Pattern Analysis and Applications (2025) 28:64
https://doi.org/10.1007/s10044-025-01447-4

ORIGINAL ARTICLE

Conversion-aware forecasting of Alzheimer's disease via featurewise attention

Elvan Karasu[1] · **İnci M. Baytaş**[1]

Received: 18 December 2024 / Accepted: 4 March 2025 / Published online: 20 March 2025
© The Author(s) 2025

Abstract
Alzheimer's disease (AD) is a neurodegenerative disorder that leads to cerebral atrophy, impacting memory and cognitive abilities. A precursor to AD known as Mild Cognitive Impairment (MCI) shows subtle symptoms that do not overwhelm the patients' daily activities. MCI patients might eventually progress to AD in later stages. Early detection of the conversion is a vital step in preventative treatment planning. However, conversion detection is challenging due to the rarity of conversion visits in public datasets and the unknown nature of the conversion. This study aims to improve conversion detection with an attention-based architecture designed to encode input biomarkers and time into a shared space where time and attribute embeddings are fused with attention. Temporal information is incorporated as a separate modality with time embeddings to capture the correlation between time and feature significance for the model's predictions. Experiments with widely used public databases (TADPOLE and NACC) show encouraging performance in conversion detection. In TADPOLE, a conversion recall of 74.3%, significantly outperforming baseline models such as logistic regression (36.9%) and Long Short-Term Memory networks (62.3%), is reported while maintaining an area under the curve (AUC) score of 82.0%. In NACC, our model demonstrates a competitive conversion recall of 71.6% and an AUC of 82.6%. The experimental results highlight the contribution of the attention between time and attributes to MCI-AD conversion recall. The experimental analyses hold promise for assisting physicians in designing targeted preventative treatment strategies for at-risk individuals. The implementation of the proposed method is available at https://github.com/ALLab-Boun/FATE-Net.

Keywords Self-attention · Time embedding · Transformer network · Alzheimer's disease · Mild cognitive impairment

1 Introduction

Alzheimer's disease (AD), known as the prevalent cause of dementia, causes a steady degradation in the cognitive and physical abilities of patients [1]. The monitoring of the early stages, which could be diagnosed as Mild Cognitive Impairment (MCI), is critical to detecting the risk of converting to AD. Predicting the progression of MCI to AD ahead of time enables treatment planning to alleviate the symptoms that complicate the daily activities of patients and their loved ones [2]. Prediction of the conversion can be posed as a classification task where the patient data constitutes a multivariate time series. Consequently, machine learning models are inevitable candidates for solving this classification problem with a temporal nature.

Machine learning is widely used in biomedical informatics to make the most effective use of patient data. Deep learning as a subfield of machine learning offers a wide range of neural network architectures and training frameworks to tackle complex healthcare tasks, including patient subtyping [3], phenotyping [4], and disease forecasting [5]. The success of deep learning models stems from learning task-specific representations of patients and their visits. In the MCI-AD conversion prediction problem, obtaining visit representations that differentiate a stable course of MCI visits from a progressive one is essential. In this regard, deep learning models offer the capacity for encoding several factors leading to conversion into visit representations. For this purpose, multi-layer perceptrons, convolutional neural networks, and recurrent neural networks could be considered.

✉ İnci M. Baytaş
inci.baytas@bogazici.edu.tr

Elvan Karasu
elvan.karasu@std.bogazici.edu.tr

[1] Computer Engineering Department, Boğaziçi University, Bebek, İstanbul 34342, Turkey

Recurrent Neural Network (RNN) and its popular variants, such as Long Short-Term Memory (LSTM) network [6], are widely used to capture temporal dependencies in temporal patient data. Since the symptoms of neurodegenerative diseases change over time, patients' conditions are tracked by their neurologists to change the course of treatment according to the stage of the disease. Therefore, sequential modeling is beneficial in capturing the temporal pattern of the disease's progression. Similarly, transformer networks successfully employed in natural language processing and vision domains can be used for temporal modeling [7]. The most prominent feature of the transformer network is the self-attention mechanism that enables learning how to attend each patient visit to construct a representation of the current visit. Although the follow-up data of an MCI patient visiting every three to six months can be considered a time series, it would be much shorter than, for instance, the periodic physiological measurements of an ICU patient. Therefore, whether an RNN could learn a meaningful temporal pattern from a short sequence of visits is ambiguous. Besides, deciphering the underlying attributes contributing to the prediction of the AD conversion is as essential as the prediction itself. For this reason, an interpretable association between the prediction, visit time, and input attributes is critical in the predictive modeling of AD.

This study proposes an inherently interpretable feed-forward architecture named Feature Attentive Network with Time Embeddings (FATE-Net) to improve the early conversion prediction between MCI and AD diagnoses. FATE-Net learns how to attend to each input biomarker by employing self-attention, where the Time2Vec [8] embedding of the visit time is treated as a query. This study, unlike the studies prioritizing biomarker selection with traditional machine learning models, focuses on designing an architecture that learns how to transform input biomarkers into a prediction. On the other hand, unlike studies proposing deep architectures as black-boxes, FATE-Net aims to improve the explainability of the model's predictions. In addition, time is integrated into the architecture in a more explicit manner compared with studies applying recurrent architectures and transformer networks to the sequential input. The visit time embedding is used to control the contributions of each biomarker while learning a latent representation of the visit. FATE-Net is trained with tabular patient data comprising cognitive test scores and volumetric measurements obtained from brain imaging. To be able to diagnose a patient with MCI and AD, cognitive tests comprising oral and written evaluations of a patient's cognition and neuroimaging are often administered in real life [5]. A patient's memory and reasoning skills and ability to perform their daily activities are mainly scored. Neurologists prefer to administer cognitive tests since they are not invasive and lead to an initial diagnosis [2]. Although the cognitive scores are beneficial for diagnosis, they are insufficient for a neurologist to determine the final verdict since cognitive test results could be affected by external factors such as the patient's education level. Therefore, neurologists often analyze neuroimaging of the brain along with the cognitive scores to be sure since volumetric changes in the brain could be observed in the progressive stages of AD [9].

The contributions of the proposed model are outlined below.

- As the importance of input biomarkers might change with time in neurodegenerative diseases, it is essential to address this aspect in predictive modeling. FATE-Net treats time as a separate modality through time embeddings and computes attention between time embedding and attribute embeddings. Thus, learning how time attends differently to input biomarkers inherently creates an opportunity to comprehend each attribute's contributions to conversion prediction.
- Time information is integrated through a time embedding module rather than employing a recurrent structure. The recurrent and transformer architectures aim to capture and incorporate the temporal pattern of the sequence. However, the explicit correlation between disease progression and time is essential in neurodegenerative diseases. For this reason, FATE-Net aims to learn how to embed the visit time that could play a significant role in the MCI-AD conversion while learning a visit's representation. Furthermore, the attention module allows the time embeddings to control the aggregation of the attribute embeddings to obtain the visit representation.
- The proposed architecture does not require sequential patient data to be trained. FATE-Net could be trained with cross-sectional data when multiple sequential visits are unavailable during training. FATE-Net is also applicable to end-to-end training with sequential patient data when coupled with sequential modeling. Therefore, FATE-Net applies to various scenarios where patients have few visits to sequential visits spanning years.
- FATE-Net's feature attention applies to any healthcare task with tabular input biomarkers and time information.

Experiments on two publicly available datasets report a significant improvement in MCI to AD conversion recall. The performance of FATE-NET is compared with commonly used machine learning and deep learning models in the AD prediction literature. A case study is provided to present the potential use of the proposed framework.

2 Related work

This section presents studies in the literature focusing on early prediction of AD. Some studies aim to discover the most clinically plausible input biomarkers facilitating the predictive tasks performed by traditional Machine Learning (ML) methods. In this regard, Irfan *et al.* proposed a feature selection method to discover vital cognitive features for early AD detection [5]. An ensemble of ML models with an adaptive voting approach is trained with the vital features. Likewise, Franciotti et al. applied feature selection and suggested some cognitive, Magnetic Resonance Imaging (MRI), and Position Emission Tomography (PET) biomarkers based on performances of Random Forest (RF) and gradient boosting classifiers [10].

A hybrid prognostic unsupervised and supervised ML framework was introduced by Bucholc et al. to identify MCI patients with high risk of conversion to AD [2]. The authors trained RF, Support Vector Machine (SVM), Logistic Regression (LR), and ensemble models with clinical scores. On the other hand, another study focused more on the input variables, using factor analysis techniques to extract latent variables from neuropsychological, MRI, and Cerebrospinal Fluid (CSF) predictors [11]. The proposed latent variables are tested with the LR classifier to predict conversion from MCI to AD.

