± 0.03	0.922 ± 0.02	0.666 ± 0.02	0.713 ± 0.05	0.651 ± 0.02

Table 5: mAUC and F1 score (balanced dataset)

Model	mAUC	F1 score
TST	0.890 ± 0.033	0.721 ± 0.045
Transformer	0.883 ± 0.024	0.746 ± 0.064
GRU	0.846 ± 0.107	0.786 ± 0.041
LSTM	0.857 ± 0.114	0.786 ± 0.048
minRNN	0.873 ± 0.114	0.766 ± 0.049

For stable sequences, the Transformer model exhibited lower accuracy, BCA, sensitivity, and specificity

compared to all RNN models. Statistical tests confirmed that the Transformer was significantly worse in accuracy than GRU ($p = 4.08\text{e-}02$), LSTM ($p = 1.63\text{e-}01$), and minRNN ($p = 6.74\text{e-}02$). However, the differences across models in BCA were not statistically significant. The differences in sensitivity and specificity were also not significant, besides that GRU was significantly better in specificity ($p = 2.79\text{e-}02$).

The Transformer achieved the highest accuracy on converter sequences, substantially outperforming each of the RNNs. This trend was also consistent for BCA, sensitivity, and specificity. This difference was statistically significant; when compared to the best RNN in this subset (LSTM), we observe significantly

higher accuracy ($p = 1.20\text{e-}05$), BCA ($p = 1.57\text{e-}05$), sensitivity ($p = 1.25\text{e-}02$), and specificity ($p = 5.16\text{e-}08$).

Table 4 includes a comprehensive list of the values for each metric pertaining to the overall, stable, and converter performance.

3.3. Visit Sequence Length Assessment

Overall, both the Transformer and RNN models exhibit similar trends in performance across different visit sequence lengths, as illustrated in Figure 5. However, a key distinction emerges when comparing stable and converter sequences, particularly among those with limited sequence history. The Transformer model consistently achieves higher BCA for converters in these early visit groups. This improvement, however, comes at the expense of lower performance in stable sequences.

Statistical analysis confirms this contrast in predictive performance. Among sequences with shorter visit histories, the Transformer significantly outperforms all RNN models in distinguishing converters, with p -values reaching as low as $4.24\text{e-}10$ in the earliest group. The Transformer maintains a significant advantage in groups 2 and 3 ($p = 1.04\text{e-}05$ for group 3). Beyond this point, differences between models become statistically insignificant. The overall trajectory of performance for converter sequences is depicted in Figure 7.

This improvement in converter classification comes with a corresponding reduction in stable sequence performance, particularly in early visit groups. The Transformer performs significantly worse than all RNN models in these cases, with statistical significance reaching $p = 3.41\text{e-}03$ in the shortest sequence group and remaining significant across subsequent groups. These can be seen in Figure 6. Outside of the earliest visit groups and a few other exceptions, model performances are statistically indistinguishable beyond group 4, with a few minor exceptions.

An ablation study was conducted in order to further investigate the importance of visit history. Additional Transformer models were trained using a limited number of visits from each sequence. Specifically, one model was trained using only the penultimate visit (final visit prior to the target visit), while another was trained using up to the last four visits, depending on how many were available. The results are summarized in Table 6.

Table 6: Transformer BCA change (%) when trained on limited visit history (raw dataset)

Visit History	BCA	% Change
Last \leq 4 Visits	0.830	-5.040%
Last Visit Only	0.816	-6.619%

3.4. Feature Importance Assessment

To assess the contribution of different feature categories to model performance, an ablation study was conducted using the Transformer model on the raw dataset. Categories of features were systematically removed, including cognitive scores, volumetric MRI data, and biomarkers. Separate models were trained with these reduced feature sets, and the resulting change in BCA was recorded. The results, contained in Table 7, illustrate the impact of each feature category on model performance.

Table 7: Transformer BCA change (%) when removing a feature category (raw dataset)

Feature Set Removed	BCA	% Change
Cognitive Scores	0.736	-15.761%
Volumetric MRI	0.874	0.064%
Biomarkers	0.868	-1.276%

In addition to the ablation study, separate models were trained using only cognitive scores, only volumetric MRI data, and only biomarkers. Comparing the BCA of these single-feature-category models to that of the full-feature model provides further insight into the relative importance of each feature category, as shown in Table 8.