Some studies categorize MCI patients based on whether they eventually convert to AD to study the progression from MCI to AD. MCI patients who convert to AD over time are classified as progressive MCI (pMCI), while those who do not convert are categorized as stable MCI (sMCI) [2, 11]. This categorization helps understand the heterogeneity of MCI and enables the development of predictive models specifically tailored to identify individuals at higher risk of progression. Similarly, Kung et al. identified the most suitable neuroimaging biomarkers to predict MCI to AD conversion, training multi-layer perceptron to distinguish between sMCI and pMCI patients [9]. Our study also focuses on MCI to AD conversion, a challenging but critical task, given that this transition is relatively rare in publicly available datasets [12–14]. While predicting stable MCI and AD cases is easier due to greater data availability, the transition from MCI to AD is a critical aspect that neurologists need to be alerted to. Therefore, a key focus of our work is providing timely and interpretable predictions for this important progression.

The studies mentioned above often rely on traditional machine learning models, such as support vector machines, logistic regression, and tree-based classifiers, primarily for benchmarking purposes. Consequently, their contributions are typically limited to extracting and evaluating new features using these well-established models. While using relevant biomarkers is crucial for prediction, the biomarker selection or discovery process often requires domain expertise. In contrast, FATE-Net employs an attribute encoder that projects each feature into a higher-dimensional space, enabling the model to learn feature importances directly from the data in a data-driven manner.

While the studies presented above commonly train shallow models with various input biomarkers, some studies train deep models to detect patients at high risk of AD conversion. For instance, Tabarestani et al. employed LSTM and gated recurrent unit (GRU) for predicting future diagnoses of MCI and cognitively normal (CN) patients [15]. The authors also trained the networks to solve the regression task of predicting future Mini-Mental State Examination (MMSE) scores. Since the scale of the patient data is often small to train a complex model, less complex deep models are preferred to avoid overfitting in healthcare-related tasks. Therefore, Nguyen *et al.* proposed an RNN model with fewer parameters than LSTM networks that can inherently handle missing values [16]. The authors used temporal patient data comprising cognitive and MRI biomarkers. Imputing patient data in a clinically plausible way can be critical. Their proposed RNN model is trained to predict the missing values of the next visit [16].

An exciting study by El-Sappagh et al. proposed to train an LSTM network in a two-stage framework [17]. The first stage predicts the disease progression, while the second stage is a regression task to predict the time of MCI to AD conversion [17]. The authors suggested that the LSTM network performed better than the shallow models, such as LR, SVM, and RF. However, the authors also discussed that the non-linear projections in the LSTM architecture constitute the model's main limitation, which is the lack of interpretability [17]. Building open these limitations, FATE-Net is designed to enhance the explainability of the model's predictions while explicitly integrating time into the architecture. Unlike recurrent architectures such as LSTM and GRU or transformer networks, FATE-Net incorporates visit time embeddings to control the contributions of each biomarker, allowing the model to learn a latent representation of the visit. This structured integration of temporal information improves interpretability and provides more detailed insights into disease progression.

Convolutional Neural Networks (CNNs) and vision transformer-based architectures are commonly used when MRI and PET images are considered [18, 19]. Kim et al. [20], for example, used CNNs with an Inception-V4 architecture for diagnosing AD from 2D MRI slices, achieving a performance level better than clinicians on average. Additionally, neuroimaging data from different modalities can be combined using fusion approaches. In one study, researchers used PET and MRI images from the Open Access Series of Imaging Studies (OASIS) dataset and developed a model

that combines Inception and ResNet50 architectures to classify patients into CN, MCI, and AD categories [21]. While these approaches achieve strong performance, they rely on high-dimensional image data, which can be expensive to acquire repeatedly in clinical settings and computationally costly to process. Additionally, such methods often lack temporal information and interpretability, both of which are critical for understanding disease progression and identifying contributing factors.

In contrast, FATE-Net is designed to take tabular input, such as cognitive scores and volumetric MRI variables. As a result, we do not compare the proposed method with frameworks that rely on medical images, which may be less accessible in some clinical environments. FATE-Net incorporates the time factor by learning temporal embeddings and inherently discovers relationships between outcomes and input biomarkers through self-attention. This design improves the MCI-to-AD recall and automatically highlights feature importance, enhancing interpretability.

3 Method

The proposed framework, given in Fig. 1, comprises time and attribute encoders that map the time variable and input biomarkers to a shared space. The time embedding is treated as a query, and attributes are projected to key and value embeddings. Thus, the time embedding specifies how to attend to each attribute before obtaining the final embedding used in the classification task. The proposed architecture aims to learn a visit representation as a weighted combination of the input biomarkers. In addition, time is integrated into the architecture as an additional modality unlike the studies that apply recurrent architectures and transformer networks to the sequential input. The visit time embedding is used to control the contributions of each biomarker while learning a latent representation of the visit. The following sections present the details of the framework.

3.1 Time encoder

Time is a critical factor in neurodegenerative diseases. Temporal change in an MCI patient's variables is an essential indicator of the disease progression. However, the visit frequency of MCI and AD patients is often around six months, considered an acceptable elapsed time to observe the changes in cognitive scores and MRI results [13]. Therefore, we could not expect to have a sufficiently large number of visits to capture a temporal pattern effectively. To address this, using a time encoder, we integrate time as a separate modality into the representation learning.

Inspired by Kazemi et al., we adopted the Time2Vec representation given below to learn a vector representation from continuous time point τ_i [8].

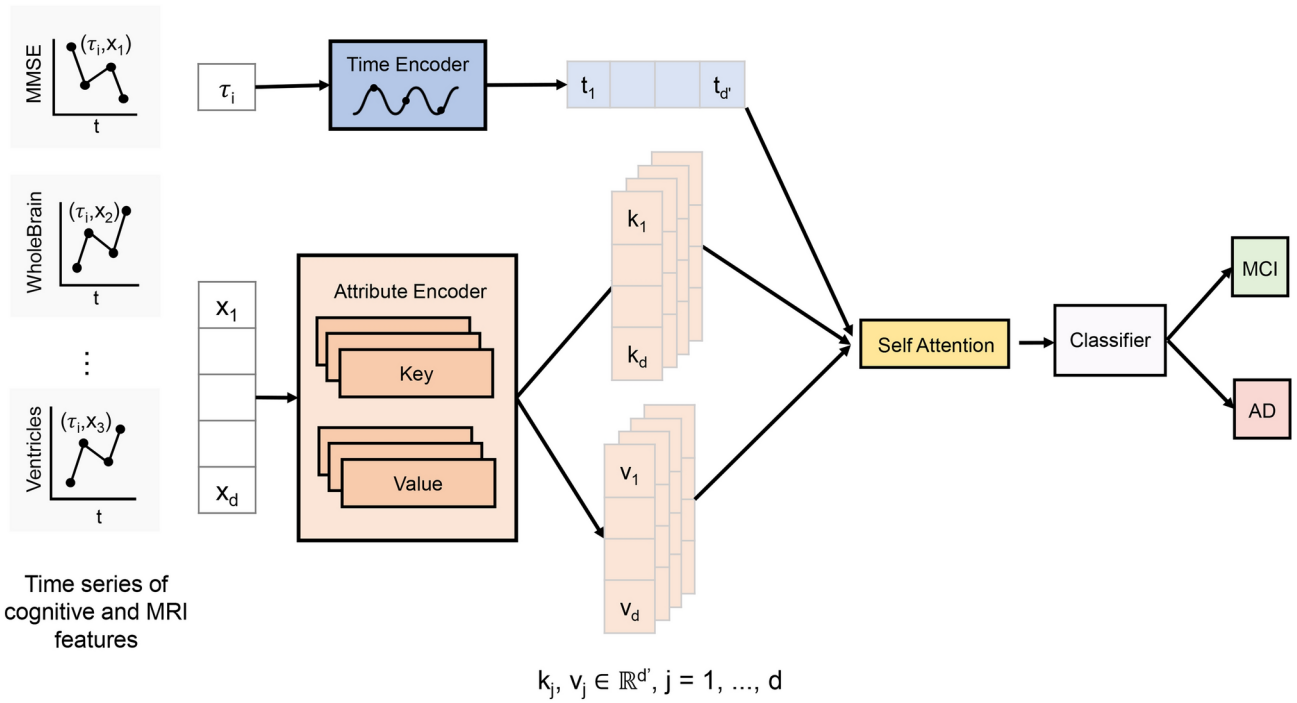


Fig. 1 FATE-Net architecture. Inspired by Time2Vec [8], the time encoder outputs a time embedding used as the query in the self-attention layer. The attribute encoder maps each input dimension into a space shared by the time embedding. The attribute embeddings are aggregated within the self-attention layer

$$\mathrm{t2v}\left(\tau_i\right)[j]=\begin{cases}\omega_j\tau+\varphi_j, & j=0\\ \sin\left(\omega_j\tau+\varphi_j\right), & 1\le j\le d'-1\end{cases} \tag{1}$$

where $\mathrm{t2v}\left(\tau_i\right)[j]$ is the *j*th element of the d'-dimensional time embedding of *i*th time step, and ω_j and φ_j are trainable parameters.

Time2Vec embeddings provide a nonlinear representation of time. Modeling the visit time as a modality on its own is particularly crucial since neurodegenerative diseases, such as AD, worsen over time. Time embeddings are learned from data in a task-specific manner. Therefore, temporal dynamics of the biomarkers contributing to the task could be encoded in the time embeddings. The importance of biomarkers evolves during disease progression since the changes in the brain structure might become more visible in neuroimaging at later stages of the disease [22]. In this study, we compute attention between time embeddings and attribute embeddings to highlight how time influences biomarker importance.

In the literature, time embeddings in attention have been used to tackle irregularly sampled time series. For instance, Shukla and Marlin proposed multi-time attention computed over the Time2Vec embeddings of observed time points [23]. Following the footsteps of Shukla and Marlin, Zhang et al. extended the multi-time attention idea to predictive models with multi-modal electronic health records [24]. The proposed FATE-Net does not compute attention weights over the embeddings of all the time points. Unlike the previous studies, we treat the learned embedding of the current time point as a query that controls how much to attend to each input biomarker. To further analyze the contribution of Time2Vec embeddings, we conducted an ablation study comparing the performance of FATE-Net with and without the Time2Vec module. Experimental results also supported the hypothesis that time embeddings improve predictive performance, particularly MCI-to-AD conversion recall. Section 3.3 presents the proposed self-attention approach.

3.2 Attribute encoder

Identifying the most relevant input biomarkers that can facilitate predictive modeling for MCI to AD conversion is essential. Feature selection before model training is one of the approaches. However, such a feature selection requires domain knowledge and might eliminate the possibility of discovering unprecedented factors contributing to disease progression modeling. Therefore, the FATE-Net aims to learn the contributions of the input attributes to construct the final representation.

The proposed attribute encoder first decouples input biomarkers by projecting each into two separate shared latent spaces. Following the self-attention terminology, we call these latent representations the key and value of each attribute. Given an input vector at time step *i*, $\mathrm{x}_i\in\mathbb{R}^d$, we project each attribute to key and value vectors as given below.

$$\mathrm{k}_{ij}=\mathrm{ReLU}\left(\mathrm{W}_{K2}\mathrm{ReLU}\left(\mathrm{w}_{K1}x_{ij}\right)\right) \tag{2}$$

$$\mathrm{v}_{ij}=\mathrm{ReLU}\left(\mathrm{W}_{V2}\mathrm{ReLU}\left(\mathrm{w}_{V1}x_{ij}\right)\right) \tag{3}$$

where k_{ij} and v_{ij} are the d'-dimensional key and value vectors of the *j*-th attribute at *i*-th time step, $\mathrm{w}_{K1}\in\mathbb{R}^{d''}, \mathrm{W}_{K2}\in\mathbb{R}^{d'\times d''}$ and $\mathrm{w}_{V1}\in\mathbb{R}^{d''}, \mathrm{W}_{V2}\in\mathbb{R}^{d'\times d''}$ are projection layers. Thus, the attribute encoder outputs a key-value pair for each input attribute. The key and the value latent spaces do not need the same dimensionality. However, the keys should be represented in the same space as the time embedding since the dot-product similarity between attribute keys and their visit time embedding needs to be computed.

One of the primary motivations in proposing attribute-wise key-value pairs is to expand the encoded amount of information in higher dimensional spaces than univariate attributes. The second and most important primary motivation is to facilitate comprehension of the individual contributions of the input biomarkers to the model's decisions, which is a key step toward model interpretability. The following section presents how the individual attribute representations are fused together by employing attention.

3.3 Self-attention and classification

The self-attention layer obtains a representation of the patient at the *i*th visit by combining the attribute embeddings. The contributions of the input attributes to the patient representation might vary based on the time point of the disease progression path. For this reason, we compute the attention weights by treating the time embedding as the query. The attention weights over input attributes are computed by the dot products between the time embedding and the keys of each attribute. Then, the weighted sum of the value vectors, given below, yields the final embeddings that can be fed to a classifier.

$$\mathrm{h}_i=\sum_{j=1}^{d}\alpha_{ij}\mathrm{v}_{ij} \tag{4}$$

$\mathrm{h}_i\in\mathbb{R}^{d'}$ is the *i*th time step's final representation before the classifier, and α_{ij} is the attention weight of the *j*th input dimension at *i*th time step computed as follows:

$$\alpha_{ij}=\frac{e^{\mathrm{t2v}(\tau_i)^T\mathrm{k}_{ij}/\sqrt{d'}}}{\sum_{k=1}^{d}e^{\mathrm{t2v}(\tau_i)^T\mathrm{k}_{ik}/\sqrt{d'}}} \tag{5}$$

The attention mechanism of FATE-Net and the self-attention proposed in Transformer networks [7] are different. In transformer-based frameworks, the input is an embedding sequence, such as a multivariate time series and a sequence of words. When constructing a representation for each element of the sequence, the self-attention method allows each element to attend to the other elements differently. Since the order of the elements matters, positional encodings are added to the input. Thus, the self-attention aims to model the temporal dynamics of a sequence. However, in this study, time and attribute encoders do not require a sequential input to output a visit representation.

Each visit has a raw representation, a vector of input biomarkers. Instead of feeding this vector directly to a fully connected layer, we first elevate each dimension corresponding to a biomarker into higher dimensional spaces. Thus, the input biomarkers are not blended right from the beginning. Instead, separate attribute representations and time embeddings are used to compute attention weights over the attributes instead of the temporal sequence. As seen in Eq. 4, the final visit representation is constructed as a weighted linear combination of the attribute embeddings. Since the framework is trained end-to-end, all the embeddings and how they are combined are tuned to improve the forecasting task. For this reason, we hypothesize that the learned attention weights should indicate the contributions of each attribute, bridging the gap between the model's predictions and input biomarkers.

We investigate the predictive performance of the FATE-Net with several classifier training options. First, we add fully connected layers to train a binary classifier with cross-sectional h_i embeddings. Secondly, we obtain a sequence of embeddings, $\{\mathrm{h}_1, \cdots, \mathrm{h}_T\}$ and train an LSTM network as the classifier. The other options are to replace LSTM with a standard transformer network or a one-dimensional Convolutional Neural Network (1D-CNN), which could be used in AD classification [25]. Each option is trained end-to-end with time and attribute encoders as given in Algorithm 1 using the binary cross-entropy loss function given below.

Table 1 Number of patients in TADPOLE and NACC datasets across training, development, and test sets

Patient group	Train	Development	Test	Total
TADPOLE				
sMCI	330	92	104	526
pMCI	193	47	58	298
AD starter	206	44	67	317
Total	729	183	229	1141
NACC				
sMCI	1525	378	463	2366
pMCI	1148	286	376	1810
AD starter	4283	1076	1335	6694
Total	6956	1740	2174	10870

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} -y_n \log\left(\text{FATE-Net}\left(\mathrm{x}_n, \tau_n\right)\right) - (1-y_n)\log\left(1-\text{FATE-Net}\left(\mathrm{x}_n, \tau_n\right)\right) \tag{6}$$

where y_n is the ground truth label of the nth visit, N is the total number of visits of all the training patients, and x_n and τ_n are the input feature vector and the time step of the nth visit, respectively.

Require: Training set of (visit, annotations, visit time) triplets: $\mathcal{D} = \{(\mathrm{x}_1, y_1, \tau_1), \cdots, (\mathrm{x}_N, y_N, \tau_N)\}$,
Ensure: FATE-Net model parameters
for $e = 1, \cdots,$ Epochs **do**
  for $i = 1, \cdots,$ Iterations **do**
    $\mathbf{t2v}_i = \texttt{TimeEncoder}\left(\tau_{ij}\right)$
    for $j = 1, \cdots, d$ **do**
      $\mathbf{k}_{ij}, \mathrm{v}_{ij} = \texttt{AttributeEncoder}\left(x_{ij}\right)$
    end for
    $\{\alpha_{ij} = \texttt{AttentionModule}\left(\mathbf{t2v}\left(\tau_i\right)[j], \mathbf{k}_{ij}\right)\}_{j=1}^{d}$
    Obtain visit representations with Eq. 5
    $\hat{y}_i = \texttt{Classifier}$ (VisitRepresentation)
    Compute Eq. 6
    Update FATE-Net parameters
  end for
end for

Algorithm 1 FATE-Net training

4 Experiments

Experiments are conducted on two publicly available datasets with different numbers of patients and varying biomarkers. All the models are implemented in Python, employing the Pytorch library for deep model implementation and training. Results are reported with 5-fold cross-validation.