Table 8: Transformer BCA change (%) when trained on a single feature category (raw dataset)

Feature Set Used	BCA	% Change
Cognitive Scores	0.868	-0.655%
Volumetric MRI	0.792	-9.314%
Biomarkers	0.821	-6.065%

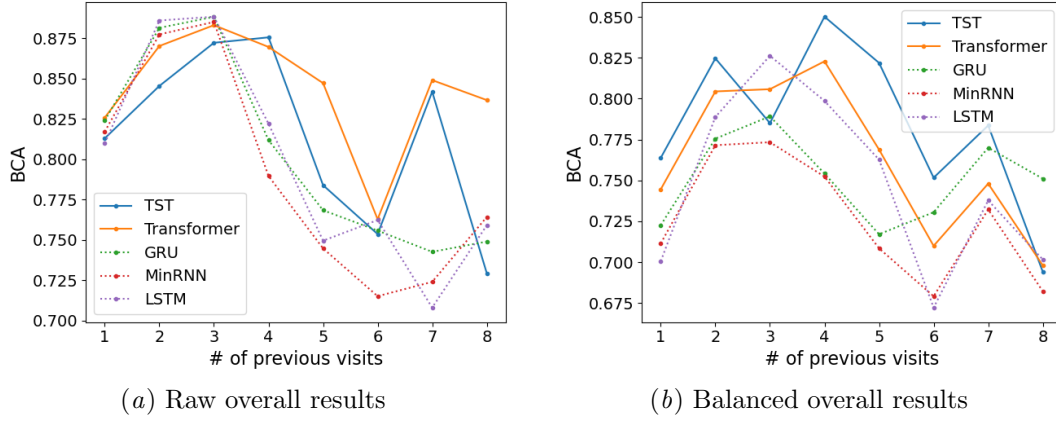


Figure 5: Model BCA on all sequences

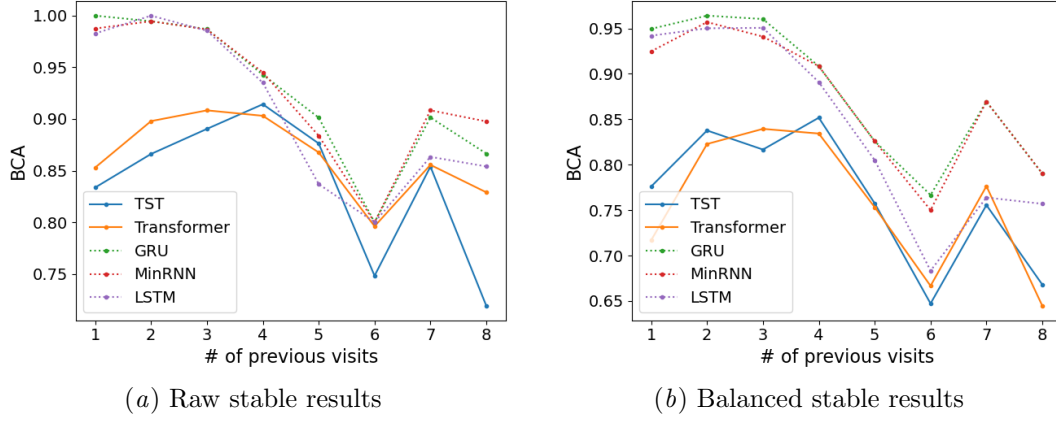


Figure 6: Model BCA on stable sequences

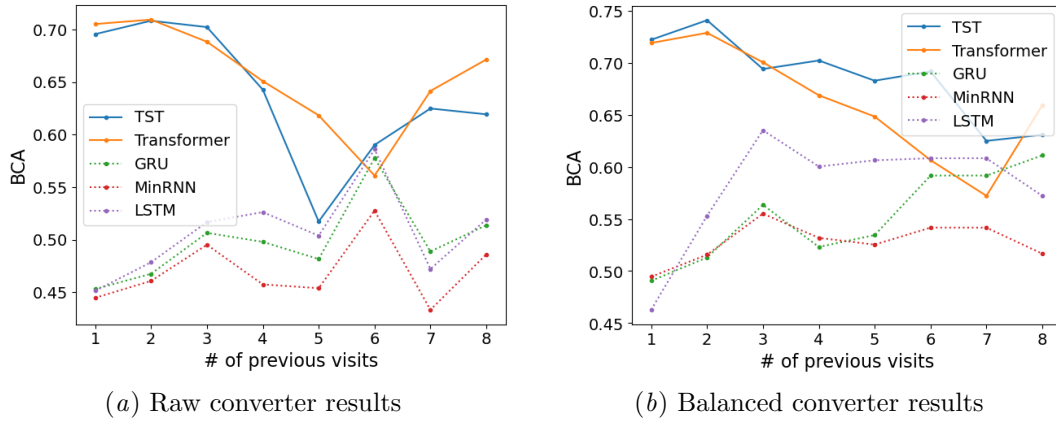


Figure 7: Model BCA on converter sequences

4. Discussion

The results indicate that Transformer-based models significantly outperformed RNNs in identifying converter sequences. In contrast, the RNN models, especially on raw data, appeared to perform well but primarily by predicting stable sequences, often failing to recognize converters. While the Transformer performed slightly worse for stable sequences, this trade-off was offset by a substantial improvement in predicting converters. Identifying converters is critical in Alzheimer’s disease research because early detection allows for timely intervention, which can slow disease progression (Livingston et al., 2020). Thus, a false negative carries a much higher cost than a false positive.