4.1 Datasets

The experiments are conducted using The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) Challenge [12, 13], and Uniform Data Set (UDS) version 3.0 from National Alzheimer's Coordinating Center (NACC) [14].

The details about the patients and their visits in the selected cohorts from both datasets are presented in Tables 1 and 2, respectively. Patients are categorized into three groups based on their disease progression path. As mentioned in Sect. 2, sMCI patients are those who begin with MCI and remain diagnosed with MCI, with no observed progression to AD; while pMCI patients convert to AD in subsequent visits. AD starters are the patients who are initially diagnosed with AD and continue with the same diagnosis over time.

Visits are examined based on both diagnostic categories, which are MCI and AD, and transition patterns, including MCI-to-MCI, MCI-to-AD, and AD-to-AD. MCI-to-MCI visits refer to visits where the patient remains diagnosed with MCI. MCI-to-AD visits are those where the patient converts from MCI to AD; that is, they were diagnosed with MCI in their previous visit, however they have progressed

Table 2 Number of visits in TADPOLE and NACC datasets across training, development, and test sets

Visit label	Train	Development	Test	Total
TADPOLE				
MCI	2694	758	856	4308
AD	1611	361	507	2479
Total	4305	1119	1363	6787
MCI-to-MCI	2694	758	856	4308
MCI-to-AD	193	47	58	298
AD-to-AD	1418	314	449	2181
Total	4305	1119	1363	6787
NACC				
MCI	7328	1863	2259	11450
AD	17936	4622	5666	28224
Total	25264	6485	7925	39674
MCI-to-MCI	7328	1863	2259	11450
MCI-to-AD	1148	286	376	1810
AD-to-AD	16788	4336	5290	26414
Total	25264	6485	7925	39674

to AD in their current visit. AD-to-AD visits refer to visits where the patient remains diagnosed with AD.

4.1.1 TADPOLE

The TADPOLE challenge dataset is a subset of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database comprising patient visit records for cognitively normal (CN), MCI, and AD patients with demographics, cognitive test scores, and MRI and PET scans [12, 13]. The database was launched as a public-private partnership led by Michael W. Weiner, MD in 2003. For the latest information, please visit www.adni-info.org. We used TADPOLE_D1_D2_spreadsheet in our analysis and selected 14 biomarkers among those recommended by the TADPOLE challenge. The selected biomarkers have less than a 50% missing rate, which are CDRSB, ADAS11, ADAS13, MMSE, RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, FAQ, Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, and MidTemp. The visit data is collected at 6-month intervals. The maximum number of visits in the selected cohort is 19, while most patients have fewer than 10 visits.

This study focuses only on visits labeled as MCI and AD. Thus, the selected cohort comprises 1141 patients with 6787 visits labeled as MCI and AD, with 4308 and 2479 visits, respectively. Out of these patients, 26% (298 patients) of them start as MCI and convert to AD after some time. The aim is to focus on these patients and capture the risk of developing AD. The patients are split into five splits, each containing 913 (80%) patients for training and 228 (20%) patients for testing. The training patients are further split into 729 (80%) and 184 (20%) patients for training and development set, respectively. For both datasets, hyperparameter tuning is done on the development set to prevent information leaks. After the model is chosen, it is trained on the larger training set (which consists of training and development sets), and the final results are reported using 5-fold cross-validation.

In the case of missing values, forward filling is used when data exists from the previous time steps; otherwise, backward filling is used. Features related to MRI (Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, MidTemp) are normalized by dividing them by the intracranial volume of each patient. Min-max scaling is applied to all the features.

4.1.2 NACC

The Uniform Data Set (UDS) version 3.0 was published by the National Alzheimer's Coordinating Center (NACC) in 2015, containing demographics, patient history, cognitive test scores, MRI scans, and clinician notes [14]. The NIA-AA criterion is used to diagnose AD. For more details, please see UDS Researchers Data Dictionary https://files.alz.washington.edu/documentation/uds3-rdd.pdf. The following 14 features are selected in the experiments: NACCAGE, CDRSUM, BILLS, TAXES, SHOPPING, GAMES, STOVE, MEALPREP, EVENTS, PAYATTN, REMDATES, TRAVEL, FAQ, MAXFAQ.

The Functional Activities Questionnaire (FAQ) measures patients' abilities in their daily life activities. It is the sum of scores for BILLS, TAXES, SHOPPING, GAMES, STOVE, MEALPREP, EVENTS, PAYATTN, REMDATES, TRAVEL features, with a score of 0 indicating normal ability and a score of 3 indicating dependence. While the total score of the FAQ is informative, clinicians pay attention to cases with an unexpectedly high score of an FAQ item, as they are critical for the decisions on the diagnosis [26]. For instance, a patient with a low FAQ total score is likely to carry out their daily activities. However, a score, e.g., PAYATTN, that is too high can be concerning for clinicians, who may take additional precautions. To address this, MAXFAQ, the highest score among the 10 FAQ features, is added as an extra attribute. Each patient has at most two MRI scans, which is insufficient for temporal models. Therefore, MRI features are excluded from the analysis.

After eliminating patients with high missing rates, the selected group has 10,870 patients with 39,674 visits in total. The patients are split into five splits, each containing 8696 (80%) patients for training and 2174 (20%) patients for testing. The training patients are further split into 6956 (80%) and 1740 (20%) patients for training and development set, respectively.

The visits are labeled according to the NACCUDSD value, with 11,450 visits labeled as MCI and 28,224 as

AD. Patients with an AD conversion constitute 17% of the selected patients, resulting in 1810 subjects. The maximum number of visits is 17, and most patient sequences have less than six visits. As mentioned in Sect. 4.1.1, missing values are filled with forward and backward filling.

4.2 Implementation details

Each visit's ground-truth label is set as the next visit's diagnosis. Both datasets have patterns of visits transitioning from MCI to AD and back to MCI. Given the progressive nature of AD, occurrences where a patient regresses from AD to MCI are considered anomalies. Such time steps are excluded to prevent any potential misleading conclusions by the model. Both datasets have a limited number of instances where MCI patients convert to AD, posing a challenge in learning this specific pattern. Conversion recall is separately reported to assess the performance in capturing the conversions. This metric explicitly considers the patients' conversion time steps and represents the AD recall at such critical time steps.

FATE-Net described in Sect. 3 is implemented as follows: Attribute encoder is designed as a 2-layer fully connected network with dimensions [32, 64] and is utilized to obtain key and value vectors. Similarly, the time encoder's dimension is initially increased to 32 and then to 64. After applying self-attention using Eqs. 3 and 5, a 3-layer network with dimensions [64,16,2] is employed for classification. The layers have the Leaky Rectified Linear Unit (ReLU) with a leakiness of 0.2 and the dropout with a probability of 0.2 between them.

An ablation study is conducted where only the attribute encoder is employed to observe the impact of the time encoder. The query vector is obtained in the same way as in Eqs. 2 and 3 for key and value vectors, and the remaining procedure is the same. The results for FATE-Net with and without Time2Vec (t2v) embeddings are given in Table 3. The time and attribute encoders are further explored by introducing multiple attention heads. Testing with 2, 3, and 4 heads reveals that employing only one head produces the best results in both datasets. Therefore, the results obtained with a single attention head are reported.

FATE-Net is also trained with temporal classifiers. For this purpose, a sequence of embeddings is obtained for each patient. Each time step corresponds to a visit and constitutes 64-dimensional embeddings obtained from Eq. 4. The

Table 3 Performance of predicting the next visit's diagnosis

Model	Precision		Recall		F1	AUC	Conversion recall
	MCI	AD	MCI	AD			
TADPOLE							
LR	0.898 (0.020)	0.879 (0.021)	0.936 (0.012)	0.817 (0.025)	0.891 (0.009)	0.876 (0.008)	0.369 (0.083)
SVM	0.899 (0.017)	**0.882 (0.016)**	**0.937 (0.012)**	0.817 (0.023)	**0.893 (0.009)**	0.877 (0.009)	0.353 (0.059)
XGBoost	0.897 (0.016)	0.847 (0.014)	0.914 (0.016)	0.818 (0.021)	0.879 (0.010)	0.866 (0.006)	0.397 (0.046)
TabNet	0.898 (0.017)	0.863 (0.023)	0.925 (0.018)	0.816 (0.037)	0.885 (0.005)	0.871 (0.011)	0.383 (0.095)
LSTM	0.917 (0.014)	0.758 (0.029)	0.839 (0.030)	0.868 (0.016)	0.852 (0.015)	0.854 (0.009)	0.623 (0.037)
1D-CNN	0.939 (0.017)	0.789 (0.034)	0.861 (0.027)	0.904 (0.016)	0.878 (0.013)	**0.882 (0.010)**	0.600 (0.057)
Transformer	**0.940 (0.019)**	0.762 (0.035)	0.835 (0.038)	**0.908 (0.022)**	0.864 (0.020)	0.871 (0.015)	0.598 (0.186)
FATE-Net + FC	0.845 (0.040)	0.781 (0.047)	0.884 (0.028)	0.722 (0.044)	0.823 (0.036)	0.803 (0.035)	0.398 (0.048)
FATE-Net + FC w/o t2Vec	0.829 (0.025)	0.764 (0.046)	0.881 (0.020)	0.678 (0.073)	0.805 (0.023)	0.779 (0.033)	0.352 (0.063)
LSTM+FATE-Net	0.929 (0.021)	0.669 (0.043)	0.740 (0.064)	0.901 (0.031)	0.803 (0.032)	0.820 (0.024)	**0.743 (0.089)**
1D-CNN+FATE-Net	0.920 (0.022)	0.644 (0.031)	0.717 (0.032)	0.890 (0.038)	0.785 (0.010)	0.803 (0.010)	0.742 (0.068)
Transformer+FATE-Net	0.893 (0.046)	0.634 (0.034)	0.719 (0.030)	0.850 (0.057)	0.771 (0.014)	0.785 (0.019)	0.562 (0.162)
NACC							
LR	0.798 (0.008)	0.930 (0.007)	0.831 (0.007)	0.915 (0.005)	0.891 (0.006)	0.873 (0.005)	0.515 (0.023)
SVM	0.794 (0.008)	0.931 (0.007)	0.833 (0.008)	0.912 (0.006)	0.890 (0.007)	0.872 (0.006)	0.512 (0.015)
XGBoost	0.791 (0.010)	0.936 (0.007)	0.847 (0.011)	0.909 (0.006)	0.892 (0.005)	0.878 (0.006)	0.476 (0.024)
TabNet	0.790 (0.018)	**0.939 (0.012)**	**0.855 (0.022)**	0.907 (0.009)	**0.893 (0.006)**	**0.881 (0.008)**	0.472 (0.032)
LSTM	**0.821 (0.010)**	0.912 (0.003)	0.778 (0.012)	**0.931 (0.010)**	0.886 (0.005)	0.854 (0.002)	0.656 (0.015)
1D-CNN	0.804 (0.006)	0.925 (0.006)	0.816 (0.004)	0.919 (0.006)	0.890 (0.006)	0.867 (0.004)	0.545 (0.029)
Transformer	0.814 (0.034)	0.902 (0.017)	0.752 (0.046)	0.929 (0.021)	0.876 (0.008)	0.840 (0.013)	**0.727 (0.260)**
FATE-Net + FC	0.773 (0.006)	0.915 (0.008)	0.795 (0.007)	0.905 (0.007)	0.874 (0.008)	0.850 (0.007)	0.548 (0.014)
FATE-Net + FC w/o t2Vec	0.775 (0.006)	0.914 (0.009)	0.791 (0.010)	0.907 (0.006)	0.849 (0.008)	0.849 (0.008)	0.557 (0.016)
LSTM+FATE-Net	0.798 (0.010)	0.893 (0.014)	0.727 (0.024)	0.925 (0.006)	0.866 (0.013)	0.826 (0.014)	0.716 (0.049)
1D-CNN+FATE-Net	0.775 (0.006)	0.913 (0.008)	0.788 (0.008)	0.907 (0.007)	0.873 (0.009)	0.847 (0.007)	0.557 (0.017)
Transformer+FATE-Net	0.772 (0.010)	0.917 (0.008)	0.800 (0.006)	0.904 (0.010)	0.874 (0.009)	0.852 (0.007)	0.624 (0.026)

The results represent the average performance from 5-fold cross-validation, with the standard deviation provided in parentheses

The highest metrics for the best performing models are highlighted in bold

sequences of the patients are zero-padded and then fed to LSTM and 1D-CNN models, where their dimensionality is increased to 128. Fully connected layers with dimensions of [64,16,2] are employed to make the final predictions. Adam optimizer is used. FATE-Net with LSTM is trained with a learning rate of 0.0005 for 1000 epochs on TADPOLE and 900 on NACC. FATE-Net with 1D-CNN is trained with a learning rate of 0.0005 for 900 epochs on TADPOLE and 100 epochs on NACC. Similarly, FATE-Net with transformer utilizes the embeddings given by the self-attention module as input. It is trained on TADPOLE with a learning rate of 0.0001 for 1500 epochs and on NACC with a learning rate of 0.0005 for 100 epochs.

4.3 Baselines

The FATE-Net models are compared with models using cross-sectional data, such as SVM, LR, eXtreme Gradient Boosting (XGBoost), and TabNet [27]. The forecasting performances of these models are presented in Table 3. For both datasets, FATE-Net with a fully connected classifier (FATE-Net+ FC) surpasses the cross-sectional baseline models regarding conversion recall without sacrificing too much from F1 and AUC scores.

Temporal models require inputs of the same sequence length; therefore, padding is applied to the sequences before feeding them to the sequential classifiers. LSTM, 1D-CNN, and transformer models are trained with and without FATE-Net. LSTM and 1D-CNN models are trained with a hidden dimensionality of 32, followed by a fully connected classifier consisting of hidden dimensionalities of [64, 16, 2]. Adam optimizer is used with a learning rate of 0.001. LSTM is trained on TADPOLE and NACC for 1000 and 500 epochs, respectively. Meanwhile, 1D-CNN is trained on TADPOLE and NACC for 400 and 1500 epochs, respectively. The transformer model first increases the dimensionality to 32. It then passes the input through two encoder layers, each containing four attention heads and a feed-forward network with a dimensionality of 64 and a dropout probability of 0.2. The model is trained using the Adam optimizer with a learning rate of 0.0005 for 250 epochs on TADPOLE and a learning rate of 0.001 for 200 epochs on NACC.

4.4 Forecasting performance

The performance of cross-sectional and temporal models with and without FATE-Net are presented in Table 3. Among the cross-sectional models, which are FATE-Net + FC, LR, SVM, XGBoost, and TabNet, FATE-Net with FC has the highest performance in terms of conversion recall for both datasets. However, the other cross-sectional models have higher F1 and AUC, implying a better overall performance. Among the temporal models, FATE-Net with LSTM has the highest conversion recall for TADPOLE, and the second-highest conversion recall comes from FATE-Net with 1D-CNN. For NACC, the highest conversion recall is 0.727 (0.260), achieved with the transformer, while the second highest conversion recall is 0.716 (0.049), achieved with FATE-Net + LSTM. It is worth noting that even though the transformer has a higher conversion recall on average, it has a high standard deviation, suggesting that it is not as stable as FATE-Net with LSTM.

Temporal models use previous time steps to make predictions, but they have not been able to perform better than the baseline cross-sectional models. However, temporal models have superior conversion recall, indicating their ability to capture the progression path of the disease and detect MCI to AD conversion. Additionally, incorporating FATE-Net into baseline temporal models, such as LSTM and 1D-CNN, improves their conversion recall for both datasets, except for the transformer model. This means that FATE-Net enhances conversion recall for temporal models, which helps them surpass temporal models without FATE-Net and cross-sectional models without significantly compromising overall performance. This demonstrates the benefits of using the visit representations from the proposed framework in temporal models.

The contribution of the time encoder is further investigated by training fully connected FATE-Net with and without Time2Vec embeddings. The results show that adding the Time2Vec module to FATE-Net for TADPOLE improves F1, AUC, and particularly conversion recall. For TADPOLE, FATE-Net without Time2Vec has the lowest conversion recall, while it shows moderate performance for NACC. The performance of models for recalling AD is better on the NACC dataset than on the TADPOLE dataset. This can be attributed to the fact that NACC has more than five times the number of visits compared to TADPOLE. Additionally, the proportion of AD patients in NACC is considerably higher than in TADPOLE.