Further analysis of the length of visit sequences reveals deeper insights into these differences. Both model types followed similar performance trends across varying sequence lengths, but key distinctions emerged in the stable and converter subsets. The Transformer exhibited a clear advantage in predicting conversion, particularly for shorter sequence histories. Statistical analysis confirmed that this advantage was most pronounced in the earliest visit groups, with significant performance differences between the Transformer-based models and all RNN baselines. This improvement in conversion prediction came at the slight expense of stable sequence classification, particularly in early visit groups.

Interestingly, model performance does not consistently improve for longer visit sequences, contrary to expectations. While this trend may suggest difficulties in generalizing to longer sequences, we propose several alternative factors that could contribute to this pattern. First, we note that there is much less data for these higher groups; as evident in Figure 1(a), groups 5-8 only account for about 25% of the raw dataset. Also, as was seen in Figure 2, the future diagnosis for groups 5 and 6 is disproportionately composed of MCI, which is notoriously challenging to predict. Additionally, we observed that longer visit histories tend to exhibit greater intervals between consecutive visits (10.02 months between visits on average, compared to 6.81), introducing more inconsistencies in the temporal structure of the data.

Another key observation is that longer visit histories tend to exhibit a weaker overall trend across visits. A linear regression analysis of each feature reveals that despite having more data points, sequences in groups 5-8 generally have smaller regression slopes,

indicating less pronounced change over time. Figure 8 illustrates this effect for ADAS13 scores across the entire raw dataset.

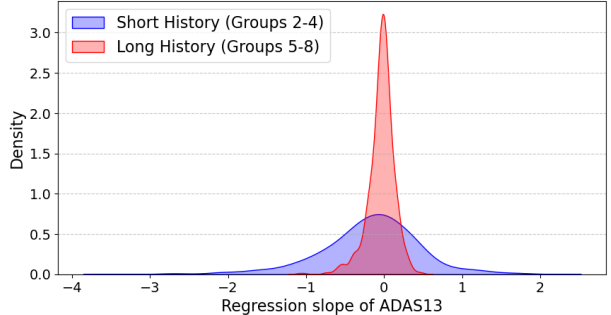


Figure 8: Regression slope of ADAS13 scores across visit sequences, short vs. long history

Despite these challenges, the model still shows evidence of learning from longer visit histories, as seen in Table 6. These findings suggest that the model leverages information from longer visit histories, as performance degrades when fewer visits are provided. However, the decrease in BCA is modest in both cases, which highlights the importance of the most recent visits in making an accurate diagnosis. This result aligns with clinical expectations, as later visits likely contain more relevant information regarding disease progression. Nonetheless, the ability of the full-sequence model to outperform the ablated versions reinforces the value of visit history, even when this may introduce additional challenges.

The results from the feature ablation studies highlight the central role of cognitive scores in predicting disease progression. The substantial change in BCA when cognitive scores were removed (-15.76%) suggests that they provide the most information, while the minimal impact of removing volumetric MRI and biomarkers indicates that these features contribute less directly to model performance. When used in isolation, cognitive scores alone achieved nearly the same BCA as the full feature set, whereas volumetric MRI and biomarker features yielded greater performance declines.

Further analysis of misclassified sequences reinforces the importance of cognitive scores. Many classification errors occurred when cognitive scores were inconsistent with the overall average for that class. Better cognitive scores often led to incorrect predictions of stability in individuals who later converted,

while low scores sometimes resulted in the opposite. An example of the former can be seen in Figure 9, where the AD sequences that were misclassified as MCI by the Transformer exhibited much lower ADAS13 scores in the penultimate visit than that of the typical AD visit. The difference in these distributions is statistically significant according to the results of a Mann-Whitney U test ($p = 1.89\text{e-}19$). We found that this is the case for many of the cognitive scores when comparing misclassified sequences to the rest of the data. These patterns suggest that while cognitive scores are highly predictive, the model may struggle to account for individual variability in testing performance.