4.5 Attribute importance

We could consider two main directions in the explainable machine learning field. In one research direction, studies aim to train inherently interpretable models. The other direction designs algorithms to explain the model's decisions after the model is traditionally trained. This study falls under the first category, where the FATE-Net's architecture is designed to facilitate comprehension of the association between the model's prediction and the input biomarkers. Using attention is one of the techniques to improve the model's interpretability. For instance, Shah et al. employs a vision transformer architecture to classify different stages

of AD using MRI images [28]. The authors also discussed enhancing the explainability and interpretability of their model using self-attention [28]. Their study focuses on revealing the relationship between local regions on the MRI images and the model's decision. Regardless of the input data modality, using attention is one of the prominent techniques recognized by the explainable AI literature to enhance the interpretability of a model.

FATE-Net constructs a visit representation as the weighted linear combinations of the latent representations of each input attribute. The weights associated with each attribute are learned based on the dot-product similarity between the latent representations of the time and attributes. Time, used as the query, is posed as a modality that controls each attribute's contributions since the input biomarkers' importance might change with time. Our experiments showed that MRI features become more important for conversion visits. At the same time, cognitive scores could be more significant in classifying the subsequent visits of stable MCI and AD patients. The attention weights are examined for conversion patients to understand the significance of the input biomarkers to show whether the attention mechanism behaves expectedly. For this reason, we investigate the attention weights of the best-performing model, FATE-Net with LSTM, on the TADPOLE dataset. MCI patients on the test set are separated into progressive MCI (pMCI) and stable MCI (sMCI) groups based on whether they convert to AD or not, respectively. The attention weights of all correctly classified visits of sMCI patients are subtracted from those of pMCI patients. The average difference of the attention weights for each attribute is presented for the TADPOLE dataset in Fig. 2.

All the MRI attributes (Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, and MidTemp) in Fig. 2 have positive sign, which indicates that their weights are higher for pMCI patients. This implies that the model gives greater importance to MRI attributes for pMCI patients who are at risk of developing AD. The highest difference in attention weights is observed for Ventricles, followed by FAQ, a cognitive test score. It is worth noting that as a patient progresses from MCI to AD, the ventricles in the brain expand [29]. Therefore, the volume of the ventricles serves as a strong indicator of AD, and it is reasonable for the model to pay more attention to this attribute for the pMCI patients.

Visits of pMCI and sMCI patients are investigated under two groups: visits where patients remained in the MCI class and visits where they progressed to AD. The attention weights of the first group are subtracted from that of the second group. Figure 3 shows the average weight differences. The figure explicitly answers how much more the model attends to the attributes at the conversion steps than stable MCI steps. The results are similar to those of Fig. 2, where all MRI attributes have positive signs and a higher average compared to cognitive score attributes. This suggests that the model gives higher importance to MRI attributes than usual while deciding on the conversion step in the TADPOLE dataset. This is, in fact, similar to a scenario where the physician first administers cognitive tests; however, for diagnosing the patient, an MRI scan is also required if it is possible to administer. In some instances, MRI scans can show signs of pathology that are not noticeable from

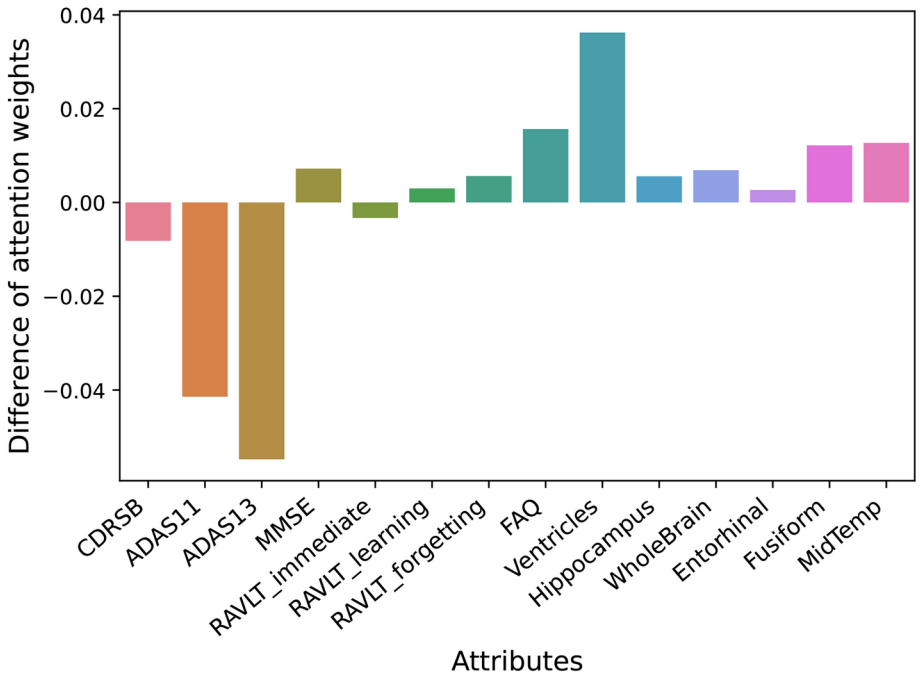


Fig. 2 Average difference of attention weights between progressive (pMCI) and stable (sMCI) MCI patients

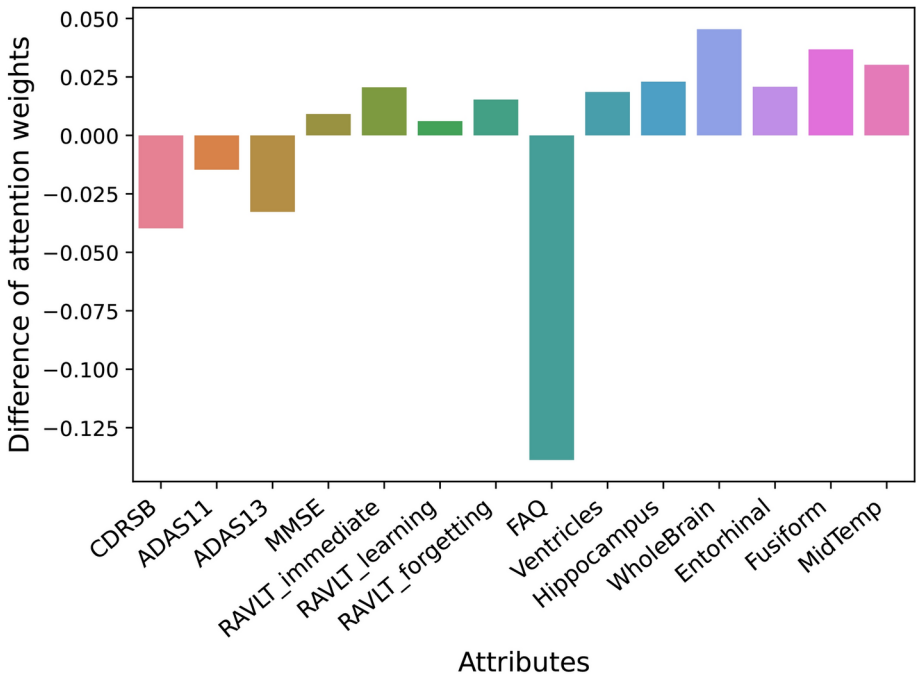


Fig. 3 Average difference of attention weights between MCI to AD and MCI to MCI visits

cognitive scores [30, 31]. This observation enhances the clinical reliability of the model.

While FATE-Net is specifically proposed to predict the conversion from MCI to AD in this study, it applies to any task concerning tabular electronic health records in broader applications. Attribute-wise attention is particularly beneficial in healthcare problems where a limited number of features are employed. Furthermore, FATE-Net can be trained with different classification heads, fully connected or recurrent for cross-sectional or longitudinal datasets. Regardless of the output layers, the FATE-Net provides feature importance in terms of attention weights for the task of interest. Thus, we propose an intrinsic interpretability with FATE-Net. This study does not claim that only using time and attribute encoders improves interpretability. However, employing attention along with the time and attribute encoders enables us to investigate further latent representations of which attributes are assigned higher importance during representation learning that impacts the classifier's decision.

4.6 Case study

This section presents an in-depth analysis of the behavior of FATE-Net with LSTM during the early, middle, and late stages of conversion for both datasets. For TADPOLE, FATE-Net with LSTM achieves the highest conversion recall, as shown in Table 3. For NACC, although the transformer model attains the highest conversion recall, it has a high standard deviation, indicating instability. FATE-Net achieves the second-highest conversion recall with LSTM, which has a much lower standard deviation, implying greater stability across different splits. Therefore, FATE-Net with LSTM is used in this analysis for both datasets. The outcomes of FATE-Net with LSTM on different patient groups are illustrated in Fig. 4a and b. This analysis provides insights into the models' performance and effectiveness in predicting conversion.