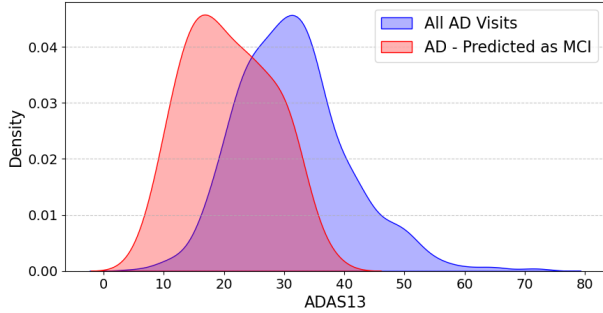


Figure 9: ADAS13 distribution: All AD visits vs. penultimate visits for AD sequences incorrectly classified as MCI.

Overall, these findings highlight the importance of model selection in predicting disease progression. A model that performs well on the majority class but fails to detect conversion cases may not be useful in certain clinical applications like Alzheimer’s disease prediction. Visit sequence analysis further emphasizes that Transformer-based models may provide a more useful approach by prioritizing early detection, even if it comes at the cost of slightly lower performance on stable sequences.

5. Limitations and Future Work

Despite the improvements demonstrated by the Transformer model, several limitations remain. One key limitation is feature selection, as the model was trained on structured clinical data but did not incorporate raw imaging modalities such as MRI or PET scans. While volumetric measures (e.g., hippocam-

pus volume, whole brain volume) and PET-derived biomarkers (e.g., amyloid and tau levels) were included, these features provide summarized information rather than the full spatial detail available in raw scans. Future work could explore the integration of deep learning-based feature extraction from imaging data to enhance the model’s predictive capabilities.

Another notable finding was that the model did not show performance gains when provided with longer visit histories, contrary to expectations. Future research could explore alternative methods to selectively prioritize the most relevant historical data.

Furthermore, conversion prediction remains a difficult task despite the improvements seen with the Transformer model. While the model significantly outperformed RNNs in identifying converters, particularly in shorter visit sequences, performance remains suboptimal, highlighting the inherent challenge of predicting disease transition. By addressing these limitations, future research can refine Transformer-based models for longitudinal disease prediction, ultimately improving their clinical utility in early detection and intervention strategies.

6. Conclusion

Our study demonstrates the effectiveness of our proposed Transformer model in predicting Alzheimer’s disease progression, particularly in identifying converter sequences. Compared to recurrent neural networks, the Transformer demonstrated a more balanced classification, reducing bias toward the majority class while improving sensitivity to early-stage transitions. The model’s improved performance, especially for sequences with shorter visit histories, suggests that attention-based architectures are better suited for early detection tasks.

These findings emphasize the role of model selection in clinical applications where early intervention is critical. By prioritizing the detection of high-risk individuals, Transformer-based models offer a promising approach to improve Alzheimer’s disease diagnosis and support timely medical decision-making.

References

- 2024 alzheimer’s disease facts and figures. *Alzheimer’s & Dementia*, 20:3708 – 3821, 2024. doi: <https://doi.org/10.1002/alz.13809>.