Figure 4 presents the forecasting performance of FATE-Net with LSTM, showing the predicted probability of developing AD at the next time step for each current time step. For instance, in Fig. 4a, the patient in the first row begins as MCI and converts to AD at the third visit (red denotes the ground truth). In the second time step, the model predicts a probability greater than 0.5, indicating that the patient is likely to develop AD in the third time step. Similarly, the patient in the third row starts as MCI and remains MCI for six visits. At the sixth time step, the model predicts conversion to AD at the seventh time step, aligning with the actual conversion.

The model's behavior is illustrated for early, middle, and late-stage conversion cases in the first, second, and third rows, respectively, while the final two rows highlight instances where the model's predictions are inaccurate. Given the progressive nature of AD, the probability of developing AD should generally increase or remain stable over time. The probability curves mostly align with this expectation, although some fluctuations are observed, as seen in the third, fourth, and fifth rows of Fig. 4b. These fluctuations may be attributed to variations in cognitive test scores, which can be affected by factors such as mood changes, stress, or other health conditions [32, 33].

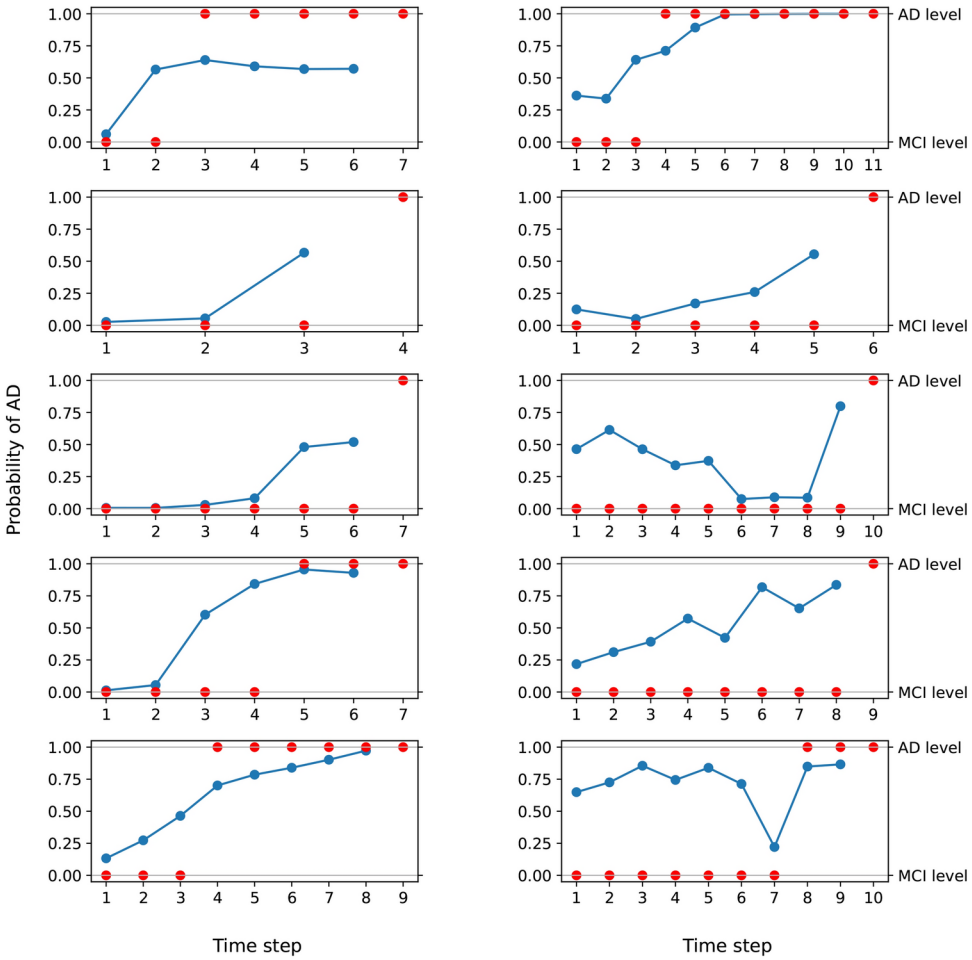


Fig. 4 Probability of developing AD at the next time step over time for different patient groups using FATE-Net with LSTM on (**a**) TADPOLE and (**b**) NACC datasets. The blue lines indicate predicted probabilities for the next time steps, while the red dots represent actual observed MCI and AD labels

4.7 Error analysis and clinical implications

The possible errors made by the model are examined and illustrated in the last two rows of Fig. 4. Specifically, the fourth row in the figure shows cases of false positives, while the fifth row depicts false negatives, each of which carries distinct clinical implications.

For the false positives, although the model's AD predictions occur somewhat earlier than the actual conversion, they are not necessarily harmful from a clinical perspective. For instance, the model predicts AD before the actual conversion for the fourth patient in Fig. 4a and the third and fourth patients in Fig. 4b. The model forecasts conversion for the fourth patient in Fig. 4b at the next visit. While the conversion does not happen immediately, the patient eventually develops AD. Such early predictions can be considered timely alerts for clinicians, prompting closer monitoring and preemptive interventions. These predictions align with the goal of early detection of Alzheimer's disease, which is critical for initiating timely therapeutic strategies that may delay progression.

Conversely, false negatives are undesirable due to their potential for delayed diagnosis and missed opportunities for early intervention. For example, the model underestimates the probability of developing AD for the fifth patient in Fig. 4a and b. In Fig. 4a, the probability of conversion at the third time step remains below 0.5, failing to flag the coming progression. Similarly, for the fifth patient in Fig. 4a, the model consistently predicts AD from the beginning but incorrectly forecasts MCI in the seventh time step, just before the patient's actual conversion to AD in the eighth step. Such errors highlight the need for cautious interpretation of predictions, as false negatives could lead to underdiagnosis and delayed treatment. Notably, baseline models such as logistic regression or standard LSTMs yield significantly lower conversion recall than FATE-Net, indicating that these methods are more prone to false negatives and may miss a greater number of high-risk cases.

The clinical applicability of FATE-Net lies in its promise to provide reliable conversion forecasts, which can support clinicians in stratifying patients based on their risk profiles and prioritizing those at higher risk for further evaluation

or monitoring. These predictions complement traditional diagnostic tools and help streamline decision-making by narrowing the focus to patients requiring closer attention. While false positives may lead to follow-up investigations for low-risk patients, they are less critical compared to false negatives, which could result in missed opportunities for timely intervention and treatment. Therefore, minimizing false negatives remains a key objective in ensuring the clinical utility of FATE-Net.

5 Conclusion

This study focuses on early prediction of the conversion from MCI to AD. The proposed approach, Feature Attentive Network with Time Embeddings (FATE-Net), comprises an attribute encoder and a self-attention mechanism, further enhanced by incorporating visit time into the network using time embeddings. Treating time embeddings as queries in the self-attention mechanism enables the identification of the attributes the model pays attention to at each step. The proposed model outperforms the baseline models in predicting the conversion steps from MCI to AD in two publicly available datasets. Moreover, it inherently supplies insights into its decision-making process with attribute-wise attention weights. While FATE-Net is specifically proposed for MCI to AD prediction, it is applicable to any task involving tabular electronic health records, especially in scenarios where feature interpretability is crucial. FATE-Net can be trained with fully connected LSTM, 1D-CNN, and transformer classifiers. In the experiments, FATE-Net achieves the best performance in terms of conversion recall when paired with the LSTM classifier for both datasets. Finally, the attribute attention shows that the MRI biomarkers substantially influence the conversion prediction. The FATE-Net could capture that progressive MCI patients undergo more significant volumetric changes in their brain regions than stable MCI patients.

In practice, FATE-Net could assist neurologists in identifying patients at risk of conversion and facilitate clinical research leading toward personalized treatments. While predicting the risk of conversion, the FATE-Net aims to highlight the input biomarkers that impact its decision in correlation with the time of the patient's visit. Thus, the primary use of the FATE-Net in practice would be presenting the extracted information from a patient cohort in the most informative and task-specific way possible, enabling neurologists to make more informed decisions without needing to examine every case. In addition, disclosing patients with a risk of conversion and its underlying factors is expected to inspire neurologists to investigate patient groups with characteristics similar to converting patients. Thus, the outcome of the proposed framework aims to pave the way for personalized treatments tailored for specific focus groups with identified risks of conversion.

Acknowledgements This work is supported by the Scientific and Technological Research Council of Turkey ARDEB 3501 under grant number 121E594. The NACC database is funded by NIA/NIH Grant U24 AG072122 and contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD). The TADPOLE data sets https://tadpole.grand-challenge.org constructed by EuroPOND http://europond.eu funded by the European Union's Horizon 2020 research and innovation program under grant No 666992 and was obtained from ADNI. Data collection and sharing was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by Laboratory for Neuro Imaging at University of Southern California. The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this paper. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Author contributions Elvan Karasu: Software, Validation, Visualization, Writing-original draft. İnci M. Baytaş: Conceptualization, Methodology, Validation, Writing-original draft.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Association A (2019) 2019 alzheimer's disease facts and figures. Alzheimer's Dementia 15(3):321–387. https://doi.org/10.1016/j.jalz.2019.01.010
2. Bucholc M, Titarenko S, Ding X, Canavan C, Chen T (2023) A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia. Expert Syst Appl 217:119541
3. Baytaş İM (2024) Predicting progression from mild cognitive impairment to alzheimer's dementia with adversarial attacks. IEEE J Biomed Health Inform 28(6):3750–3761. https://doi.org/10.1109/JBHI.2024.3373703
4. Ren Y, Li Y, Loftus TJ, Balch J, Abbott KL, Ruppert MM, Guan Z, Shickel B, Rashidi P, Ozrazgat-Baslanti T et al (2024) Identifying acute illness phenotypes via deep temporal interpolation and clustering network on physiologic signatures. Sci Rep 14(1):8442
5. Irfan M, Shahrestani S, Elkhodr M (2023) Early detection of alzheimer's disease using cognitive features: a voting-based ensemble machine learning approach. IEEE Eng Manag Rev 51(1):16–25. https://doi.org/10.1109/EMR.2022.3230820
6. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 6000–6010. Curran Associates, Inc., Long Beach, California, USA
8. Kazemi SM, Goel R, Eghbali S, Ramanan J, Sahota J, Thakur S, Wu S, Smyth C, Poupart P, Brubaker M (2019) Time2vec: learning a vector representation of time. arXiv preprint arXiv:1907.05321
9. Kung T-H, Chao T-C, Xie Y-R, Pai M-C, Kuo Y-M, Lee GGC (2021) Neuroimage biomarker identification of the conversion of mild cognitive impairment to Alzheimer's disease. Front Neurosci 15. https://doi.org/10.3389/fnins.2021.584641
10. Franciotti R, Nardini D, Russo M, Onofrj M, Sensi SL, Initiative ADN et al (2023) Comparison of machine learning-based approaches to predict the conversion to Alzheimer's disease from mild cognitive impairment. Neuroscience 514:143–152
11. Chang M, Brainerd C (2022) Predicting conversion from mild cognitive impairment to alzheimer's disease with multimodal latent factors. J Clin Exp Neuropsychol 44(4):316–335
12. Marinescu RV, Oxtoby NP, Young AL, Bron EE, Toga AW, Weiner MW, Barkhof F, Fox NC, Golland P, Klein S et al (2019) Tadpole challenge: accurate Alzheimer's disease prediction through crowdsourced forecasting of future data. In: Predictive intelligence in medicine: second international workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2, pp. 1–10. Springer
13. Petersen R, Aisen P, Beckett L, Donohue M, Gamst A, Harvey D, Jack C, Jagust W, Shaw L, Toga A, Trojanowski J, Weiner M (2010) Alzheimer's disease neuroimaging initiative (adni): clinical characterization. Neurology 74(3):201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25
14. Besser L et al (2018) Version 3 of the national alzheimer's coordinating center's uniform data set. Alzheimer Dis Assoc Disord 32(4), 351–358. https://doi.org/10.1097/WAD.0000000000000279
15. Tabarestani S, Aghili M, Shojaie M, Freytes C, Cabrerizo M, Barreto A, Rishe N, Curiel RE, Loewenstein D, Duara R, Adjouadi M (2019) Longitudinal prediction modeling of alzheimer disease using recurrent neural networks. In: 2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–4. https://doi.org/10.1109/BHI.2019.8834556
16. Nguyen M, He T, An L, Alexander DC, Feng J, Yeo BTT (2020) Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage 222:117203. https://doi.org/10.1016/j.neuroimage.2020.117203
17. El-Sappagh S, Saleh H, Ali F, Amer E, Abuhmed T (2022) Two-stage deep learning model for alzheimer's disease detection and prediction of the mild cognitive impairment time. Neural Comput Appl 34(17):14487–14509
18. Gao X, Cai H, Liu M (2023) A hybrid multi-scale attention convolution and aging transformer network for Alzheimer's disease diagnosis. IEEE J Biomed Health Inform 27(7):3292–3301. https://doi.org/10.1109/JBHI.2023.3270937
19. Chen Q, Fu Q, Bai H, Hong Y (2024) Longformer: longitudinal transformer for alzheimer's disease classification with structural mris. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 3575–3584
20. Kim JS, Han JW, Bae JB, Moon DG, Shin J, Kong JE, Lee H, Yang HW, Lim E, Kim JY, Sunwoo L, Cho SJ, Lee D, Kim I, Ha SW, Kang MJ, Suh CH, Shim WH, Kim SJ, Kim KW (2022) Deep learning-based diagnosis of Alzheimer's disease using brain magnetic resonance images: an empirical study. Sci Rep 12(1). https://doi.org/10.1038/s41598-022-22917-3
21. Subramanyam Rallabandi VP, Seetharaman K (2023) Deep learning-based classification of healthy aging controls, mild cognitive impairment and alzheimer's disease using fusion of mri-pet imaging. Biomed Signal Process Control 80:104312. https://doi.org/10.1016/j.bspc.2022.104312
22. Odusami M, Maskeliūnas R, Damaševičius R, Misra S (2024) Machine learning with multimodal neuroimaging data to classify

stages of alzheimer's disease: a systematic review and meta-analysis. Cogn Neurodyn 18(3):775–794
23. Shukla SN, Marlin B (2021) Multi-time attention networks for irregularly sampled time series. In: International conference on learning representations. https://openreview.net/forum?id=4c0J6lwQ4_
24. Zhang X, Li S, Chen Z, Yan X, Petzold LR (2023) Improving medical predictions by irregular multimodal electronic health records modeling. In: International conference on machine learning, pp. 41300–41313. PMLR
25. Wang R, Zeng L, Li H, Lu L, Li J, Luo X (2020) Alzheimer's disease classification based on one dimensional convolutional neural network. In: 2020 8th International conference on digital home (ICDH), pp. 288–293. IEEE
26. Vik A, Kociński M, Rye I, Lundervold A, Lundervold A (2023) Functional activity level reported by an informant is an early predictor of alzheimer's disease. BMC Geriatr 23. https://doi.org/10.1186/s12877-023-03849-7
27. Arik SO (2021) Pfister, T: Tabnet: attentive interpretable tabular learning. Proc AAAI Conf Artif Intell 35(8):6679–6687. https://doi.org/10.1609/aaai.v35i8.16826
28. Shah SMAH, Khan MQ, Rizwan A, Jan SU, Samee NA, Jamjoom MM (2024) Computer-aided diagnosis of alzheimer's disease and neurocognitive disorders with multimodal bi-vision transformer. Pattern Anal Appl 27(3):76
29. Nestor SM, Rupsingh R, Borrie M, Smith M, Accomazzi V, Wells JL, Fogarty J, Bartha R (2008) Alzheimer's Disease Neuroimaging Initiative: Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. Brain 131(9):2443–2454. https://doi.org/10.1093/brain/awn146https://academic.oup.com/brain/article-pdf/131/9/2443/792447/awn146.pdf
30. Johnson K, Fox N, Sperling R, Klunk W (2012) Brain imaging in alzheimer disease. Cold Spring Harbor Perspect Med 2:006213. https://doi.org/10.1101/cshperspect.a006213
31. Szczesniak D, Rymaszewska J, Zimny A, Sasiadek M, Połtyn-Zaradna K, Smith E, Zatońska K, Zatoński T, Rangarajan S, Yusuf S, Szuba A (2020) cerebral small vessel disease and other influential factors of cognitive impairment in the middle-aged: a long-term observational cohort pure-mind study in poland. GeroScience 43:1–17. https://doi.org/10.1007/s113f57-020-00271-4
32. Blair M, Coleman K, Jesso S, Desbeaumes Jodoin V, Smolewska K, Warriner E, Finger E, Pasternak SH (2016) Depressive symptoms negatively impact montreal cognitive assessment performance: a memory clinic experience. Can J Neurol Sci 43(4):513–517. https://doi.org/10.1017/cjn.2015.399
33. Bartfai A, Åsberg M, Beser A, Sorjonen K, Wilczek A, Warkentin S (2021) Impaired cognitive functioning in stress-induced exhaustion disorder: a new tablet-based assessment. BMC Psychiatry 21(1):459. https://doi.org/10.1186/s12888-021-03454-1

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.