- Samsuddin Ahmed, Kyu Yeong Choi, Jang Jae Lee, Byeong C. Kim, Goo-Rak Kwon, Kun Ho Lee, and Ho Yub Jung. Ensembles of patch-based classifiers for diagnosis of alzheimer diseases. *IEEE Access*, 7:73373–73383, 2019.
- Sait Alp, Taymaz Akan, Md. Shenuarin Bhuiyan, Elizabeth A. Disbrow, Steven A. Conrad, John A. Vanchiere, Christopher G. Kevil, and Mohammad Alfrad Nobel Bhuiyan. Joint transformer architecture in brain 3d mri classification: its application in alzheimer’s disease classification. *Scientific Reports*, 14, 2024.
- Xin Bi, Xiangguo Zhao, Hong Huang, Deyang Chen, and Yuliang Ma. Functional brain network classification for alzheimer’s disease detection with deep features and extreme learning machine. *Cognitive Computation*, 12:513 – 527, 2019.
- Minmin Chen. Minimalrnn: Toward more interpretable and trainable recurrent neural networks, 2018. URL <https://arxiv.org/abs/1711.06788>.
- Qihui Chen and Yi Hong. Longformer: Longitudinal transformer for alzheimer’s disease classification with structural mris, 2023. URL <https://arxiv.org/abs/2302.00901>.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.
- Ruoxuan Cui and Manhua Liu. Hippocampus analysis by combination of 3-d densenet and shapes for alzheimer’s disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 23:2099–2107, 2019.
- Ruoxuan Cui, Manhua Liu, and Alzheimer’s Disease Neuroimaging Initiative. Rnn-based longitudinal analysis for diagnosis of alzheimer’s disease. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 73:1–10, 2019.
- David J. Hand and Robert J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.
- Gia-Minh Hoang, Ue-Hwan Kim, and Jae Gwan Kim. Vision transformers for the prediction of mild cognitive impairment to alzheimer’s disease progression using mid-sagittal smri. *Frontiers in Aging Neuroscience*, 15, 2023. doi: <https://doi.org/10.3389/fnagi.2023.1102869>.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.
- Clifford R. Jack, Matt A. Bernstein, Nick C Fox, Paul M. Thompson, Gene E. Alexander, Danielle J. Harvey, Bret J. Borowski, Paula J. Britson, Jennifer L Whitwell, Chadwick P. Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L. G. Hill, Ronald J. Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Philip Mallozzi, Daniel J. Blezek, Joshua Levy, Josef P. Debbins, Adam Fleisher, Marilyn S. Albert, Robert C. Green, Georgeartzokis, Gary H. Glover, John P. Mugler, and Michael Weiner. The alzheimer’s disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27, 2008. doi: <https://doi.org/10.1002/jmri.21049>.
- Huanhuan Ji, Zhenbing Liu, Wei Qi Yan, and Reinhard Klette. Early diagnosis of alzheimer’s disease based on selective kernel network with spatial attention. In *Asian Conference on Pattern Recognition*, 2019.
- Yang Li, Yaping Wang, Guorong Wu, Feng Shi, Luping Zhou, W. Lin, and Dinggang Shen. Discriminant analysis of longitudinal cortical thickness changes in alzheimer’s disease using dynamic and network features. *Neurobiology of Aging*, 33: 427.e15–427.e30, 2012.
- Gill Livingston, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, Sergi G Costafreda, Amit Dias, Nick Fox, Laura N Gitlin, Robert Howard, Helen C Kales, Mika Kivimäki, Eric B Larson, Adesola Oggunniyi, Vasiliki Orgeta, Karen Ritchie, Kenneth Rockwood, Elizabeth L Sampson, Quincy Samus, Lon S Schneider, Geir Selbæk, Linda Teri, and Naaheed Mukadam. Dementia prevention, intervention, and care: 2020 report of the

- lancet commission. *The Lancet*, 396, 2020. doi: 10.1016/S0140-6736(20)30367-6.
- Razvan V. Marinescu, Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof, Nick C. Fox, Stefan Klein, Daniel C. Alexander, and the Euro-POND Consortium. Tadpole challenge: Prediction of longitudinal evolution in alzheimer’s disease, 2018. URL <https://arxiv.org/abs/1805.03909>.
- Susanne G. Mueller, Michael W. Weiner, Leon Thal, Ronald C. Petersen, Clifford R. Jack, William J. Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel A Beckett. Ways toward an early diagnosis in alzheimer’s disease: The alzheimer’s disease neuroimaging initiative (adni). *Alzheimer’s & Dementia*, 1:55–66, 2005.
- Minh-Quan Nguyen, Tong He, Lijun An, Daniel C. Alexander, Jiashi Feng, and B. T. Thomas Yeo. Predicting alzheimer’s disease progression using deep recurrent neural networks. *NeuroImage*, 222: 117203 – 117203, 2019.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
- M. Khojaste Sarakhsi, Seyedhamidreza Shahabi Haghighi, Seyyed M. T. Fatemi Ghomi, and Elena Marchiori. Deep learning for alzheimer’s disease diagnosis: A survey. *Artificial intelligence in medicine*, 130:102332, 2022. doi: <https://doi.org/10.1016/j.artmed.2022.102332>.
- Philip Scheltens, Kaj Blennow, Monique M B Breteler, Bart de Strooper, Giovanni B Frisoni, Stephen Salloway, and Wiesje Maria Van der Flier. Alzheimer’s disease. *The Lancet*, 388 (10043):505–517, 2016. ISSN 0140-6736. doi: [https://doi.org/10.1016/S0140-6736\(15\)01124-1](https://doi.org/10.1016/S0140-6736(15)01124-1). URL <https://www.sciencedirect.com/science/article/pii/S0140673615011241>.
- Simeon E. Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio’, and Nicola Toschi. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer’s disease. *NeuroImage*, 189:276–287, 2018.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.
- George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD ’21*, pages 2114–2124, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467401. URL <https://doi.org/10.1145/3447548.3467401>.
- Daoqiang Zhang and Dinggang Shen. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PLoS ONE*, 7, 2012.